UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE INFORMÁTICA

DEPARTAMENTO DE ARQUITECTURA DE COMPUTADORES Y AUTOMÁTICA

RECONOCIMIENTO DE PATRONES Y APRENDIZAJE AUTOMÁTICO EN BASES DE DATOS DE FUSIÓN NUCLEAR

PATTERN RECOGNITION AND MACHINE LEARNING ON NUCLEAR FUSION DATABASES

TESIS DOCTORAL DE:

GONZALO FARIAS CASTRO

DIRIGIDA POR:

SEBASTIÁN DORMIDO CANTO
MATILDE SANTOS PEÑAS

Madrid, 2013

Ph.D. Dissertation

Reconocimiento de Patrones y Aprendizaje Automático en Bases de Datos de Fusión Nuclear

Pattern Recognition and Machine Learning on Nuclear Fusion Databases

Gonzalo Farias Castro

Computer Science Engineer

Departamento de Arquitectura de Computadores y Automática

Facultad de Informática

Universidad Complutense de Madrid

Madrid, 2013

Ph.D. Dissertation

Pattern Recognition and Machine Learning on Nuclear Fusion Databases

Title	Pattern Recognition and Machine Learning on Nuclear Fusion Databases
Author	Gonzalo Alberto Farias Castro
Degree	Computer Science Engineering Faculty of Engineering, Sciences and Management De la Frontera University
Supervisors	Sebastián Dormido Canto, UNED Matilde Santos Peñas, UCM
Department	Architecture of Computers and Automatic Faculty of Computer Science Complutense University of Madrid
Program	Computer Science Engineering

Madrid, 2013

Tesis Doctoral

Reconocimiento de Patrones y Aprendizaje Automático en Bases de Datos de Fusión Nuclear

Título	Reconocimiento de Patrones y Aprendizaje Automático en Bases de Datos de Fusión Nuclear
Autor	Gonzalo Alberto Farías Castro
Grado	Ingeniería Civil Industrial Mención Informática Facultad de Ingeniería, Ciencias y Administración Universidad de la Frontera
Supervisores	Sebastián Dormido Canto, UNED Matilde Santos Peñas, UCM
Departamento	Arquitectura de Computadores y Automática Facultad de Informática Universidad Complutense de Madrid
Programa	Ingeniería Informática

Madrid, 2013

To my family...

Acknowledgements

I wish to thank all those who, in different ways, have contributed to the completion of this Thesis:

- To my supervisors Sebastián Dormido Canto and Matilde Santos Peñas, who have supported and helped me during all these years. I really appreciate your friendship and invaluable guidance to complete this work. I was lucky to meet both of you in my stay in Spain, two truly honest and good persons. This Thesis was possible just by your continuous encouragement.

- To Dr. Jesús Vega, who has given me the possibility to discover the fascinating worlds of patterns recognition and nuclear fusion. Jesús, deep thanks and sincere gratitude for your support and help all these years. I really enjoyed my short stay at CIEMAT, everyday it was a challenge, for a better science. I extend these thanks to the people of CIEMAT, and especially to Augusto, Sergio, Rodrigo, Ignacio and Ana.

- To Prof. Sebastián Dormido Bencomo, who has given me the opportunity to live nine wonderful years in Spain. It was an honor to work with you at UNED. Thanks a lot also for your friendship and to invite me so many times to watch Real Madrid matches at the fantastic Santiago Bernabeu Stadium!. A dream comes true since Zamorano times.

- To the people of the DACYA Department at UCM. Special thanks to Prof. Jesús Manuel de la Cruz for helping me on my arrival to the UCM.

- To Dr. David P. Schissel, who allowed me to work at General Atomics facilities in San Diego. This was a great experience for me, and I tried to do my best to take advantage of all your hospitality in order to improve this Thesis. Thanks also to all the people in General Atomics who helped me during my stay in the beautiful city of San Diego. I also would like to give special thanks to Xia Lee and Punit Gohil for their help to build the L-H transition time predictor.

- To the people of the DIA Department at the UNED. Thank you very much for your help and support during all these years. Special thanks to Pilar for all

the help given with administrative issues. Similarly, I wish to thank professors José Sánchez, María Antonia Canto, Joaquín Aranda, Fernando Morilla, José Luis Fernández Marrón, Raquel Dormido, Natividad Duro, Alfonso Urquía, and José Manuel Díaz.

- To my colleagues and friends of the DIA Department at the UNED. Thanks Rocío, Carla, Carlos, Arnoldo, Dictino, Victor, Miguel Angel, David, María, Luis, Alejandro, Jesús, Oscar and Ernesto. I have a really nice remember of my stay there.

- To all the people of the School of Electrical Engineering at PUCV. Special thanks to my colleagues and friends Héctor Vargas, Gabriel Hermosilla, Domingo Ruíz, Ariel Leiva, Jorge Mendoza, Sebastián Fingerhuth and Miguel López. My gratitude also to Prof. Paulino Alonso to give me the opportunity to work here.

- To my dear friend Elizabeth, who started this adventure in Spain with me. Eli, I will always owe you. I hope to have the opportunity to give back at least a small part that you gave me. Eli, people like you make this world a better place.

- To all my family, especially my mother Emelina, my sister Nitcy, my wife Ruth, my aunts Cecilia and Teresa, and my uncles José and Juan for their unconditional support and love. Tuty all makes sense with you to my side.

- Finally, to God who looks after my family and friends.

Contents

List of Tables

List of Figures

Abstract

Modern civilization increases its demand for energy year after year. Rapid urbanization in developing countries, and a population estimated over 10000 millions of people at the middle of this century, will require a large-scale electricity generation in next decades. Nowadays fossil fuels are the main source of energy because of their relative low cost of production and high energetic capacity. However, environmental requirements for low CO_2 emission sources and, the need to invest in long-term options, forces to develop new kinds of energy sources.

Nuclear sources can provide great quantities of energy. Although fusion energy is still developing, its potential is enormous, even compared with nuclear fission. Nuclear fusion, the energy source of the sun and stars, could be cheaper, cleaner and safer. Fusion power would provide much more energy than any other technology currently in use, and the fuel required for fusion, mainly deuterium, exits abundantly in the ocean.

In order to study the process of nuclear fusion, many experiments are performed in the experimental fusion devices. Every experiment produces thousands of signals, with enormous amounts of data. A typical discharge, an experiment of about tents of seconds, can generate until 10GB of data. It is estimated that similar experiments could storage until 1 TByte in more advanced devices in the future. However, nowadays only 10% of data is treated by computer algorithms. The rest 90% is not processed at all.

Hence, the current databases of experimental devices should be completely analyzed, in order to make fusion a future energy option by the middle of this century. For this reason, the PhD Thesis proposes the use of advanced pattern recognition and machine learning techniques to analyze fast and efficiently the massive fusion databases.

Although great efforts have been done so far, there is still room for improvements on this research line. Specifically, noise reduction on time series and images databases are needed to improve plasma diagnosis. Besides, automatic feature selection and multilayer approaches are required to have better classifiers and predictors of plasma behavior. Finally, information retrieval and searching like-patterns is also needed to reduce time analysis of entire databases. This document resumes the work performed in this research topic. The published journal articles and the main conclusions obtained during the PhD Thesis are descried in detail.

Resumen

Nuestra civilización incrementa su demanda de energía cada año. El rápido desarrollo de los centros urbanos en los paises emergentes, y una población estimada sobre los 10000 millones de personas hacia la mitad de este siglo, requerirá de una generación de electricidad a gran escala en las próximas décadas. Hoy en día los combustibles fósiles son la principal fuente de energía. Sin embargo, los actuales requerimientos de bajos niveles de emisión de CO_2 obligan al desarrollo de nuevos tipos de fuentes energéticas.

Las fuentes nucleares proporcionan grandes cantidades de energía. Aunque la energía de fusión aún esta en desarrollo su potencial es enorme, incluso comparada con la fisión nuclear. La fusión nuclear, la fuente de energía del sol y las estrellas, podría ser más barata, limpia y segura que la energía generada por las plantas nucleares. La fusión podría proporcionar un nivel mucho mayor de energía que cualquier otra tecnología existente, y el combustible requerido se encuentra en el océano de forma abundante.

Con el fin de estudiar el proceso de fusión, se realiza una gran cantidad de ensayos en dispositivos de fusión nuclear. Cada experimento genera miles de señales y una enorme cantidad de datos. Una descarga típica, un experimento que dura unas decenas de segundos, puede producir hasta 10GB de datos. Se prevé que en el futuro, en dispositivos de fusión más avanzados, se lograría alcanzar hasta 1TB. A pesar de la inmensa cantidad de información obtenida, se estima que hoy en día sólo un 10% de los datos son procesados. El restante 90% no es tratado en absoluto. Así, existe una necesidad de analizar por completo las actuales bases de datos, con el fin de convertir la fusión nuclear en una alternativa real de energía en la mitad de este siglo.

Por esta razón, la Tesis doctoral propone el uso de técnicas avanzadas para el reconocimiento de patrones y aprendizaje automático, con el fin de analizar de una forma más rápida y eficiente las inmensas bases de datos de fusión nuclear. Aunque se han realizado esfuerzos en este sentido, todavía existen grandes problemas a considerar. Específicamente, se requiere eliminar o reducir ruido de las señales obtenidas, se necesita construir clasificadores y predictores del comportamiento del plasma, y por último se necesita disminuir los tiempos de búsqueda de patrones similares que indiquen comportamientos equivalentes en el pasado. Este documento resume el trabajo realizado en esta línea de investigación, incluyendo los artículos publicados y las conclusiones obtenidas.

I

Summary of the Research

Chapter 1

Introduction

Energy is a crucial element for the subsistence of our modern civilization. Almost all human activities require energy to work. This requirement is increased year after year, especially due to the growing population, which is estimated about 10000 millions of people at the middle of this century.

Nowadays fossil fuels are the main source of energy because of their relative low cost of production and high energetic capacity. Nevertheless they are not a long-term option. Alternatives like renewable energies are increasing its participation in modern life. However, current technology of renewable sources is still not able to supply all energy needed. On the contrary nuclear sources can provide great quantities of energy.

1.1 Nuclear Fusion

Although fusion energy is still developing, its potential is enormous, even compared with nuclear fission. Nuclear fusion, the energy source of the stars, could be cheaper, cleaner and safer.

Nuclear fusion is the process by which two or more atomic nuclei join together to form a single heavier nucleus. This is usually accompanied by the release of large quantities of energy. Fusion is the process that powers active starts, the hydrogen bomb and some experimental devices. Fusion power would provide much more energy than any other technology currently in use, and the fuel required for fusion, mainly deuterium, exits abundantly in the ocean. Fusion could, in theory, supply all the energy needs of the world for millions of years.

In order to reproduce on the Earth the fusion power, some fusion reactions can be used. One of the most important is the deuterium-tritium cycle (Sheffield 1994), which release 17.58MeV, as it is shown in Equation (1.1)

$$ {}^{2}_{1}D + {}^{3}_{1}T \longrightarrow {}^{4}_{2}He\,(+3.52MeV) + {}^{1}_{0}n\,(+14.06MeV) \tag{1.1} $$

In a fusion device this reaction is produced at very high temperatures, about 150 million degrees Celsius. At this temperature, the matter inside fusion devices is found as plasma, which is a state of matter similar to gas with a portion of its particles ionized (Reitz & Milford 1996, Lawson 2002). Magnetic fields are used to confine plasma in the shape of a torus. Most common configurations for magnetic confinement of plasma are *stellerators* (Wakatani 1998) and *tokamaks* (Lister et al. 1997).

The International Thermonuclear Experimental Reactor (ITER) is an international nuclear fusion research and engineering project, which is currently building the world's largest and most advanced experimental tokamak nuclear fusion reactor at the Cadarache (France). ITER is expected to demonstrate that more energy is obtained than is used to initiate fusion process, something that has not been achieved by any experimental fusion reactor. After ITER, the first commercial demonstration fusion power plant, named DEMO, will be tried.

Currently there are many experimental fusion devices in operation. JET, the Joint European Torus, (EFDA 2013*a*) is an experimental tokamak reactor located in Oxfordshire (UK). It is currently the largest facility of its kind in operation. TJ-II (Alejaldre et al. 1999) is a medium size stellarator located at CIEMAT in Madrid (Spain). DIII-D (General Atomics 2013) is another tokamak machine developed by General Atomics in San Diego (USA).

1.2 Motivation and General Problem Formulation

Experiments on fusion reactors are carried out by producing discharges or shots, in which plasma exists inside the torus. The duration of a shot is normally tens of seconds (Ongena 2006). ITER would keep the shot for about 30 minutes.

During the discharges many diagnostics at several places on the reactor acquire data at high sampling frequencies. About 10 GBytes per discharge can be acquired in JET

(Vega et al. 2007). ITER could even storage 1 TByte per shot. Bolometry, density, temperature, and soft X-rays are just some examples of the thousands of data acquired during a discharge. Huge databases, with enormous amount of data, are a common situation in experimental fusion reactors.

However, nowadays only 10% of data is processed. The rest 90% is not processed at all. Therefore, in order to achieve fusion energy as a clean, inexhaustible, safe and cheap energy source, the current databases of experimental devices (tokamaks and stellarators) should be analyzed completely. Performing complete analysis will involve an optimal operation planning of ITER, and in turn, will be basic for a successful design of DEMO.

For that reason this Thesis propose the use of advanced pattern recognition and machine learning techniques in order to analyze in a faster and more efficient way massive fusion databases. Much work has been done before in this sense, but there is still room for improvements on this research line. Specifically, noise reduction on time series and images databases are needed to improve plasma diagnosis. Besides, automatic feature selection and multi-layer approaches are required to have better classifiers and predictors of plasma behavior. Finally, information retrieval and searching like-patterns is also needed to reduce time analysis of entire databases. All these tasks are considered in this work, which could contribute to a faster and deeper analysis in the huge fusion databases.

1.3 Thesis Objectives

The general objective of this thesis is to develop advanced pattern recognition methods and apply innovative machine learning techniques to perform automatic data analysis of massive nuclear fusion databases. Given the extremely wide range of topics to be considered, this thesis focuses on some particular issues of the nuclear fusion devices: TJ-II, JET and DIII-D. Nevertheless, many of the results obtained in the Thesis could be adapted or reproduced in other fusion machines such as ITER.

The following specific objectives are addressed in this thesis:

- Application of advanced classification and clustering techniques on fusion databases.
- Development of efficient searching and information retrieval methods of diagnostics from massive databases of fusion devices.

- Design and validation of algorithms to eliminate noise pattern in images. Applications to the Thomson Scattering diagnostic of a fusion device.

- Study and development of automatic feature selection procedures. Potential applications are the confinement transitions in thermonuclear plasmas.

1.4 Main Contributions

The contributions of this Thesis are summarized in the main conclusions of this work, and can be organized in the following developments and publications.

1.4.1 Developed Software Components

The concrete results of this PhD Thesis include the implementation of advanced pattern recognition & machine learning algorithms in MATLAB, and the development of graphical user interfaces for easy testing of data. The most important results are summarized below:

- MATLAB programming of custom pattern recognition tools.

- MATLAB graphical user interfaces for teaching and research purposes.

- Application of support vector machines and Wavelet transform for pattern recognition in fusion databases.

- Structural pattern recognition applied to the search of specific waveforms.

- Image classification in Thompson Scattering diagnostic.

- Detection of different physical plasma events in nuclear fusion signals.

1.4.2 Publications

During the PhD Thesis several articles have been published in specialized journals and international conferences. Most of the papers have been obtained as direct result of this Thesis. Others works, however, have been developed in collaboration by the author with different research groups.

Journal Papers Published

The following articles have been published in journals and are directly related with the PhD Thesis:

- G. Farias, S. Dormido-Canto, J. Vega, I. Pastor, M. Santos (2013) Application and validation of image processing algorithms to reduce the stray light on the TJ-II Thomson Scattering diagnostic, *Fusion Science and Technology*, ISSN 1536-1055, Volume 63, Number 1, Pages 20–25.

- G. Farias, J. Vega, S. González, A. Pereira, X. Lee, D. Schissel, P. Gohil (2012) Automatic determination of L/H transition times in DIII-D through a collaborative distributed environment, *Fusion Engineering and Design*, ISSN 0920-3796, Volume 87, Issue 12, Pages 2081–2083.

- S. Dormido-Canto, G. Farias, J. Vega, I. Pastor (2012) Image processing methods for noise reduction in the TJ-II Thomson Scattering diagnostic, *Fusion Engineering and Design*, ISSN 0920-3796, Volume 87, Issue 12, Pages 2170-2173.

- L. Makili, J. Vega, S. Dormido-Canto, I. Pastor, A. Pereira, G. Farias, A. Portas, D. Pérez-Risco, M.C. Rodríguez-Fernández, P. Busch (2010) Upgrade of the automatic analysis system in the TJ-II Thomson Scattering diagnostic: New image recognition classifier and fault condition detection, *Fusion Engineering and Design*, ISSN 0920-3796, Volume 85, Issues 34, Pages 415-418.

- M. Santos, G. Farias (2010) Laboratorios virtuales de procesamiento de señales, *Revista Iberoamericana de Automática e Informática Industrial (RIAI)*, ISSN 1697-7912, Volume 7, Number 1, Pages 91-100.

- J.A. Martín, M. Santos, G. Farias, N. Duro, J. Sánchez, R. Dormido, S. Dormido-Canto, J. Vega, H. Vargas, (2009) Dynamic clustering and modeling approaches for fusion plasma signals, *IEEE Transactions on Instrumentation and Measurement*, ISSN 0018-9456, Volume 58, Number 9, Pages 2969-2978.

- N. Duro, R. Dormido, J. Vega, S. Dormido-Canto, G. Farias, J. Sánchez, H. Vargas, A. Murari and JET-EFDA Contributors (2009) Automated recognition system

for ELM classification in JET, *Fusion Engineering and Design*, ISSN 0920-3796, Volume 84, Issues 2-6, Pages 712-715.

- S. Dormido-Canto, G. Farias, J. Vega, R. Dormido, J. Sánchez, N. Duro, H. Vargas, A. Murari, and JET-EFDA Contributors (2008) Classifier based on support vector machine for JET plasma configurations, *Review of Scientific Instruments*, ISSN 0034-6748, Volume 79, Pages 10F326-1/10F326-3.

- S. Dormido-Canto, G. Farias, R. Dormido, J. Sánchez, N. Duro, H. Vargas, J. Vega, G. Ratta, A. Pereira, A. Portas (2008) Structural pattern recognition methods based on string comparison for fusion database, *Fusion Engineering and Design*, ISSN 0920-3796, Volume 83, Issue 2-3, Pages 421-424. Ed. Elsevier.

- G. Rattá, J. Vega, A. Pereira, A. Portas, E. De la Luna, S. Dormido-Canto, G. Farias, R. Dormido, J. Sánchez, N. Duro, H. Vargas, M. Santos, G. Pajares, A. Murari, and JET EFDA Contributors (2008) First applications of structural pattern recognition methods to the investigation of specific physical phenomena at JET, *Fusion Engineering and Design*, ISSN 0920-3796, Volume 83, Issue 2-3, Pages 467-470. Ed. Elsevier.

- J. Vega, A. Pereira, A. Portas, S. Dormido-Canto, G. Farias, R. Dormido, J. Sánchez, N. Duro, M. Santos, E. Sánchez, G. Pajares (2008) Data mining technique for fast retrieval of similar waveform in Fusion massive databases, *Fusion Engineering and Design*, ISSN 0920-3796, Volume 83, Issue 1, Pages 132-139. Ed. Elsevier.

- S. Dormido-Canto, G. Farias, J. Vega, R. Dormido, J. Sánchez, N. Duro, M. Santos, J.A. Martín, G. Pajares (2006) Search and retrieval of plasma waveforms: structural pattern recognition approach, *Review of Scientific Instruments*, ISSN 0034-6748, Volume 77, Pages 10F514-1/10F514-4.

- G. Farias, S. Dormido-Canto, J. Vega, J. Sánchez, N. Duro, R. Dormido, M. Ochando, M. Santos, G. Pajares (2006) Searching for patterns in TJ-II time evolution signals, *Fusion Engineering and Design*, ISSN 0920-3796, Volume 81, Pages 1993-1997, Ed. Elsevier.

- N. Duro, J. Vega, R. Dormido, G. Farias, S. Dormido-Canto, J. Sánchez, M. Santos, G. Pajares (2006) Automated clustering procedure for TJ-II experimental signals, *Fusion Engineering and Design*, ISSN 0920-3796, Volume 81, Pages 1987-1991, Ed. Elsevier.

- S. Dormido Canto, J. Vega, Sánchez, G. Farias (2005) Information retrieval and classification with wavelets and support vector machines, *Lecture Notes in Computer Science*, ISSN 0302-9743, Volume 3562, Part 2, Pages 548-557, Springer-Verlag.

- G. Farias, R. Dormido, M. Santos, N. Duro (2005) Image classifier for the TJ-II Thomson Scattering diagnostic: Evaluation with a feed forward neural network, *Lecture Notes in Computer Science*, ISSN 0302-9743, Volume 3562, Part 2, Pages 604-612, Springer-Verlag.

- S. Dormido-Canto, G. Farias, R. Dormido, J. Vega, J. Sánchez, M. Santos and The TJ-II Team (2004) TJ-II wave forms analysis with wavelets and support vector machines, *Review of Scientific Instruments*, ISSN 0034-6748, Volume 75, Pages 4254-4257.

The following papers are the result of collaboration with research groups. The articles are related to the use of pattern recognition techniques to the biomedical applications.

- G. Farias, M. Santos, V. Loópez (2010) Making decisions on brain tumor diagnosis by soft computing techniques, *Soft Computing*, ISSN 1432-7643, Volumen 14, Number 12, Pages 1287-1296.

- G. Farias, M. Santos (2007) A computational fusion of wavelets and neuronal networks in a classifier for biomedical applications, *Lecture Series on Computer and Computational Sciences*, ISSN 1573- 4196, Volume 8, Pages 66-70, Brill Academic Publishers.

Papers in Conferences

The following articles have been published in national and international conferences mainly related to pattern recognition on fusion databases. Given the high number of contributions, only the most relevant ones are listed.

- Farias G., Vega J., Gonzalez S., Pereira A., Lee X., Schissel D., Gohil P. (2011) *Automatic determination of L/H transition times in DIII-D through a collaborative distributed environment*, 8^{th} IAEA Technical Meeting on Control, Data Acquisition, and Remote Participation for Fusion Research, June 20-24, 2011, San Francisco, USA.

- Dormido-Canto S., Farias G., Vega J., Pastor I. (2011) *Image processing methods for noise reduction in the TJ-II Thomson Scattering Diagnostic*, 8^{th} IAEA Technical Meeting on "Control, Data Acquisition, and Remote Participation for Fusion Research", June 20-24, 2011, San Francisco, USA.

- G. Farias, S. Dormido, F. Esquembre, H. Vargas, S. Dormido-Canto (2008) *Laboratorio virtual para la enseñanza de técnicas de reconocimiento de patrones*, XIII Latin-American Congress on Automatic Control. Mérida, Venezuela.

- G. Farias, M. Santos, V. López (2008) *Brain tumour diagnosis with wavelets and support vector machines*, 3rd International Conference on Intelligent System and Knowledge Engineering, Proceedings of the 3rd ISKE, IEEE Press, ISBN: 978-1-4244-2197-8, pp: 1453-1459, November 17-19, Xiamen, China.

- Martin, J. A. Santos, M., Farias, G., Duro, N., Sánchez, J., Dormido, R., Dormido-Canto, S., Vega, J. (2007) *Dynamic clustering and neuro-Fuzzy identification for the analysis of fusion plasma signals*, Proc. of the IEEE International Symposium on Intelligent Signal Processing. ISBN: 1-4244-0829-6. pp: 979-984.

- Vega, J., Rattá, G., Murari, A., Castro, P., Dormido-Canto, S., Dormido, R., Farias, G., Pereira, A. Portas, A., de la Luna, E., Pastor, I., Sánchez, J., Duro, N., Castro, R., Santos, M., Vargas, H. (2007) *Recent result on structural pattern recognition for fusion massive database*, Proc. of the IEEE International Symposium on Intelligent Signal Processing. ISBN: 1-4244-0829-6. pp: 949-954.

- Farias G., Dormido-Canto S., Vega J., Sánchez J., Duro N., Dormido R., Ochando M., Pajares G., Santos M. (2005) *Searching patterns in TJ-II temporal evolution signals with support vector machines*, Fifth IAEA Technical Meeting on Control, Data Acquisition, and Remote Participation for Fusion Research, Budapest, Hungary.

- Duro N., Vega J., Dormido R., Farias G., Dormido-Canto S., Sánchez J., Santos M., Pajares G. (2005) *Automated clustering procedure for TJ-II experimental signals*, Fifth IAEA Technical Meeting on Control, Data Acquisition, and Remote Participation for Fusion Research, Budapest, Hungary.

- Farias G., Santos M. (2005) *Aplicación de técnicas de inteligencia artificial y tratamiento de señales en fusión*, 1er Simposio de Control Inteligente, 1- 3 Junio, Huelva, España.

- Farias G., Santos M., Dormido-Canto S. (2005) *Desarrollo de una aplicación para la integración de técnicas de reconocimiento de patrones*, XXVI Jornadas de Automática, Alicante-Elche, España, ISBN: 84-689-0730-8.

- Vega J., Pastor I., Cereceda J. L., Pereira A., Herranz J., Pérez D., Rodríguez M. C., Farias G., Dormido-Canto S., Sánchez J., Dormido R., Duro N., Dormido S. (2005) *Application of intelligent classification techniques to the TJ-II Thomson Scattering diagnostic*, 32nd EPS Plasma Physics Conference, 8th International Workshop on Fast Ignition of Fusion Targets. 27 June- 1 July, Tarragona- Spain.

- Farias G., Santos M., Marrón J. L., Dormido-Canto S. (2004) *Determinación de parámetros de la transformada wavelets para la clasificación de señales del diagnóstico Scattering Thomson*, XXV Jornadas de Automática, Ciudad Real, España, ISBN: 84-688-7460-4.

- Dormido S., De la Cruz J.M., Vega J., Santos M., Dormido-Canto S., Sánchez J., Dormido-Canto R., Farias G. (2004) *Análisis de formas de onda de plasmas con wavelets y support vector machines*, 3ra. Conferencia Iberoamericana en Sistemas, Cibernética e Informática CISCI. Orlando, USA.

1.4.3 Short Research Stays

The PhD Thesis has involved two short research stays in the following fusion nuclear laboratories:

- **Period**: January, 2011.

  Laboratory: General Atomics (San Diego, USA).

  Objective: L/H transition studies on DIII-D fusion reactor.

Supervisors: Jesús Vega (CIEMAT) and David Schissel (General Atomics).

Publication G. Farias, J. Vega, S. Gonzalez, A. Pereira, X. Lee, D. Schissel, P. Gohil *Automatic determination of L/H transition times in DIII-D through a collaborative distributed environment*, 8^{th} IAEA Technical Meeting on "Control, Data Acquisition, and Remote Participation for Fusion Research", June 20-24, 2011, San Francisco, USA.

- **Period**: February to June, 2011.

  Laboratory: National fusion laboratory, CIEMAT (Madrid, Spain).

  Objective: L/H transition studies on DIII-D fusion reactor, and reduction of Thomson scattering stray-light.

  Supervisors: Jesús Vega (CIEMAT) and David Schissel (General Atomics).

  Publication S. Dormido-Canto, G. Farias, J. Vega, I. Pastor *Image processing methods for noise reduction in the TJ-II Thomson Scattering diagnostic*, 8^{th} IAEA Technical Meeting on "Control, Data Acquisition, and Remote Participation for Fusion Research", June 20-24, 2011, San Francisco, USA.

Chapter 2

Plasma Diagnostics and Physical Events

2.1 Plasma Diagnostics

The CIEMAT, and specifically the association EURATOM/CIEMAT for magnetic confinement fusion, has obtained from many different experiments a large number of signals in the nuclear fusion device TJ-II.

TJ-II (Alejaldre et al. 1999) is a medium sized stellarator fusion device (Heliac type, magnetic field $B_0 = 1.2T$, average major radius $R(0) = 1.5m$, average minor radius $\leq 0.22m$) located at CIEMAT (Madrid, Spain) that can explore a wide rotational transform range (see Figure 2.1). TJ-II plasmas are produced using electron cyclotron resonance heating (ECRH) (two gyrotrons, 300 kW each, 53.2 GHz, second harmonic, X-mode polarization) and additional neutral beam injection (NBI, 300 kW). At present, 940 digitization channels are available for experimental measurements in TJ-II. Fusion devices generate a massive database. Typically, thousands of signals with high dimensionality are collected per discharge.

In TJ-II, the magnetic trap is obtained by means of various sets of coils that completely determine the magnetic surfaces before plasma initiation. The toroidal field is created by 32 coils. The three-dimensional twist of the central axis of the configuration is generated by means of two central coils: one circular and one helical. The horizontal position of the plasma is controlled by the vertical field coils. The combined action of these magnetic fields generate bean-shaped magnetic surfaces that guide the particles

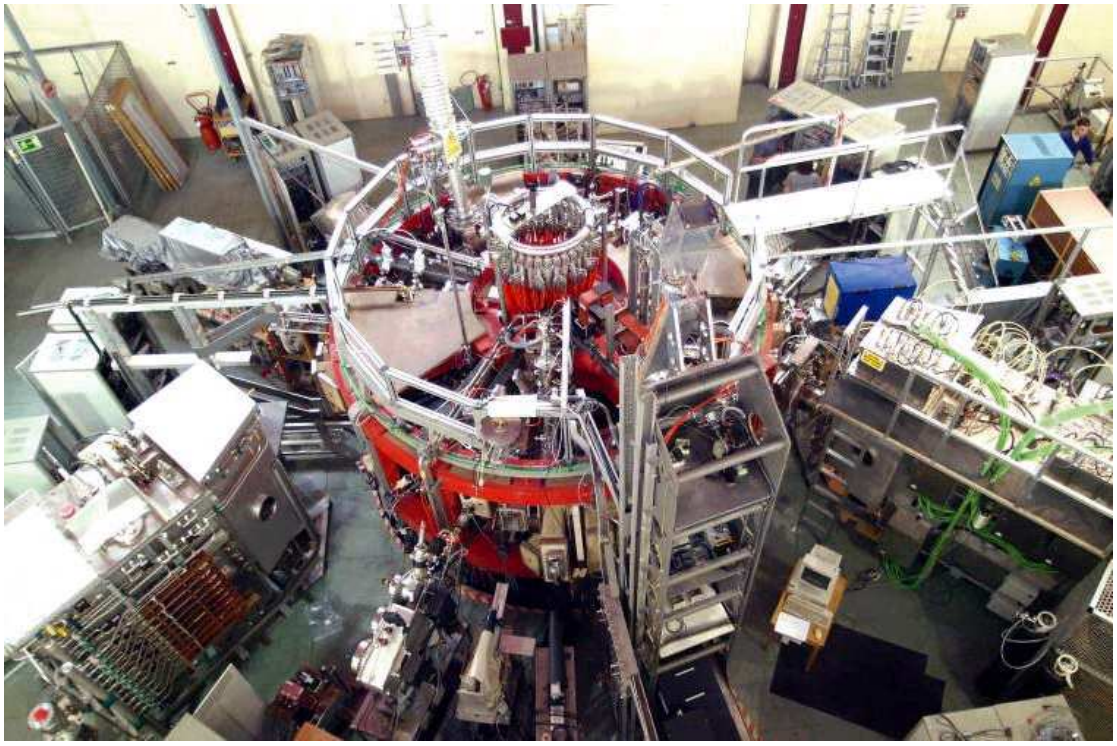


Figure 2.1: TJ-II Fusion device stellerator.

of the plasma so that they do not collide with the vacuum vessel wall (Alejaldre et al. 1999). As an example, Figure 2.2 shows an scheme of two sensors arrays to acquire bolometer signals.

TJ-II discharges last between 150-250ms, with a repetition frequency of about 7 minutes. Depending on the sampling rate, the number of samples could be in the range of 4000-16000 per discharge.

In this Thesis two type of signals have been analyzed. Temporal series (unidimensional data) from different diagnostics, and Images (two-dimensional data) from the Thomson Scattering diagnostic. The methodologies used to perform pattern recognition solutions could be extended to other kind of signals or fusion devices. Thus, data mining techniques are also used in the fusion devices DIII-D (General Atomics, San Diego, USA) and JET (EFDA, Oxfordshire, UK). This ability of the pattern recognition methods open the possibility to apply the same results and works on ITER data as well.

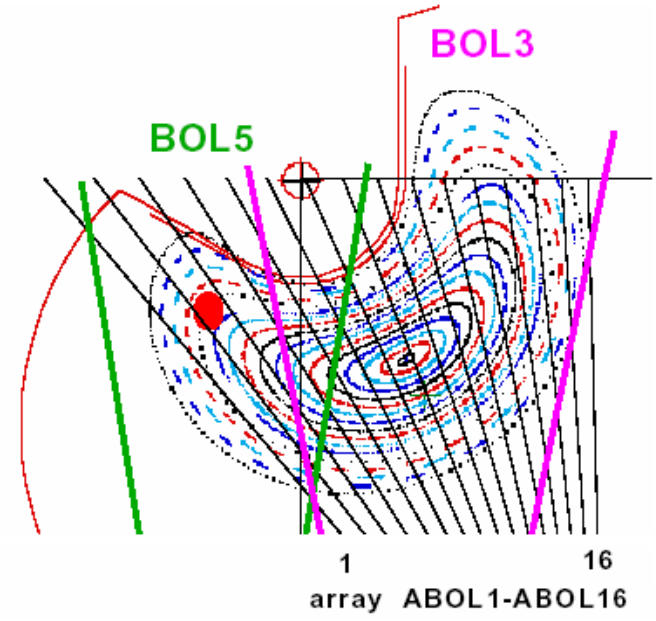


Figure 2.2: Diagram of sensors arrays in the bean-shaped magnetic surface.

2.1.1 Temporal Series Diagnostics

Table 2.1 shows some temporal signals of the TJ-II database. Each of them describes a particular measurement of a physical characteristic of the plasma. For instance, a combination of the bolometer and X-ray systems can be used to characterize the temporal evolution of the plasma density. The data that these sensors provide are time-series values, where one of the coordinates is time, and the other coordinate corresponds to the amplitude. These signals can be made up of millions of samples.

Table 2.1: Some temporal signal classes acquired in TJ-II.

Signal class	Description
RX306	Soft X-ray
ACTON275	Spectroscopic signal (CV)
HALFAC3	H_α
DENSIDAD2	Line averaged electron density
BOL5	Bolometer signal
ECE7	Electron cyclotron emission

Figure 2.3 shows some temporal signals from the TJ-II diagnostics ACTON275, BOL5, Densidad2, and ECE7 respectively for the discharge N^o 10108. Note that the time is given in milliseconds.

Although it can be found representative patterns of each class of signal, the signals of a class are not similar for different discharges. This leads to the study of subclasses

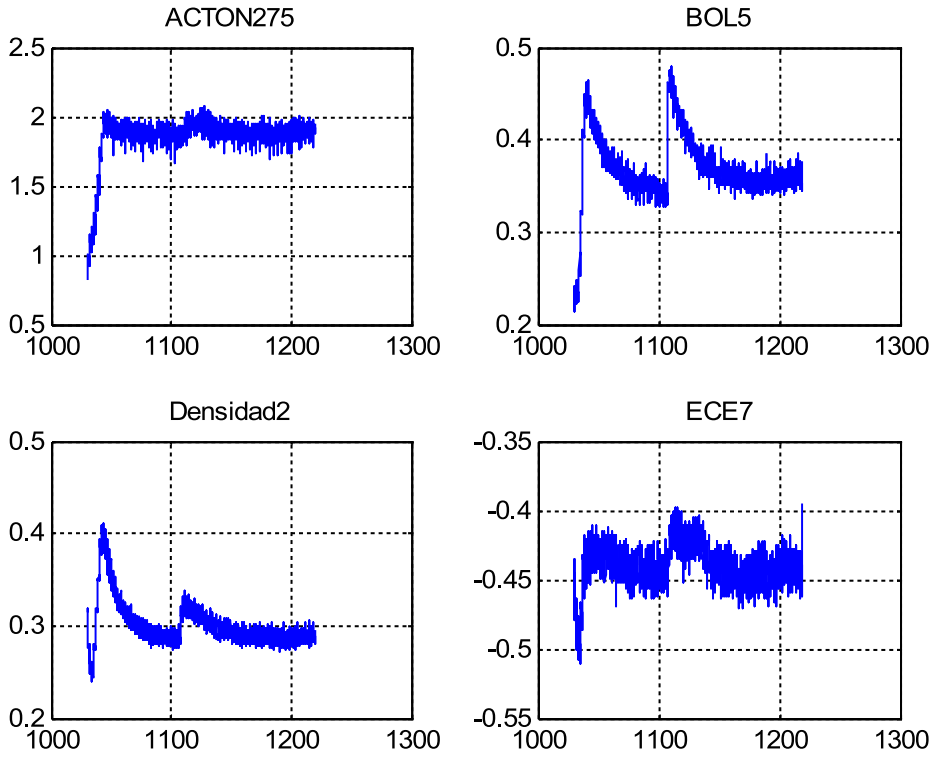


Figure 2.3: Temporal signals per discharge N^o 10108.

inside each signal class could indicate the existence of different physical behavior. More details about the study of subclasses on diagnostics can be found in 3.1.3.

2.1.2 TJ-II Thomson Scattering Diagnostic

The Thomson Scattering (TS) Diagnostic in plasma consists in the re-emission of incident radiation (from very powerful lasers) by free electrons. Electron velocity distribution generates a spectral broadening of the scattered light (by Doppler effect) related to the electronic temperature. The total number of scattered photons is proportional to the electronic density. Figure 2.4 shows the diagram of the Thomson Scattering Diagnostic implemented in the TJ-II.

Every laser shot produces an image (a bidimensional data) from which it is possible to obtain radial profiles of temperature and density. Only a restricted number of pattern images appear in the TJ-II. Each image has (385x576) pixels, *i.e.* 221760 possible attributes.

The images represent different physical situations related to either the plasma heating or the system calibration. The TJ-II Thomson Scattering diagnostic (Farias et al. 2005,

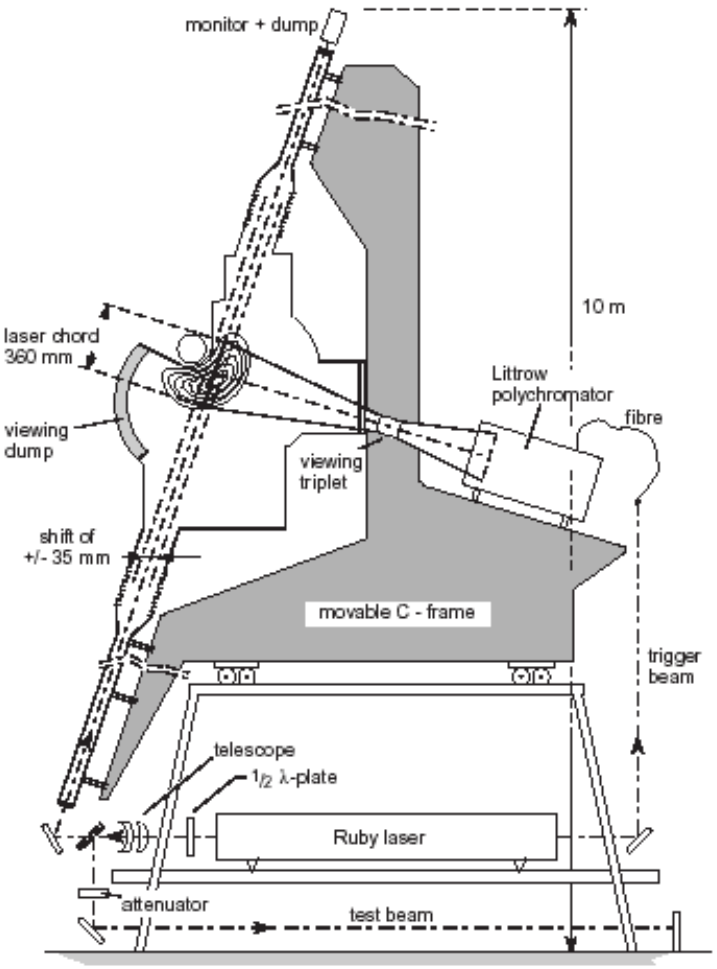


Figure 2.4: Diagram of the TJ-II Thomson Scattering diagnostic.

Makili et al. 2010) collects five different types of 2D spectra (see Figure 2.5): CCD camera background (BKG), measurement of stray light without plasma or in a collapsed discharge (STR), images during ECH phase (ECH), during NBI phase (NBI) and after reaching the cut-off density during ECH heating (COFF). From the point of view of plasma physics, the most important images are ECH and NBI because they correspond to high temperature plasmas. In both cases, the image is processed to obtain the radial profiles of the electron density and electron temperature.

Stray-light on TJ-II Thomson Scattering Diagnostic

The CCD camera collects frequently images corrupted with noise. In Thomson Scattering diagnostic the main source of noise is the so-called stray-light. Controlling stray light has always been important in optical design (Breault 1995). Caused by phenomena such as Fresnel reflection from lens surfaces, air bubbles in glass, dust, diffraction at aperture

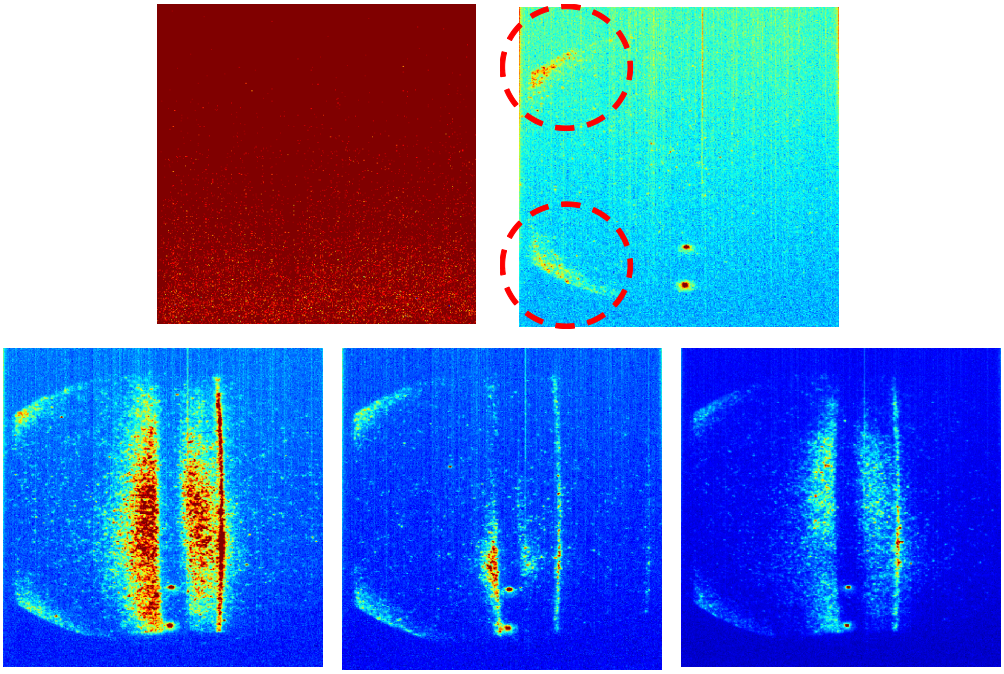


Figure 2.5: Classes of images acquired by the TJ-II Thomson Scattering: BKG, STR at the top, and NBI, COFF and ECH at the bottom. Red circles on the STR image show the noise of stray-light. Note that noise appears on all classes except BKG.

edges, and numerous other effects, its presence frequently degrades both image contrast and measurement accuracy. In particular, the CCD camera in the TJ-II Thomson Scattering diagnostic acquires images corrupted with stray light that, in some cases, can produce unreliable profiles (see Figure 2.5). Therefore the application of approaches to reduce this disturbance will increase the accuracy of the Thomson Scattering analysis.

2.2 Physical Plasma Events

Plasma behavior on fusion devices is not easily predicted, in fact, there is a great effort to understand how to control and stabilized the plasma during a discharge (Schuller 1999). However, there are well known plasma events when the fusion reactor is working. This Thesis has considered two important events: transition from Low to High confinement modes (L-H transitions) and Edge localized modes (ELMs). For both types the detection and localization have been considered. Figure 2.6 shows both phenomena in the Dα emission signal.

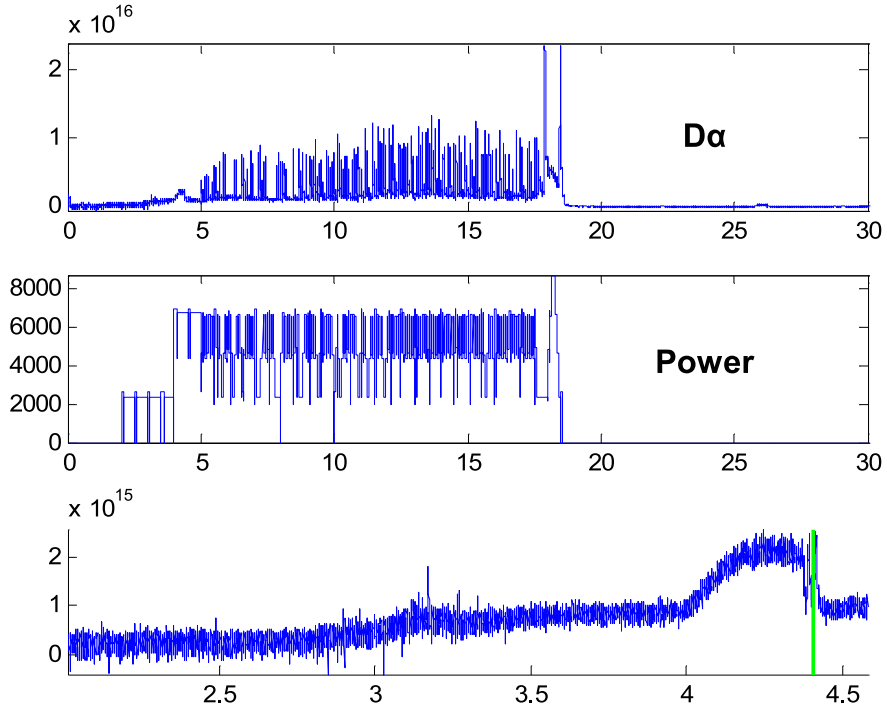


Figure 2.6: Upper plot shows a Dα emission signal where the L-H transition takes place around 4.5s and ELMs appear in the range [5,18]s approximately. Medium plot shows the power injected to heat the plasma (specifically PINJ). Lower plot shows an enlarged section of first plot where the L-H transition occurs (marked by the green line).

2.2.1 L-mode to H-mode transitions

The H-mode (high-energy mode) is one of the main confinement regimes in present and future tokamaks and stellerators. The H mode was firstly detected in the ASDEX tokamak (Wagner et al. 1982). The sudden variation of the plasma parameters from the L-mode (low energy mode) to the H-mode is known as L-H transition. The L-H transition is characterized by the creation of an Edge Transport Barrier (ETB). When the ETB is lost, the plasma returns to L-mode. This transition is known as the H-L transition.

A L-H transition can be identified as a fast drop of the Dα emission between the start of the NBI heating system and the first type I ELM of the pulse. Upper plot of Figure 2.6 shows a Dα emission signal for a particular discharge. The duration of the discharge is approximately 16 seconds. In the lower plot it can be noticed that the L-H transition (fast drop) occurs about at 4.5 seconds, and the first type I ELM is located at around 5 seconds.

In this Thesis, pattern recognition techniques have been applied to predict L-H transitions on DIII-D tokamak databases. To this purpose a multi-layer predictor was

developed, which used data around L-H transitions (previously located by experts) to train a system capable of recognizing L-H transitions in new discharges (Farias et al. 2012). Similar approaches have also been used in other works for L-H and H-L transition estimation on JET tokamak (González et al. 2012, Vega et al. 2009).

2.2.2 Edge localized modes

Plasma instabilities should be successfully controlled in order to produce fusion energy efficiently and without compromising the material boundary. Edge localized modes are one of these instabilities that are not fully known and further theoretical and experimental analysis are required.

ELMs can be observed in the plasma edge as repetitive *peaks*, e.g. in light intensity or in voltage measured at an electric probe. The development of edge-localized modes poses a major challenge in magnetic fusion research with tokamaks, as these instabilities can damage wall components, particularly divertor plates, due to their extremely high energy transfer rate.

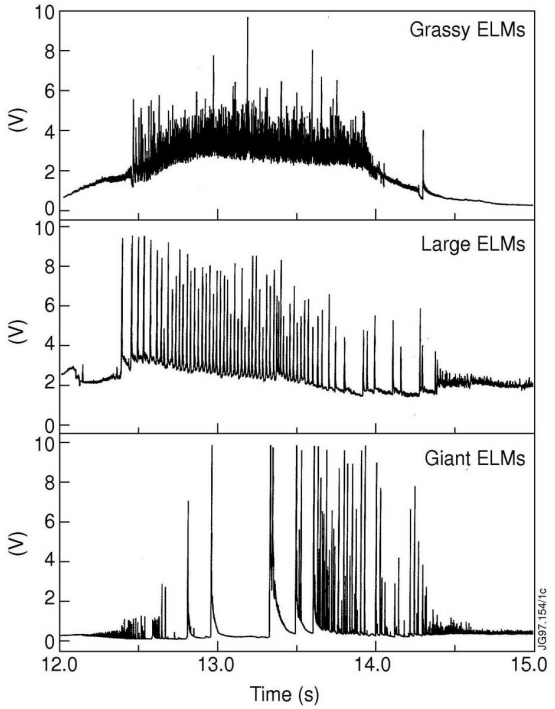


Figure 2.7: ELMs can be observed in the plasma edge as repetitive peaks, plots show the three type of ELMs. Image taken from (EFDA 2013*b*).

A way to examine ELMs is to study the global behaviour of the plasma during ELMs (Liang 2011, EFDA 2013*b*). While some of the features are common to all ELMs,

there are also distinctive differences such as frequency and amplitude. Consequently three types of ELMs have been defined. **Type I** ELMs: The Dα radiation shows large isolated bursts and, therefore, Type I ELMs are also called *large* or even *giant* ELMs. **Type II** ELMs: These are observed only in strongly-shaped plasmas, i.e. with high elongation and triangularity of plasma cross-section. The magnitude of the ELM bursts is lower and the frequency is higher than that of type I ELMs. Type II ELMs are normally called *grassy* ELMs. **Type III** ELMs: The bursts are small and frequent. Type III ELMs are also called *small* ELMs. The ELMs repetition frequency is found to decrease with the increasing heating power. More details about features of ELMs can be found in (Liang 2011, EFDA 2013*b*, Saibene et al. 2002, Bellizio et al. 2011). Figure 2.7 shows a plot for each type of ELM.

Chapter 3

Pattern Recognition on Fusion

3.1 Classification and Clustering

The aim of objects classification is to find a rule, based on external observations or training elements, that allows assigning each object to anyone of several possible classes. There are two big stages to implement in a classification process: features extraction and objects sorting (Duda et al. 2001, Santos & Farias 2010). The first one consists of performing some pre-processing on the objects trying to extract specific differentiating features. The second stage groups the objects into a set of classes.

Note that the approach *one versus the rest* allows to build multi-class classifiers by using a set of binary classifiers. Thus, each classifier is trained to separate one class from the rest, and to combine them by doing the multi-class classification according to the maximal output before applying the *sign* function.

On the contrary to supervised classification, clustering tries to group data into clusters but without the knowledge of how many groups or classes really exist. This approach is quite useful in pattern recognition to reveal the existence of similar patterns into databases, which could indicate that certain conditions or behavior is repeated. In order to group data a similarity criteria or distance is required. In fusion databases clustering can be used to discover the presence of similar plasma's physics.

3.1.1 Classification of Thomson Scattering Images

As it was said before, Thomson Scattering images represent different physical situations related to either the plasma heating or the system calibration. Depending on the pattern

obtained, data are processed in different ways. Therefore, to perform an automated data analysis, a computerized classification system must provide the kind of pattern obtained in order to execute the proper analysis routines. Thus, several classification systems were implemented by using mainly two automatic learning approaches: Neuronal Networks (Farias et al. 2005) and Support Vector Machines (Makili et al. 2010).

Support vector machines (SVM) is a universal constructive learning procedure based on the statistical learning theory (Vapnik 1999, Sebald & Bucklew 2000, Schölkopf & Smola 2001, Hearst et al. 1998, Cherkassky & Mulier 2007, Weston et al. 2001). The SVM maps input data into a high-dimensional space using a non-linear function. Once input data are mapped into the high dimensional space, linear functions with constraints on complexity (i.e., hyper-planes) are used to discriminate the inputs. A quadratic optimization problem must then be solved to determine the parameters of these functions. However, for high-dimensional feature spaces, the large number of parameters makes this problem intractable. For this reason, duality theory of optimization is used in SVM to make the parameter estimation in the high-dimensional feature space computationally affordable. The linear approximation function corresponding to the solution of the dual problem is given in the kernel representation, $k(x, x')$, and it is called the optimal separating hyperplane. The solution in the kernel representation is written as a weighted sum of the support vectors. Figure 3.1 shows the SVM method, the data points at the margin (indicated in grey) are the support vectors.

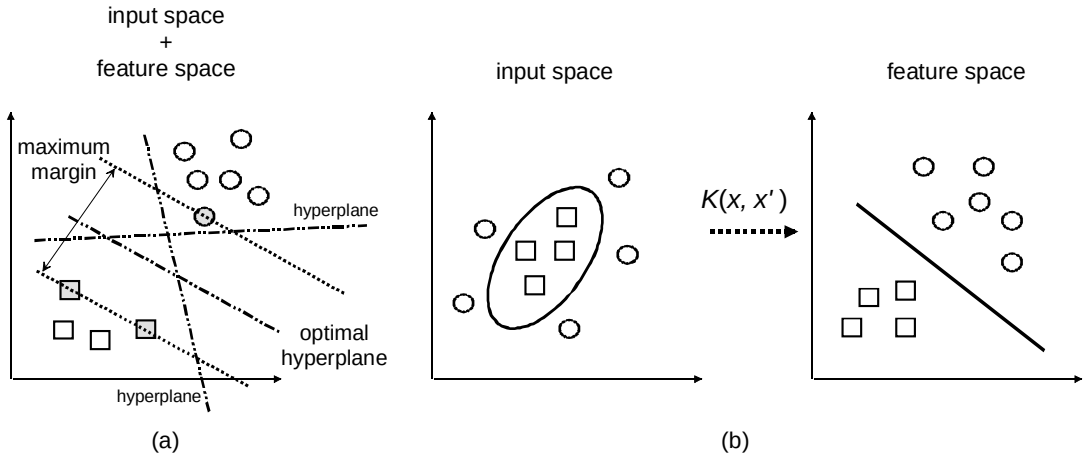


Figure 3.1: The idea of SVMs: mapping the training data into a higher-dimensional feature space via K, and construct a separating hyperplane with maximum range there. This yields a nonlinear decision boundary in the input space. By the use of kernel functions, it is possible to compute the separating hyperplane without explicitly carrying out the mapping into the feature space. (a) Linearly separable case. (b) Non-linearly separable case.

Neuronal networks (NN) have been successfully applied in a great number of classification problems (Duda et al. 2001, Farias et al. 2010, Farias & Santos 2007). There are many types of NN with different structure that can be applied depending on the application characteristics. In any case, a NN consists of some basic processing elements called neurons, which are grouped in layers and connected by synapses connections that are weighted by a factor (Freeman & Skapura 1991, Haykin 2004, Rojas 1996, Hilera & Martínez 1995).

In both classifiers, Wavelet transform (WT) was used to perform feature extraction. The Wavelet transform (Daubechies 1992, Mallat 2008, Misiti et al. 2004) has been used extensively in this Thesis to reduce the dimensionality of signals without a significant loss of information. Some examples of application of wavelets on fusion databases can be found in (Farias et al. 2004, Dormido-Canto et al. 2004, 2005, Farias & Santos 2005).

Analysis of bi-dimensional signals shows great improvements by using Wavelet based methods. Due to the fact that the WT decomposition is multi-scale, images can be characterized for a set of approximation coefficients and three sets of detailed coefficients (horizontal, vertical and diagonals). The approximation coefficients represent coarse image information (they contain the most part of the images energy), whereas the details are close to zero, but the information they represent can be relevant in a particular context.

In relation to the Thomson Scattering images, it has been found (Farias et al. 2004) that the best coefficient to apply the Wavelet transform is the vertical detail, when the selected mother wavelet is Haar at level 4. When applying the mentioned Wavelet transform to the signals of the Thomson Scattering diagnostic, the attributes are reduced from 221.760 to 900 pixels. So, the obtained attributes with the Wavelet transform represent the 0.4% of the complete original image.

Regarding to the classifier with neuronal network (WT+NN), a Feed Forward scheme has been used since it has shown to be successful for also other classification problems (Farias & Santos 2007, Farias et al. 2010). One of the possibilities of this type of neuronal network is to use it for the supervised learning, where it is necessary to train the neuronal network indicating to the input layer the attributes of a signal (the wavelet transform of the original image) and the desired values to the output layer (the class of the original image).

Figure 3.2 shows the neuronal network that generated the best results. Note that the neuronal network has an input layer of 900 attributes which come from the previous processing stage (generated by Wavelet transform). The hidden layer uses 90 nodes with functions of activation Tansig, whereas the output layer has 5 neurons with functions of activation Logsig. After the training of the neuronal network, every signal is associated with its class through the activation from its output neuron and resetting the remaining ones. This classifier shows an average percentage of hits of 90.89%.

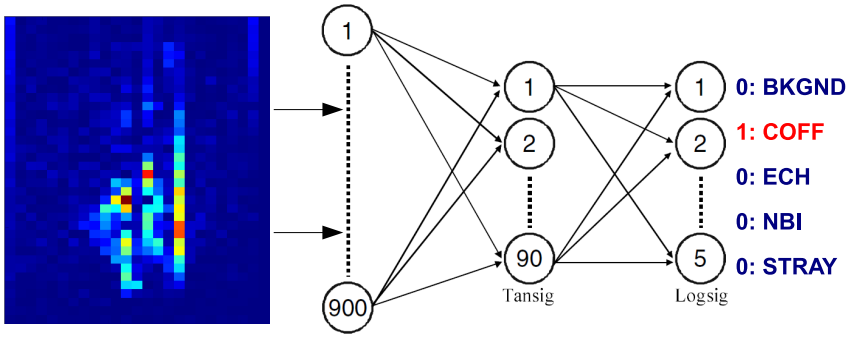


Figure 3.2: WT+NN classifier scheme and structure of the proposed neuronal network.

Regarding to the classifier with support vector machines (WT+SVM), a linear kernel has been used with satisfactory results, although the selection of other kernels improve the classification rate.

The WT+SVM technique is very robust and success rate was 92.7% in the last TJ-II experimental campaign (over a 98% in earlier campaigns). The classifier did not know to assign a class in a 5.5% of the cases and the rate of wrong classifications was 1.8%. This technique has shown a great robustness than other techniques, which were based on statistics properties of Thomson Scattering images.

3.1.2 Classification of Plasma Diagnostics and Configurations

Since the fusion plasma experiment generates hundreds of signals, it is essential to have automatic mechanisms for searching similarities and retrieving of specific data in the wave form database.

Similar to previous Thomson Scattering classifiers, Wavelet transform was applied in order to mapping signals to spaces of lower dimensionality. Besides, support vector machine is a very effective method for general purpose pattern recognition, specifically for multi-classification case. The combination WT+SVM has been proposed for searching

and retrieving similar wave forms in TJ-II databases (Dormido-Canto et al. 2004, 2005), whereas the selection of SVM and *a priori* knowledge of the geometrical parameters of the plasma can be used to automatically identify the plasma configuration of discharges in JET (Dormido-Canto et al. 2008*a*).

Classification of Temporal Series Diagnostics in TJ-II

A proof of these approaches was based on classifying and recognizing temporal evolution signals from the TJ-II database. As before, it is accomplished in a two-step process. The first step provides signal conditioning (to ensure the same sampling period, number of samples, etc) and the reduction of the signal dimensionality with the Wavelet transform (approximation coefficient with wavelet mother Haar at level 8). The second step is performed by using support vector machines with different kernel functions.

Figure 3.3 displays the positive support vectors for 4 classes (ECE7, BOL5, RX306, and Densidad2) using a linear kernel, the training signal corresponding to the original signal in TJ-II (gray line), and the wavelet approach which is the signal resampled to 16384 samples after the wavelet transform (black line).

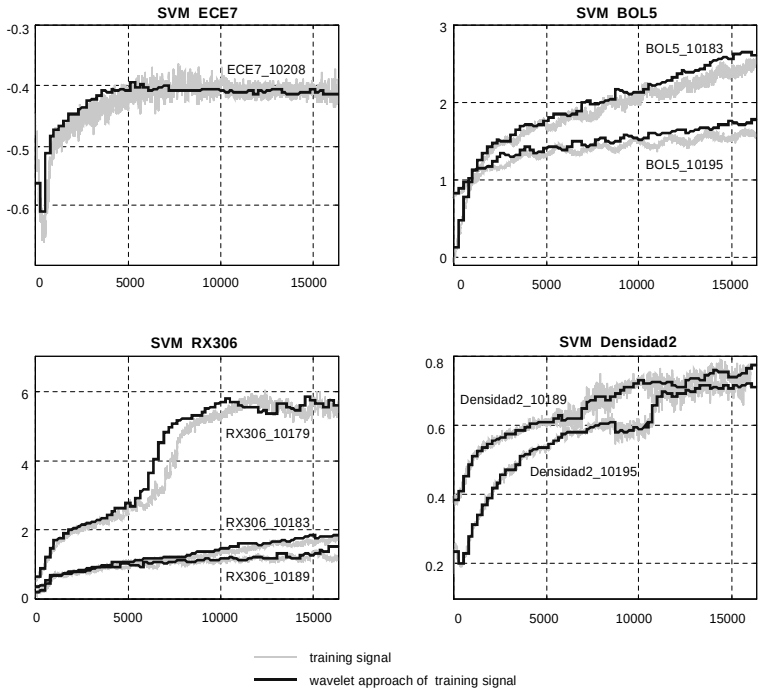


Figure 3.3: Support vectors and wavelet approximation coefficients of four time evolution signals in TJ-II.

Two multi-classifiers have been built for 4 and 6 different classes with 100% (radial

basis function kernel) and 92% (exponential radial basis function kernel) of hits as best results, respectively (Dormido-Canto et al. 2004, 2005).

Classification of Plasma Configuration in JET

Feature extraction of fusion databases is difficult to be implemented since there is not a generic solution. Normally, the reduction of dimensionality of a signal, such as the use of Wavelet transform, is a good first attempt to perform pattern recognition in fusion. However, best option is to take advantage of previous or *a priori* known information. This was the case of the plasma configurations in JET, where the boundary of the last flux surface can be used to identify the plasma configuration of discharges.

The shape of the last flux surface is an essential ingredient in the definition of the JET operation scenarios and several ones can be present during a discharge. Some kinds of data analysis are sensitive to the plasma configuration (for example, to the location of the X-point and strike points) and, therefore, proper identification (classification) of the plasma configuration is important.

JET configurations were primarily identified by referring to an identifying keyword describing the request made, prior to the pulse, to the plasma control system. This has the disadvantage of being nonspecific, as several different identifiers can refer to the same configuration; cumbersome, as this data cannot be accessed automatically; incomplete, as some discharges are not assigned an identifier; and potentially wrong, as the identifier describes the intended rather than the resulting configuration. These problems motivated the development of an automatic classifier.

Results with two classification systems based on geometrical parameters of the last flux surface and SVM were both successful (Dormido-Canto et al. 2008*a*).

A first classifier was implemented to discriminate discharges belonging to the following three classes: VH_3M5_HT, HIXR_GB, and SEPTUM. A two-dimensional feature vectors and a linear kernel function was enough to classify the 100% of the 102 tested configuration. A second classifier was also implemented for the automatic recognition of 8 classes. Since this a more difficult problem than before, the feature vector was expanded to incorporate twelve geometrical parameters. The results of the classifier was also encouraged, because the hit rate of the tested configurations was over 96% in average for the 8 classes.

3.1.3 Clustering of Temporal Series Diagnostics

Diagnostics provide temporal evolution signals that translate plasma physical properties. Thus, it can be assumed that similar signals correspond to similar plasma behaviors and, therefore, it is possible to find representative patterns for the same physical conditions.

The plasma's physics can be described by the many kinds of acquired signals (density, temperature, soft X-rays, bolometry, etc.) of each pulse. Thus, a method for finding similar waveforms for each kind of signal would be very helpful to reveal, in an automatic way, the set of discharges that show comparable behaviors.

The goal of this problem is then to classify waveforms into a number of categories (or clusters) and to apply proximity measures to evaluate the similarity between the acquired signals. The clustering method is responsible for r̈evealingẗhe organization of signals into s̈imilarg̈roups. This is so-called unsupervised classification or simply clustering.

Clustering of Diagnostics

The article (Duro et al. 2006) shows the experience of applying clustering to TJ-II databases. In this work it was selected waveforms of 194 discharges corresponding to *Hα emission*, *line average electron density*, *bolometer* and *soft X-ray* signals. All the signals were pre-processed in order to be able to analyse the data within the same time window (258 ms) with identical sampling period (10 μs). Before to apply the clustering, the 4 signals have been processed by two feature extractors: Wavelet (Haar at level 8, 64 approximation coefficients) and Fourier (24 first coefficients) transforms.

Four techniques have been applied to perform clustering: *Hierarchical*, *K-means*, *Adaptive resonance theory (ART)*, and *Gran Tour (GT)*.

Hierarchical (Johnson 1967) starts by assigning each item to a cluster, then merges most similar pair of clusters, reducing one cluster. The process continuous until get one final cluster.

K-means (MacQueen et al. 1967) begins from a fixed number of clusters, them assign each item to the group that has the closest centroid, when all items have been assigned, the centroids are recalculated. The process ends until all centroids are not modified.

Adaptive resonance theory (Carpenter & Grossberg 1993) is applied to artificial neural network to develop a kind of competitive learning neural net. In this case when

the information is presented to the input just one output neuron is activated. The idea is to resonate the input information with prototypes of classes those the net recognizes.

The Grand tour of a multi-dimensional data set is an interactive visualization technique for examining structure of high dimensional data using dynamic graphics. The idea is to project the n-dimensional data to a plane and to rotate the plane through all possible angles, searching for "structure" in the data. Structure is defined to be departure from normality (Martinez & Martinez 2001) and includes such things as clusters, linear structures, holes and outliers.

Each clustering method (hierarchical, K-means, ART and GT) gives a set of clusters. However, it was only considered clusters that included at least a 5% of the waveforms. Clusters with less than 5% are grouped together in a miscellaneous cluster.

The results analyzed the number of clusters found and the percentage of signals included in each one. First of all, it must be pointed out that equivalent results are obtained without feature extraction. In particular, using Hierarchical, ART and GT the percentage obtained are very similar, not only in the main clusters but also in the miscellaneous cluster. It can be noted that at least the 50% of signals belong to the same cluster. The inspection of the K-means results showed that it produces a different behavior: more number of clusters is generated. Besides, the number of signals in each cluster is smaller. Analyzing the signals that constitute these clusters it can be concluded that the signals for two or three clusters (depending on the experience) in the K-means method are integrated into a bigger cluster for the other three methods.

The several methods group the same signals into the same clusters, independently on features. Roughly speaking, all families provide two clusters. Firstly, the big one symbolizes that most signals translate an average physical behavior of the measured plasma property. Secondly, the rest of the waveforms can be integrated into a single cluster. The latter includes non-average behaviors and, therefore, signals classified in this group reveal non-standard plasma properties.This fact helps diagnosticians because they can find, in an automated way, interesting data to be analyses, instead of having to search for them manually. See more details of clustering in TJ-II databases in (Duro et al. 2006).

Clustering and Modeling of Diagnostics

The article (Martín et al. 2009) shows also the experience of applying clustering to TJ-II databases. The work considers the signals described in Table 2.1. The clustering technique proposed is based on a partitioning method. The strategy consists of generating a triangular matrix with the values of a mathematical measurement of the similarity, i.e., the normalized scalar product (NSP), between each couple of waveforms and the application of a threshold to generate dynamic clusters based on it. From other works in which it was used these signals (Dormido-Canto et al. 2004, 2006, Duro et al. 2006, Farias et al. 2006, Martín et al. 2007), it can be derived that the most efficient procedure for the real-time measurement of similarity of the TJ-II plasma fusion signals is the NSP. Thus, there is not extraction of discriminatory features as it is only based on the NSP.

The information provided by the clustering method can be also used to obtain a concise and representative model of each class of plasma signals by applying different modeling approaches. In that way, the expected patterns of each group are obtained, and they make it possible to detect anomalies or unexpected physical events. Neuro-fuzzy identification and time domain approaches have been used for modeling purposes.

The result of applying this procedure as a clustering method is a set of different groups. For each of these groups, a model will be generated by means of fuzzy inference systems (FIS)(Jang 1993), and using this model, it will be possible to detect unexpected events.

From some experiments, it can be concluded that there is always a stable group of waveforms where at least 75% of the signals are included. There are also some other clusters with fewer signals. If there are only one or two signals in a group, then that may mean that those signals are outliers. In this sense, this clustering method allows the detection of anomalies in an immediate way.

The goal of the neuro-fuzzy modeling and time-domain strategies is to identify natural groups of data from a large data set to produce a concise representation of a type of signal. It can be seen as a pattern with which a new signal will be compared to classify it. On the other hand, these models will help in the searching and retrieval of similar signals, as each model represents a cluster, and, therefore, the searching space in which similar signals are more likely to be found is reduced. More details about results and experiment carried out by this work can be found in (Martín et al. 2009).

3.1.4 Classification and Clustering of ELMs

Edge localised modes (ELMs) are plasma instabilities that can affect the material boundary. Although many advances have been done from theoretical and experimental point of views, ELMs are still not fully understand (see Section 2.2.2 for more details). In order to advance in the study of the physics behind ELMs, data-driven methods seem to be powerful techniques to extract knowledge from the experimental signals. This knowledge could be combined with theoretical models for both exploratory and confirmatory analysis.

In (Duro et al. 2009) it was developed a data-driven approach for the characterization and automatic classification of ELMs as type I or type III (Liang 2011, EFDA 2013*b*, Saibene et al. 2002, Bellizio et al. 2011). To this end, three steps are accomplished. The first one is to identify, isolate and extract individual ELMs from JET signals (each individual ELM is analysed instead of a temporal segment containing many ELMs). Second step is a feature extraction process to represent the ELMs with a minimum set of relevant characteristics. Finally, three classifications methods (supervised and unsupervised) have been applied to classify the ELMs.

The ELM recognition and isolation is carried out using three signals: stored diamagnetic energy (corresponding to the JET signal MG3F/WPD), line integrated electron density (JET signal KG1V/ LID4) and Dα (JET signal S3AD/ AD34). ELMs are recognized by an abrupt change in the diamagnetic energy and a simultaneous drop in the line integrated electron density. As a consequence of the ELM instability, a typical peaked shape appears in the Dα. See similar peaked shapes in Figures 2.6 and 2.7.

Regarding to the feature extraction process, from visual inspection point three attributes have been considered: the drop from the ELM peak, the period of each individual ELM, and the crest measure (which is a custom feature to measure the shape of ELMs). The first two features are calculated from diamagnetic energy signals, while the latter attribute is computed from the Dα emission signal.

After feature extraction, the clustering and classification techniques must group the ELMs into two subsets: type I and type III. Training data set is composed by 122 individual ELMs (97 of type I and 25 of type III). The test set is made up of 143 ELMs isolated from JET signals. The supervised method was implemented with support vector machines. To perform the clustering (or unsupervised classification), the k-means and

hierarchical techniques have been selected.

Similar results are obtained when using any of the other classification methods, either supervised or unsupervised. In both cases the number of classes that provides the vast classification performance is two. Moreover, the success rates (over 93% in most cases) with different techniques allow us concluding that the feature selection strategy adopted is quite robust. In particular, using K-means and hierarchical method, the percentage of signals included in each cluster is the same.

Results also show that ELMs of type I have bigger drop than those of type III but this characteristic is not enough to definitely differentiate between them. Moreover, the period is also not discriminant feature for the classification process. However classification depends strongly on the crest measure. More details about this approach can be found in (Duro et al. 2009).

3.2 Information Retrieval

Different plasma physical behaviors are shown by the different signals acquired by the diagnostics. In general, a linear mapping can be established to connect the time evolution of a physical phenomenon with the kind of signal that it generates. Therefore, it is possible to speak about patterns. To analyze plasma properties, pattern search can be very helpful. However, an experimental database of a fusion device contains thousands of signals, so automatic information retrieval methods are required.

3.2.1 Searching for Entire Waveforms

In (Farias et al. 2006), an automated technique to search for and retrieve similar time evolution signals to a reference waveform (input signal) is described.

The procedure is divided in three stages. First, a feature extraction is carried out. After that, a classification system performs a coarse filter to reduce the search space. Finally, similarity query methods are used to retrieve the waveforms that are more similar to the input signal. This technique has been applied to temporal series diagnostics on the TJ-II databases.

Wavelet transform makes possible to reach a desired decomposition level preserving signal information. Redundant information is minimized and the computational load is

substantially cut down. Thus, in the first phase wavelets are used for noise reduction. Besides, signals processed by wavelet transform reduced the samples from 16384 to only to 64. In some previous works (Rafiei & Mendelzon 1998, Nakanishi et al. 2004, 2006), similar waveform recognition methods have been applied by using discrete Fourier transform, but since many fusion waveforms have a non-stationary behavior, using WT seems to be a better option for data characterization.

After feature extraction, the searching process begins. The search procedure implies the two following steps. The first one is aimed to narrow down the search space, i.e. to limit the search of the signals to a proper subset of the database. This is carried out by means of a SVM classifier, which has been previously trained to distinguish among different kind of signals.

Having reduced the search space by the use of the classification system, the last step consists of finding the most similar signals to the reference one. Two different methods can be applied for this purpose: Euclidean distance and bounding envelope. Euclidean distance simply computes the distance sample by sample of the input signal against the rest of database. Bounding envelope method is based on the construction of two bounds around a signal (upper and lower bounds). A distance measurement can be done by counting the number of samples that are outside the bounds. Both similarity query methods can be used to find the minimum distance between the reference signal and the database, and therefore to find the most similar signals.

Experiments showed that bounding envelope method is more robust technique than Euclidean distance. This is due to the accumulated error with Euclidean distance. The bounding envelope method considers more distant points as outliers independently of their values. More details can be found in (Farias et al. 2006).

Another approach to find similar signals have been applied in (Vega et al. 2008). This article describes a technique where, given a waveform, it is possible to retrieve similar ones from large databases in a fast and automated way. The method uses an input signal to look for the most similar waveforms within the database. The technique is based on the development of a tree-structure classification system that groups waveforms into clusters in accordance to certain rules (each node is a cluster). Waveform clustering is the essential element to speed up the searching and to save computational resources. The searching process is carried out by means of a one by one comparison method but

only within those waveforms inside a cluster, rather than by computing the similarity between all database signals.

The searching process of the most similar signals is carried out in four steps. Given a waveform, the first step performs feature extraction. The second one is the classification of the feature vector into one of the existing clusters. The third step is the computation of the similarity factor between the input feature vector and the rest of the cluster feature vectors. Finally, waveforms are sorted according to their similarity measurement in descending order.

The absolute value of the normalized inner product was used as similarity function. Note that similarity between waveforms does not imply that they are almost equal (similarity near 1). Thus, some advantages of this method over Euclidean or bounding envelope come from the fact that the method does not depend on either amplification gains (i.e. waveforms whose difference is a gain factor are recognized as equal signals), or signal polarity (i.e. inverted waveforms are perceived as equal signals). The method finds the most similar signals but the similarity factor can be low (close to 0). In other words, the technique can get a list of signals from the database although the waveforms do not resemble between them.

Results of this approach show that the system is able to find similar waveforms (bolometry and soft x-ray signals) in a couple of seconds from a total of hundreds of waveforms. More details can be found in (Vega et al. 2008).

3.2.2 Searching for Pattern Within Plasma Waveforms

Visual data analysis is an essential tool in plasma physics. A simple visual inspection of signals is enough to recognize a typical plasma evolution or to distinguish the presence of interesting events. A researcher identifies the plasma behavior through the recognition of patterns inside wave forms: bumps, unexpected amplitude changes, abrupt peaks, or sinusoidal components. Therefore, a big challenge in data access is the creation of fast ways to look for patterns within waveforms.

There are some previous works on pattern recognition in fusion databases. In the approaches shown in Section 3.2.1, efforts were concentrated in looking for similar full wave forms, i.e., signals covering the full plasma life. The pioneer work described in (Nakanishi et al. 2006) is centered in searching for patterns within wave forms, which

was based on one major frequency component. However, more general methods are required to look for general patterns in non-stationary waveforms such as most of the temporal fusion signals.

The searching for patterns within waveforms has been considered in the articles (Dormido-Canto et al. 2006, 2008*b*, Rattá et al. 2008, Vega et al. 2007). To this purpose the syntactic approach was used. The syntactic approach takes the view that a pattern is composed of simpler subpatterns (Fu & Albus 1982). The most elementary subpatterns are known as primitives. A complex pattern is then expressed in terms of relationships between its primitives. An analogy between the structures of patterns and the theory of formal languages is used to establish the foundation of syntactic pattern recognition. The patterns represent the sentences in a language, while the primitives constitute the alphabet of the language. A grammar of the language generates and identifies sentences belonging to that language by employing its rules. The idea that a potentially large set of related complex patterns can be described just by a finite number of primitives and grammatical rules makes this approach appealing. However, sometimes a grammar is not suitable for a pattern class description because the patterns under consideration lack regularities and cannot be described by rules. In such cases, the structural approach can be adopted.

In structural pattern recognition, primitives are represented by strings. Consequently, the recognition problem turns into a pattern-matching problem. For example, given a pattern decomposed into primitives (set of characters: string), the final goal is to find the most similar pattern from a database of strings. Figure 3.4 depicts primitive labels and a coded waveform. The classification of the angle gives all the elementary structural information needed to construct more complex subpatterns in wave form recognition. Note that the waveforms are divided in segments of fixed length.

The types of searching mentioned before can be applied with a relational database management system using Structure Query Language (SQL). For experimental purposes, the database was implemented with MICROSOFT ACCESS^{TM} (Feddema 2001). The information retrieval algorithm is performed as follows: First the user selects a shot, then chooses a section of the signal (pattern), and asks for the application of a type of searching. The MATLAB application carries out the preprocessing and primitive computation. After that, an SQL query is carried out with regard to the searching type

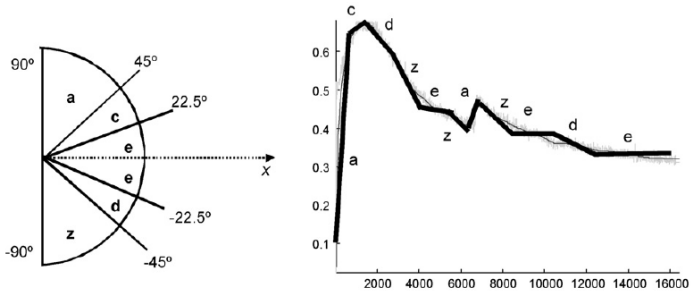


Figure 3.4: Waveform coded by primitives: Left plot shows primitives and labels for the classification of the angle of the signal segment. Right plot shows a coded waveform with primitives.

selected. Finally, ACCESS sends back to MATLAB the SQL results, and the application shows all matches in the returned signals. Figure 3.5 shows an example of this string matching approach to search patterns within plasma waveforms.

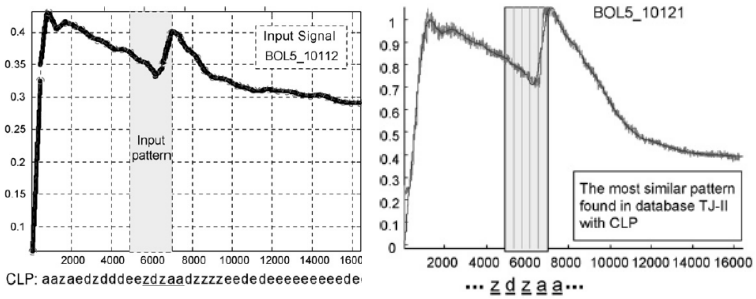


Figure 3.5: Example of information retrieval with structural pattern recognition.

More details and variants of this approach applied to TJ-II and JET databases can be found in (Dormido-Canto et al. 2006, 2008*b*, Vega et al. 2007, Rattá et al. 2008).

3.2.3 Detection of L-H Transition Times

Machine learning methods have been developed to automatically determine the time instants of L/H transitions in the DIII-D tokamak (Farias et al. 2012). A training dataset is used to generate a non-parametric model to distinguish between the L and H confinement modes at any time of a discharge. The only requirement to create the model is to assume that all samples are independent and they are identically distributed according to a fixed but unknown distribution. The model also provides the uncertainty (error bar) in the prediction of the transition time. To this end, conformal predictors are used. Conformal predictors qualify their predictions with a couple of values, confidence and credibility, that provide information about how accurate and reliable the predictions are (Vovk et al. 2005).

The system is implemented in two steps within a distributed computing environment. Firstly, a multi-layer SVM model is created by using a training dataset. The SVM model uses a combination of several signals to determine the L-H transitions. The selection of the dataset is accomplished in an automatic way by means of a morphological pattern recognition (MPR) technique of the Dα emission signal. The morphological algorithm looks for the fast drop by using only structural information of the waveform of the Dα signal (González et al. 2010). WT and SVM regression techniques are used in the whole process for the MPR algorithm.

Secondly, the SVM multi-layer model and the MPR algorithm are combined to predict separately the L-H transition time of new discharges. Figure 3.6 shows the use of SVM classifier to estimate L-H transition times.

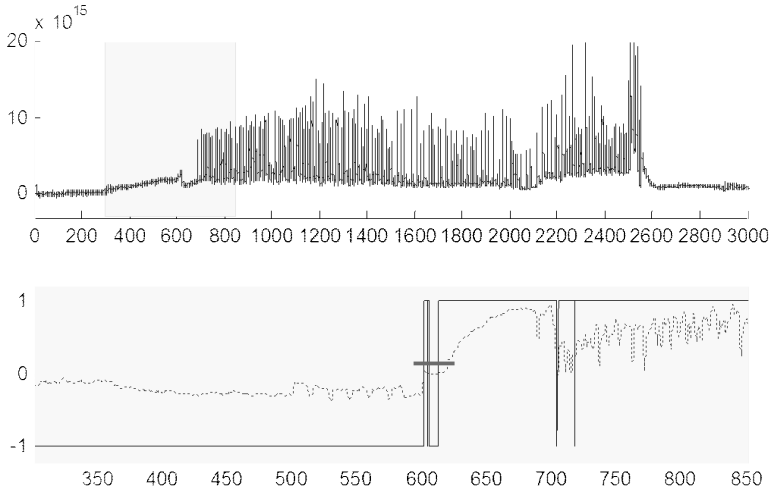


Figure 3.6: Prediction of transition L-H with SVM: Upper plot shows an interval of a Dα emission signal where the L-H transition takes place at 600ms approx. Lower plot depicts the distance to the SVM separating hyperplane and the classification of a SVM model for the interval.

On one hand, the morphologic algorithm takes into account a Dα emission and power signals to predict the transition. On the other hand, the multilayer SVM model performs the prediction. Note that the SVM models put the focus of the searching only on the zone before ELMs start. If the difference of both predictions is less than 100ms then the final prediction is given by MPR, otherwise it is given by SVM. Figure 3.7 shows the scheme used to predict L-H transitions.

The predictor has been tested with the initial 354 discharges. The combined predictor has an average of prediction error of 6 ms and a standard deviation of 49 ms. The successful rate is 95.6% . Figure 3.8 shows the histogram with the frequency of the

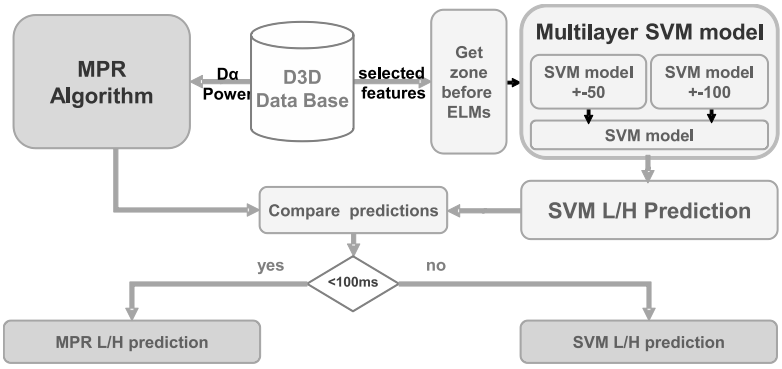


Figure 3.7: Combination of MPR algorithm and SVM models to predict L-H transition times.

prediction error.

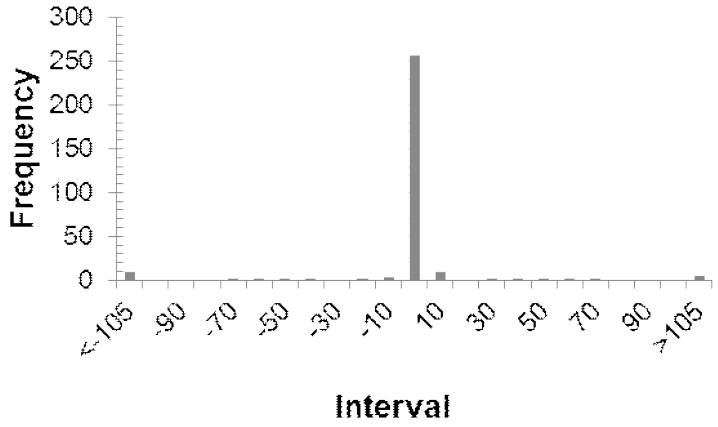


Figure 3.8: Histogram of the prediction error of the MPR + SVM system.

3.3 Noise Reduction

This work has also put the focus on the noise reduction in the TJ-II Thomson Scattering diagnostic. As it was said before, the CCD camera in the Thomson Scattering diagnostic acquires images (spectra of laser light scattered by plasma) corrupted with stray light that, in some cases, can produce unreliable profiles of temperature and density. One example is the light from the ruby laser which reaches the spectrometer, and there is no possibility to distinguish it from the light scattered by the electrons. So far, different hardware techniques have been tried to remove/decrease the stray light contribution but only with partial success. For example, to place a notch filter in front of the spectrometer or inside the spectrometer, or to carry out a correct alignment of the system.

The noise on images can be reduced by applying many classical and advanced techniques such as low pass filters or wavelets. However, in some cases the presence of noise

is not global but located only in particular regions of the image. In these situations the application of global filters over the entire image is not a suitable option since the noise and the information are equally reduced. Alternatives to the *global* techniques come from region segmentation theory.

In recent years, there has been a lot of interest in using results of a generic image segmentation algorithm to obtain pixel-precise object segmentation (Fulkerson et al. 2009, Gu et al. 2009). Segmentation is used to subdivide an image into a set of regions. These regions do not have predefined shapes and their boundaries are irregulars. Therefore the shape and boundary properties of the regions can be used for feature extraction. Another advantage of using image regions is scalability and potential savings in computational efficiency. Image regions usually provide a much smaller set of hypothesis to be examined in comparison to the sliding window approach, and at "natural" scales that are obtained through segmentation. Thus, this work has used the approach based on extraction of regions with connected-components (ERCC) in order to remove some specifics regions associated to the noise.

As it has been said, the ERCC method is based on segmentation theory. Segmentation refers to the process of partitioning a digital image into multiple segments (sets of pixels). More precisely, image segmentation is the process of assigning a label to every pixel in an image, so pixels with the same label share certain visual characteristics. In general, a segmentation of an image is a partition into connected subimages (regions) R_1, R_2,..., R_n such that all regions are disjointed, and the union of all of them makes up the image. Each subimage satisfies a predicate such as all pixels in any subimage R_i must not differ by more than Δ_x gray levels, all pixels in any subimage R_i must be joined by a connectivity factor, etc.

The procedure proposed for noise removal by using connected components is shown in Figure 3.9.

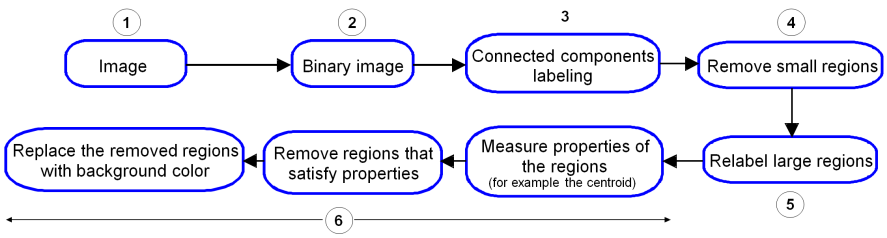


Figure 3.9: Flowchart for extraction regions with connected-components.

The approach starts with the conversion of the original image into a binary version (pixels of 1's and 0's) by using a given threshold. Then, the regions of the binary image are labeled as regions, where 1's belong to the same region if they are connected (i.e. neighbors) in any direction. The next step removes all regions that have fewer than a given number of pixels (small regions), and then the remaining regions are relabeled. After that, all regions whose centroid is less than 100 (i.e. left part of the image) are removed. The eliminated regions are replaced by the pixels' values of the symmetric right-hand half of the image. Finally, small regions removed in the fourth step are restored.

Figure 3.10 shows the result for each stage of the procedure described with an image of type NBI. More details can be found in (Dormido-Canto et al. 2012).

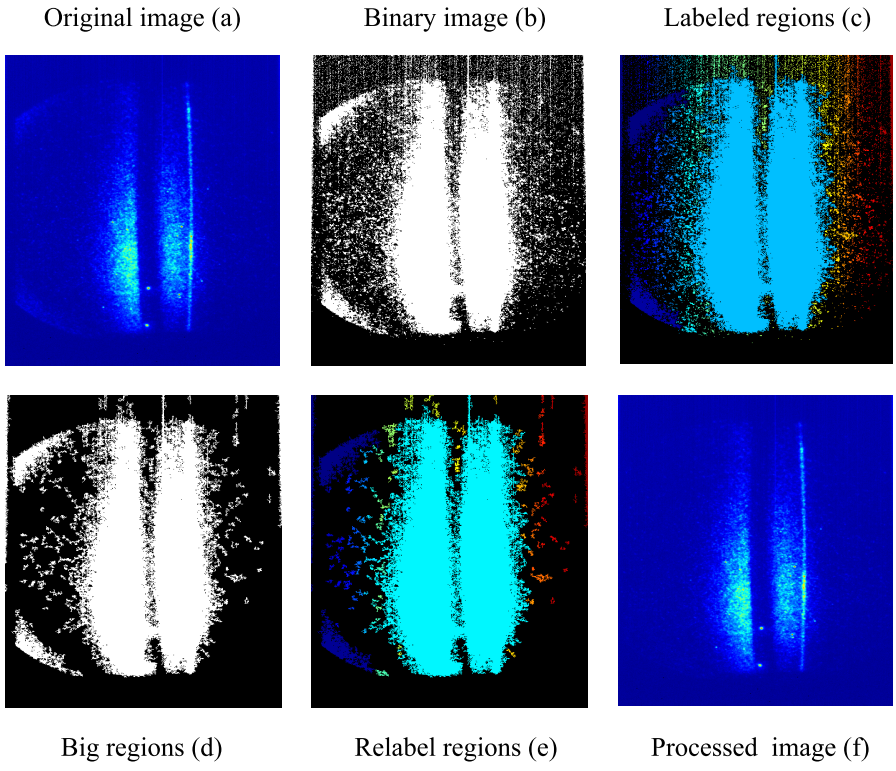


Figure 3.10: Images showing the steps of the ERCC algorithm: (a) original image, (b) binarized image, (c) labeled image, (d) image with regions greater than P pixels, (e) relabeled image, and (f) image with noise removed.

ERCC can reduce significantly the stray-light in TJ-II Thomson Scattering images, however the predicate of connection for a pixel is sometimes too strong. For instance, pixels quite near but not connected to the region are not considered as noise in this approach. This is where region growing (RG) comes in handy. Region growing allows

addition of pixels to a region by using a custom predicate. Thus, a pixel could be considered part of a region although the location of the pixel is near but not connected to the region.

In order to validate how well the noise is reduced in an image, it was defined the *denoised function*. The results of both approaches, ERCC and RG, are shown in Table 3.1.

Table 3.1: Success rate of the denoised function for ERCC and RG Algorithms.

	BKG	STR	NBI	COFF	ECRH
ERCC	98%	96%	91%	97%	94%
RG	98%	97%	95%	97%	96%

The validation process has been tested for 1146 Thomson Scattering images. Note that the results for RG are slightly better than for the ERCC algorithm. In the case of NBI images, the difference is mainly because ERCC is not able to reduce the stray light when the noise is "connected" to the central part of the image (the significant information). More details about the denoised function can be found in (Farias et al. 2013).

Chapter 4

Conclusions and Future Works

4.1 Conclusions

There are many experimental fusion devices that study the process of nuclear fusion. Every experiment produces thousands of signals, with enormous amounts of data. For instance, in JET (the biggest European fusion reactor) every discharge, of about tents of seconds, can generate 10GB of data. ITER (an international nuclear fusion project) could storage until 1 TByte per shot. However, not all data is processed, in fact nowadays only 10% of the data is processed. The rest 90% is not processed at all.

Since the fusion plasma experiment generates very large databases, it is essential to have automatic mechanisms for searching similarities and retrieving specific data in the waveform database.

First applications of pattern recognition and machine learning techniques involved classification and clustering of temporal and images diagnostics.

The development of Thomson Scattering classifier showed the potential of the combination of wavelet transform and support vector machines without the necessity of expert knowledge. On the one hand, wavelets allow to reduce the high dimensionality of TS images. On the other hand, SVM can be used to classify processed images with high hit rates. Although similar results (over 90%) were found with the combination of WT+NN, the structural risk minimization by selecting the optimal hyperplane in SVMs allows a generic decision rules (i.e without over-fitting), which is a desired feature of any learning algorithm. Besides, the training and testing times are normally longer when NN are used. One disadvantage of SVM with respect to NN could be the difficulty to

find a suitable kernel function.

The combination of wavelet transform and support vector machines has been also applied to the classification of temporal signals. From the observation of several experiments, the WT+SVM method is viable and very efficient, and the results present high rate of hits, reaching 100% for a classifier of 4 classes in TJ-II diagnostics. Similar approach was developed to classify JET configurations. Although apparently a simple visual inspection could be enough to discriminate a limited number of different discharges, when a bigger number of categories are considered, it is necessary to resort to a general purpose system as SVM. High success rate in spite of the reduced number of training data should be emphasized. Results with eight classes are promising (over 90% for a 8 classes classifier) even for a future real-time application of the method.

Localization of physical plasma events have been also analyzed in this Thesis. An example of this is the automatic predictor of L-H transition times implemented for DIII-D databases. The system was trained in the CIEMAT supercomputers and the operation was carried out in the DIII-D site. The predictor has been tested with the initial 354 discharges. The combined predictor has an average of error prediction of 6ms and a standard deviation of 49ms. The successful rate is 95.6%.

In this Thesis an approach to classify Edge localized modes was considered as well. The classifier for individual ELM analysis was built. Although the physical basis to this approach is not yet established, the method seems to work on a restricted database of 300 ELMs. In addition, the selected dataset has been classified with very high success rates and very low dimensionality (in fact it can be reduced to a single feature, the crest measure). Supervised and unsupervised methods have been implemented. Both methods group the same signals into the same sets which means that the features selected are robust enough to represent the ELM instability. Although results are promising, it should be noted that the databases is not completely general and a comparison with a more extended database is needed in order to draw definitive conclusions on the approach proposed here.

Unsupervised classification has also applied in fusion. A clustering approach in TJ-II databases shows that, typically, most waveforms of a signal family are grouped into one big cluster, but there also appear to be reduced number of clusters with few signals. Different methods to perform clustering show similar results, grouping the

same signals into the same clusters, independently of the features. The application of clustering shows that the problem of finding the most similar waveforms to a given input signal can be solved very efficiently. Clustering approaches provide normally only two clusters. Firstly, the big one symbolizes that most signals translate an average physical behavior of the measured plasma property. Secondly, the rest of the waveforms can be integrated into a single cluster. The latter includes non-average behaviors and, therefore, signals classified in this group reveal non-standard plasma properties. This fact helps diagnosticians because they can find, in an automated way, interesting data to be analyzed, instead of having to search for them manually.

Information retrieval is important for fast analysis of similar plasma behavior. Automatic mechanisms for searching similarities and retrieving entire or specific data from the signal database is highly desired in fusion databases. This Thesis considers some approaches. Structural pattern recognition techniques are an efficient way to implement a pattern oriented data retrieval paradigm. Since the approach is quite flexible, many variants have been implemented. All the methods are based on the computation of primitives, translating the searching of similar waveforms to a string matching problem. Thus, the power of relational database management system and SQL pattern searching mechanism can be used to perform similar waveform retrieval.

Noise reduction is always a problem in nuclear fusion signals. Diagnostic acquisition in this environment is very hostile to experimental measurements from the electromagnetic point of view. Segmentation-based methods have proven to be useful for removing the stray light in the TJ-II TS diagnostic without eliminating significant information. ERCC shows a great performance even though the predicate, the connectivity, which defines when a pixel belongs to a region, is very simple. Region growing improves the region extraction approach of ERCC, involving a more complex predicate that allows distinguishing in a better way when a pixel belongs to a region. The performance of both algorithms have been tested by a validation method that allows to quantify the removed information. Over 90% of images are cleared of stray light.

Although an important number of problems have been considered in this work, there is still room for many other advances. Next section explains some problems that could be addressed in the future.

4.2 Future works

All problems presented in this work can be extended in many aspects. The application of the proposals in other fusion devices is one of the first future work that can be tackled. However, many other problems could have a higher interest from a research point of view. Three important issues could be considered: Firstly, a proposal research to search pattern-like on Thomson Scattering databases is described. Secondly, a proposal research to perform automatic selection of feature and building multi-layer approaches is introduced. Finally, a third proposal research is presented. The goal of this proposal is to build models when databases are unbalance or there is a small number of data samples.

Regarding the searching pattern-like on TS databases, the structural pattern recognition approach, used in this Thesis, could be extended to be applied on images. The key idea is to translate each pixel or set of pixels into a predefined primitive. This primitive or code, normally represented as a letter, will transform the original problem (searching similar image shapes) into a much more reduced one (searching similar text strings). This string-based search will require, firstly, reducing the dimensionality of images, probably by means of wavelets and secondly, to perform the searching by using a query language on fusion databases.

A second interesting problem is the automatic feature selection and the enhancement of multi-layer approaches (Weston et al. 2001). Before creating a model to predict a specific fusion event, it is necessary to define a suitable set of signals (features) in order to get a high success rate. This process will start from a large set of candidate features, which are normally selected by some previous experience or by the opinion of experts. The using of such set of features normally implies a high amount of data and computational power. So instead of using all features of the original set, it is better to perform a selection process in order to get a suitable subset of signals. The feature selection process evaluates a subset of signals as a group of suitability.

After the attributes are chosen, the training dataset has to be used in order to build classifiers and predictors. Traditionally, the output of this process corresponds to one single model. The prediction of such single model could be suitable for simple problems, but its accuracy is in general not enough for detecting fusion events. For that reason,

the detection of confinement transitions or plasma disruptions have been addressed by combining several classifiers trained with different subsets of features. Although the results so far are good, there is still room for improvements (Cannas et al. 2003, 2006, Murari et al. 2008, 2009, Rattá 2012, Ruiz et al. 2010). For instance, there is not a general methodology to build multi-layer classifiers for the discussed fusion events. Besides subsets of features are isolated instead of to group features according to their physical meaning or domain (time, frequency). New research could be done in this sense to propose a multi-layer solution for predicting a specific fusion event.

The third research problem comes which is called learning from scratch. In fusion, there are normally huge databases with an enormous number of samples available to be used by machine learning methods. However, there are some situations where databases have small training sets, and maybe the class distribution of available data does not match the target distribution (Chi et al. 2008, Forman & Cohen 2004). Such unwanted situations could appear when the fusion device has suffered a change or modification on its mechanical/electrical configuration, and old data could not be totally suitable to build an updated model.

In fact, ITER will not have any data available to build models in a classical way, until after of a couple of months of normal operation. Important fusion events such as the L-H transitions, or plasma disruptions could involve severe damage on the device if there is not a suitable model for predictions. Thus, a great challenge is to train models with small datasets. The proposal research is to speed up the building of new models in this context. A primary research line is to obtain new models by using the small training datasets, but also considering old databases and old models.

II

Resumen de la investigación

Capítulo 1

Introducción

La energía es un elemento crucial para la subsistencia de nuestra civilización. En la práctica cualquier actividad humana requiere energía para funcionar. Esta necesidad aumenta año tras año, especialmente debido al crecimiento de la población, la cual se estima en cerca de 10000 millones de personas hacia la mitad de este siglo.

Hoy en día, los combustibles fósiles son la principal fuente de energía debido a su bajo coste de producción y alta capacidad energética. Sin embargo éstos no representan una opción a largo plazo. Una alternativa más sustentables viene dada por el uso de las energías renovables, aunque la tecnología actual de tales fuentes todavía no es capaz de suplir todas las necesidades energéticas del planeta. Sin embargo, las energías nucleares tienen el potencial de proporcionar grandes cantidades de energía.

1.1 Fusión Nuclear

Aunque la energía de fusión aún esta en desarrollo su potencial es enorme, incluso comparada con la fisión nuclear. La fusión nuclear, la fuente de energía de las estrellas, podría ser más barata, limpia y segura que la energía generada por las actuales plantas nucleares.

La fusión nuclear es el proceso por el cual dos o más núcleos atómicos se unen para formar un único nucleo más pesado. Este proceso es acompañado por la liberación de grandes cantidades de energía. La fusión es la fuente que proporciona la energía de las estrellas activas, la bomba de hidrógeno y algunos dispositivos experimentales. La fusión podría proporcionar un nivel mucho mayor de energía que cualquier otra

tecnología existente en la actualidad, y el combustible que se requiere para esta fuente, principalmente deuterio, se encuentra en el océano de forma abundante. La fusión podría, en teoría, suplir todas las necesidades energéticas del mundo durante millones de años.

Con el fin de lograr reproducir la fusión nuclear en la Tierra, se pueden utilizar algunos métodos de producción bastante conocidos. Uno de los más importantes es el ciclo deuterio-tritio (Sheffield 1994), el cual libera la cantidad de 17.58 MeV, tal como muestra la Ecuación (1.1).

$$ {}^{2}_{1}D + {}^{3}_{1}T \longrightarrow {}^{4}_{2}He\,(+3.52MeV) + {}^{1}_{0}\,n\,(+14.06MeV) \tag{1.1} $$

En un dispositivo experimental de fusión la reacción se produce a temperaturas muy elevadas, cerca de 150 millones de grados Celsius. A esta temperatura, la materia contenida en los dispositivos de fusión se encuentra en forma de plasma, un estado de la materia similar a un gas con un porción de partículas ionizadas (Reitz & Milford 1996, Lawson 2002). Para confinar el plasma dentro un dispositivo de fusión en forma de toroide, se utilizan campos magnéticos. Las configuraciones de dispositivos más comunes de confinamiento magnético son los *stellerators* (Wakatani 1998) y *tokamaks* (Lister et al. 1997).

ITER, siglas en inglés de International Thermonuclear Experimental Reactor, es un proyecto de ingeniería internacional para la investigación sobre fusión nuclear que actualmente esta en construcción en Cadarache (Francia). ITER será el mayor y más avanzado reactor experimental de fusión nuclear de tipo tokamak. Se espera que ITER demuestre que es posible obtener una mayor cantidad de energía a través de la fusión, que la energía requerida para iniciar el proceso. Si el experimento resulta un éxito, se prevé que sea posible construir el primer dispositivo, denominado DEMO, que demuestre la factibilidad comercial de producir energía de fusión en forma continuada.

Actualmente, existen varios dispositivos de fusión en funcionamiento. JET (EFDA 2013*a*), del inglés Joint European Torus, es un reactor experimental también de tipo tokamak ubicado en Oxfordshire (UK). JET es actualmente el dispositivo de mayor dimensión en operación. Por su parte, el dispositivo TJ-II (Alejaldre et al. 1999) es un dispositivo de tipo stellarator de dimensiones medianas que se encuentra localizado en

Madrid (España). Otra máquina del tipo tokamak, el dispositivo fusion nuclear DIII-D (General Atomics 2013) se encuentra ubicado en las dependencias del centro General Atomics en San Diego (USA).

1.2 Motivación y Formulación General del Problema

Los experimentos en dispositivos de fusión son realizados mediante lo que se denomina descarga, pulso o disparo, y en los cuales el plasma existe en el interior del toroide. La duración de un pulso es normalmente de decenas de segundos (Ongena 2006). Se espera sin embargo que en el proyecto ITER se pueda alcanzar una descarga de aproximadamente 30 minutos.

Durante la descarga muchos de los diagnósticos realizan la medición de una gran variedad de variables físicas a altas tasas de muestreo. En JET, una descarga puede generar alrededor de 10GBytes de información (Vega et al. 2007). ITER podría incluso almacenar 1 TByte por disparo. Bolometría, densidad, temperatura, y rayos X suaves son sólo algunos ejemplos de los miles de datos que son adquiridos durante una descarga. Por ello, es común que los experimentos realizados en los dispositivos de fusión generen bases de datos masivas con enormes cantidades de datos.

A pesar de la inmensa cantidad de información obtenida, se estima que hoy en día sólo un 10% de los datos son procesados. El restante 90% no es tratado en absoluto. Por lo tanto existe una necesidad importante de analizar por completo las actuales bases de datos que se obtienen en los dispositivos experimentales de fusión nuclear (tokamaks y stellartors). Este análisis en profundidad podría permitir alcanzar una fuente de energía limpia, inagotable, segura y barata para la humanidad en un futuro cercano. Como primer paso, el tratamiento intensivo de la información podría permitir la operación óptima de ITER, lo cual redundaría en un diseño exitoso para el funcionamiento de DEMO, el primer reactor de fusión nuclear comercial.

Por esta razón, la Tesis doctoral propone el uso de técnicas avanzadas para el reconocimiento de patrones y aprendizaje automático, con la finalidad de analizar de una forma más rápida y eficiente las inmensas bases de datos de fusión nuclear. Hasta ahora, aunque se ha realizado una gran esfuerzo en este sentido, todavía existe un gran espacio para trabajar en esta línea de investigación. Específicamente, se requiere de la elimi-

nación o reducción de ruido en señales temporales o imágenes para mejorar el análisis de los diagnósticos. Así mismo, la selección automática de características o atributos y los enfoques multi-capa podrían ser útiles para obtener mejores clasificadores y predictores del comportamiento del plasma. Por último, lograr la disminución de los tiempos asociados a la búsqueda de patrones similares, que podrían indicar comportamientos equivalentes del plasma en diferentes descargas, ayudaría a incrementar la información tratada en las bases de datos. Todas estas tareas son consideradas en este trabajo, lo que en definitiva haría que el análisis de la enorme información existente fuera realizado de manera más rápida y profunda.

1.3 Objetivos de la Tesis

El objetivo general de esta Tesis es desarrollar métodos avanzados de reconocimiento de patrones y aplicar técnicas innovadoras de aprendizaje automático para el análisis asistido por computador de grandes bases de datos en fusión nuclear. Dado el muy amplio rango de tópicos que pueden ser considerados, esta Tesis se enfoca en algunos problemas particulares de los dispositivos de fusión nuclear: TJ-JJ, JET y DIII-D. Sin embargo, muchos de los resultados obtenidos en la Tesis podrían ser adaptados o reproducidos en otras máquinas de fusión tales como ITER.

Los objetivos específicos de la Tesis son los siguientes:

- Aplicar técnicas avanzadas de clasificación y agrupamiento a bases de datos de fusión nuclear.

- Desarrollar métodos eficientes de búsqueda y recuperación de información en bases de datos masivas.

- Diseñar y validar algoritmos para eliminar patrones de ruido en imágenes. Aplicaciones al diagnóstico Thomson Scattering.

- Estudiar y desarrollar procedimientos de selección automática de características. Aplicaciones a transiciones de modo de confinamiento en plasmas termonucleares.

1.4 Contribuciones Principales

Las contribuciones principales de esta Tesis se resumen en las conclusiones principales de este trabajo, y además ha dado lugar a los siguientes desarrollos y publicaciones.

1.4.1 Componentes de Software Desarrollados

Los resultados concretos de esta Tesis doctoral incluyen la implementación de algoritmos avanzados de reconocimiento de patrones y aprendizaje automático en MATLAB, y el desarrollo de interfaces gráficas de usuario para realizar pruebas de datos de manera rápida y sencilla. Los resultados más importantes son descritos a continuación:

- Herramientas propias de reconocimiento de patrones programadas en MATLAB.
- Interfaces gráficas de usuario programadas en MATLAB para propósitos de investigación y enseñanza.
- Aplicación de las técnicas de máquinas de vectores soporte y transformada Wavelet para el reconocimiento de patrones en fusión nuclear.
- Aplicación de técnicas de reconocimiento de patrones estructural para la búsqueda de formas de onda específicas.
- Clasificación automatizada de imágenes en el diagnóstico Thomson Scattering.
- Detección de diferentes eventos físicos del plasma en señales de fusión.

1.4.2 Publicaciones

Durante la Tesis doctoral se han publicado aportaciones en conferencias internacionales y en revistas especializadas. Muchos de los artículos han sido obtenidos como un resultado directo de esta Tesis. Otros han sido desarrollados en colaboración por el autor con diversos grupos de investigación.

Artículos Publicados en Revistas

Los siguientes artículos fueron publicados en revistas especializadas y tienen directa relación con la Tesis doctoral:

- G. Farias, S. Dormido-Canto, J. Vega, I. Pastor, M. Santos (2013) Application and validation of image processing algorithms to reduce the stray light on the TJ-II Thomson Scattering diagnostic, *Fusion Science and Technology*, ISSN 1536-1055, Volume 63, Number 1, Pages 20–25.

- G. Farias, J. Vega, S. González, A. Pereira, X. Lee, D. Schissel, P. Gohil (2012) Automatic determination of L/H transition times in DIII-D through a collaborative distributed environment, *Fusion Engineering and Design*, ISSN 0920-3796, Volume 87, Issue 12, Pages 2081–2083.

- S. Dormido-Canto, G. Farias, J. Vega, I. Pastor (2012) Image processing methods for noise reduction in the TJ-II Thomson Scattering diagnostic, *Fusion Engineering and Design*, ISSN 0920-3796, Volume 87, Issue 12, Pages 2170-2173.

- L. Makili, J. Vega, S. Dormido-Canto, I. Pastor, A. Pereira, G. Farias, A. Portas, D. Pérez-Risco, M.C. Rodríguez-Fernández, P. Busch (2010) Upgrade of the automatic analysis system in the TJ-II Thomson Scattering diagnostic: New image recognition classifier and fault condition detection, *Fusion Engineering and Design*, ISSN 0920-3796, Volume 85, Issues 34, Pages 415-418.

- M. Santos, G. Farias (2010) Laboratorios virtuales de procesamiento de señales, *Revista Iberoamericana de Automática e Informática Industrial (RIAI)*, ISSN 1697-7912, Volume 7, Number 1, Pages 91-100.

- J.A. Martín, M. Santos, G. Farias, N. Duro, J. Sánchez, R. Dormido, S. Dormido-Canto, J. Vega, H. Vargas, (2009) Dynamic clustering and modeling approaches for fusion plasma signals, *IEEE Transactions on Instrumentation and Measurement*, ISSN 0018-9456, Volume 58, Number 9, Pages 2969-2978.

- N. Duro, R. Dormido, J. Vega, S. Dormido-Canto, G. Farias, J. Sánchez, H. Vargas, A. Murari and JET-EFDA Contributors (2009) Automated recognition system for ELM classification in JET, *Fusion Engineering and Design*, ISSN 0920-3796, Volume 84, Issues 2-6, Pages 712-715.

- S. Dormido-Canto, G. Farias, J. Vega, R. Dormido, J. Sánchez, N. Duro, H. Vargas, A. Murari, and JET-EFDA Contributors (2008) Classifier based on support vector

machine for JET plasma configurations, *Review of Scientific Instruments*, ISSN 0034-6748, Volume 79, Pages 10F326-1/10F326-3.

- S. Dormido-Canto, G. Farias, R. Dormido, J. Sánchez, N. Duro, H. Vargas, J. Vega, G. Ratta, A. Pereira, A. Portas (2008) Structural pattern recognition methods based on string comparison for fusion database, *Fusion Engineering and Design*, ISSN 0920-3796, Volume 83, Issue 2-3, Pages 421-424. Ed. Elsevier.

- G. Rattá, J. Vega, A. Pereira, A. Portas, E. De la Luna, S. Dormido-Canto, G. Farias, R. Dormido, J. Sánchez, N. Duro, H. Vargas, M. Santos, G. Pajares, A. Murari, and JET EFDA Contributors (2008) First applications of structural pattern recognition methods to the investigation of specific physical phenomena at JET, *Fusion Engineering and Design*, ISSN 0920-3796, Volume 83, Issue 2-3, Pages 467-470. Ed. Elsevier.

- J. Vega, A. Pereira, A. Portas, S. Dormido-Canto, G. Farias, R. Dormido, J. Sánchez, N. Duro, M. Santos, E. Sánchez, G. Pajares (2008) Data mining technique for fast retrieval of similar waveform in Fusion massive databases, *Fusion Engineering and Design*, ISSN 0920-3796, Volume 83, Issue 1, Pages 132-139. Ed. Elsevier.

- S. Dormido-Canto, G. Farias, J. Vega, R. Dormido, J. Sánchez, N. Duro, M. Santos, J.A. Martín, G. Pajares (2006) Search and retrieval of plasma waveforms: structural pattern recognition approach, *Review of Scientific Instruments*, ISSN 0034-6748, Volume 77, Pages 10F514-1/10F514-4.

- G. Farias, S. Dormido-Canto, J. Vega, J. Sánchez, N. Duro, R. Dormido, M. Ochando, M. Santos, G. Pajares (2006) Searching for patterns in TJ-II time evolution signals, *Fusion Engineering and Design*, ISSN 0920-3796, Volume 81, Pages 1993-1997, Ed. Elsevier.

- N. Duro, J. Vega, R. Dormido, G. Farias, S. Dormido-Canto, J. Sánchez, M. Santos, G. Pajares (2006) Automated clustering procedure for TJ-II experimental signals, *Fusion Engineering and Design*, ISSN 0920-3796, Volume 81, Pages 1987-1991, Ed. Elsevier.

- S. Dormido Canto, J. Vega, Sánchez, G. Farias (2005) Information retrieval and classification with wavelets and support vector machines, *Lecture Notes in Computer Science*, ISSN 0302-9743, Volume 3562, Part 2, Pages 548-557, Springer-Verlag.

- G. Farias, R. Dormido, M. Santos, N. Duro (2005) Image classifier for the TJ-II Thomson Scattering diagnostic: Evaluation with a feed forward neural network, *Lecture Notes in Computer Science*, ISSN 0302-9743, Volume 3562, Part 2, Pages 604-612, Springer-Verlag.

- S. Dormido-Canto, G. Farias, R. Dormido, J. Vega, J. Sánchez, M. Santos and The TJ-II Team (2004) TJ-II wave forms analysis with wavelets and support vector machines, *Review of Scientific Instruments*, ISSN 0034-6748, Volume 75, Pages 4254-4257.

Los siguientes artículos son fruto de la colaboración con otros grupos de investigación. El tópico de estos trabajos es la aplicación de técnicas de reconocimiento de patrones en el ámbito de la biomedicina.

- G. Farias, M. Santos, V. López (2010) Making decisions on brain tumor diagnosis by soft computing techniques, *Soft Computing*, ISSN 1432-7643, Volumen 14, Number 12, Pages 1287-1296.

- G. Farias, M. Santos (2007) A computational fusion of wavelets and neuronal networks in a classifier for biomedical applications, *Lecture Series on Computer and Computational Sciences*, ISSN 1573-4196, Volume 8, Pages 66-70, Brill Academic Publishers.

Artículos Publicados en Conferencias

Los siguientes artículos han sido publicados en conferencias nacionales e internacionales relacionadas principalmente con reconocimiento de patrones en fusión nuclear. A continuación se presentan las publicaciones más importantes:

- Farias G., Vega J., Gonzalez S., Pereira A., Lee X., Schissel D., Gohil P. (2011) *Automatic determination of L/H transition times in DIII-D through a collabora-*

tive distributed environment, 8^{th} IAEA Technical Meeting on Control, Data Acquisition, and Remote Participation for Fusion Research, June 20-24, 2011, San Francisco, USA.

- Dormido-Canto S., Farias G., Vega J., Pastor I. (2011) *Image processing methods for noise reduction in the TJ-II Thomson Scattering Diagnostic*, 8^{th} IAEA Technical Meeting on "Control, Data Acquisition, and Remote Participation for Fusion Research", June 20-24, 2011, San Francisco, USA.

- G. Farias, S. Dormido, F. Esquembre, H. Vargas, S. Dormido-Canto (2008) *Laboratorio virtual para la enseñanza de técnicas de reconocimiento de patrones*, XIII Latin-American Congress on Automatic Control. Mérida, Venezuela.

- G. Farias, M. Santos, V. López (2008) *Brain tumour diagnosis with wavelets and support vector machines*, 3rd International Conference on Intelligent System and Knowledge Engineering, Proceedings of the 3rd ISKE, IEEE Press, ISBN: 978-1-4244-2197-8, pp: 1453-1459, November 17-19, Xiamen, China.

- Martin, J. A. Santos, M., Farias, G., Duro, N., Sánchez, J., Dormido, R., Dormido-Canto, S., Vega, J. (2007) *Dynamic clustering and neuro-Fuzzy identification for the analysis of fusion plasma signals*, Proc. of the IEEE International Symposium on Intelligent Signal Processing. ISBN: 1-4244-0829-6. pp: 979-984.

- Vega, J., Rattá, G., Murari, A., Castro, P., Dormido-Canto, S., Dormido, R., Farias, G., Pereira, A. Portas, A., de la Luna, E., Pastor, I., Sánchez, J., Duro, N., Castro, R., Santos, M., Vargas, H. (2007) *Recent result on structural pattern recognition for fusion massive database*, Proc. of the IEEE International Symposium on Intelligent Signal Processing. ISBN: 1-4244-0829-6. pp: 949-954.

- Farias G., Dormido-Canto S., Vega J., Sánchez J., Duro N., Dormido R., Ochando M., Pajares G., Santos M. (2005) *Searching patterns in TJ-II temporal evolution signals with support vector machines*, Fifth IAEA Technical Meeting on Control, Data Acquisition, and Remote Participation for Fusion Research, Budapest, Hungary.

- Duro N., Vega J., Dormido R., Farias G., Dormido-Canto S., Sánchez J., Santos

M., Pajares G. (2005) *Automated clustering procedure for TJ-II experimental signals*, Fifth IAEA Technical Meeting on Control, Data Acquisition, and Remote Participation for Fusion Research, Budapest, Hungary.

- Farias G., Santos M. (2005) *Aplicación de técnicas de inteligencia artificial y tratamiento de señales en fusión*, 1er Simposio de Control Inteligente, 1- 3 Junio, Huelva, España.

- Farias G., Santos M., Dormido-Canto S. (2005) *Desarrollo de una aplicación para la integración de técnicas de reconocimiento de patrones*, XXVI Jornadas de Automática, Alicante-Elche, España, ISBN: 84-689-0730-8.

- Vega J., Pastor I., Cereceda J. L., Pereira A., Herranz J., Pérez D., Rodríguez M. C., Farias G., Dormido-Canto S., Sánchez J., Dormido R., Duro N., Dormido S. (2005) *Application of intelligent classification techniques to the TJ-II Thomson Scattering diagnostic*, 32nd EPS Plasma Physics Conference, 8th International Workshop on Fast Ignition of Fusion Targets. 27 June- 1 July, Tarragona- Spain.

- Farias G., Santos M., Marrón J. L., Dormido-Canto S. (2004) *Determinación de parámetros de la transformada wavelets para la clasificación de señales del diagnóstico Scattering Thomson*, XXV Jornadas de Automática, Ciudad Real, España, ISBN: 84-688-7460-4.

- Dormido S., De la Cruz J.M., Vega J., Santos M., Dormido-Canto S., Sánchez J., Dormido-Canto R., Farias G. (2004) *Análisis de formas de onda de plasmas con wavelets y support vector machines*, 3ra. Conferencia Iberoamericana en Sistemas, Cibernética e Informática CISCI. Orlando, USA.

1.4.3 Estancias Breves de Investigación

Durante la realización de la Tesis doctoral se han realizado estancias de investigación en los siguientes laboratorios de fusión nuclear:

- **Periodo**: Enero, 2011.

  Laboratorio: General Atomics (San Diego, USA).

  Objetivo: Estudios de la transición L-H en el dispositivo de fusión DIII-D.

  Supervisores: Jesús Vega (CIEMAT) y David Schissel (General Atomics).

Publicación: G. Farias, J. Vega, S. Gonzalez, A. Pereira, X. Lee, D. Schissel, P. Gohil *Automatic determination of L/H transition times in DIII-D through a collaborative distributed environment*, 8^{th} IAEA Technical Meeting on "Control, Data Acquisition, and Remote Participation for Fusion Research", June 20-24, 2011, San Francisco, USA.

- **Periodo**: Febrero-Junio, 2011.

  Laboratory: Laboratorio Nacional de Fusión, CIEMAT (Madrid, Spain).

  Objetivo:Estudios de la transición L-H en el dispositivo de fusión DIII-D, y reducción de luz parásita en el diagnóstico Thomson Scattering en el TJ-II.

  Supervisores: Jesús Vega (CIEMAT) y David Schissel (General Atomics).

  Publicaciones: S. Dormido-Canto, G. Farias, J. Vega, I. Pastor *Image processing methods for noise reduction in the TJ-II Thomson Scattering diagnostic*, 8^{th} IAEA Technical Meeting on "Control, Data Acquisition, and Remote Participation for Fusion Research", June 20-24, 2011, San Francisco, USA.

Capítulo 2

Diagnósticos y Eventos Físicos del Plasma

2.1 Diagnósticos del Plasma

El CIEMAT, y específicamente la asociación EURATOM/CIEMAT para la fusión por confinamiento mágnetico, obtiene a través de un gran número de experimentos una enorme cantidad de señales en el dispositivo de fusión nuclear TJ-II (ver Figura 2.1).

El TJ-II (Alejaldre et al. 1999) es un dispositivo de fusión de tamaño medio de tipo stellerator (Tipo helicoidal, campo magnético $B_0 = 1.2T$, promedio radio mayor $R(0) = 1.5m$, promedio radio menor $\leq 0.22m$) localizado en el CIEMAT (Madrid, España). Los plasmas del TJ-II son producidos utilizando calentamiento por resonancia ciclotrónica de electrónes (ECRH) (dos girotrones, de 300 kW cada uno, 53.2 GHz, segundo armónico, polarización modo-X) y de inyección de haces de átomos neutros de hidrógeno (NBI, 300 kW). Actualmente, existen 940 canales digitales para las mediciones experimentales en el TJ-II. Los dispositivos de fusión generan enormes cantidades de datos. Típicamente, cada descarga genera miles de señales con una alta dimensionalidad de muestras y atributos.

En el TJ-II, el confinamiento magnético es obtenido mediante un conjunto de bobinas que determinan completamente las superficies magnéticas antes de la existencia del plasma. El campo toroidal es creado por 32 bobinas. El giro tridimensional del eje central de la configuración se genera mediante dos bobinas centrales: una circular y otra helicoidal. La posición horizontal del plasma se controla mediante las bobinas de campo

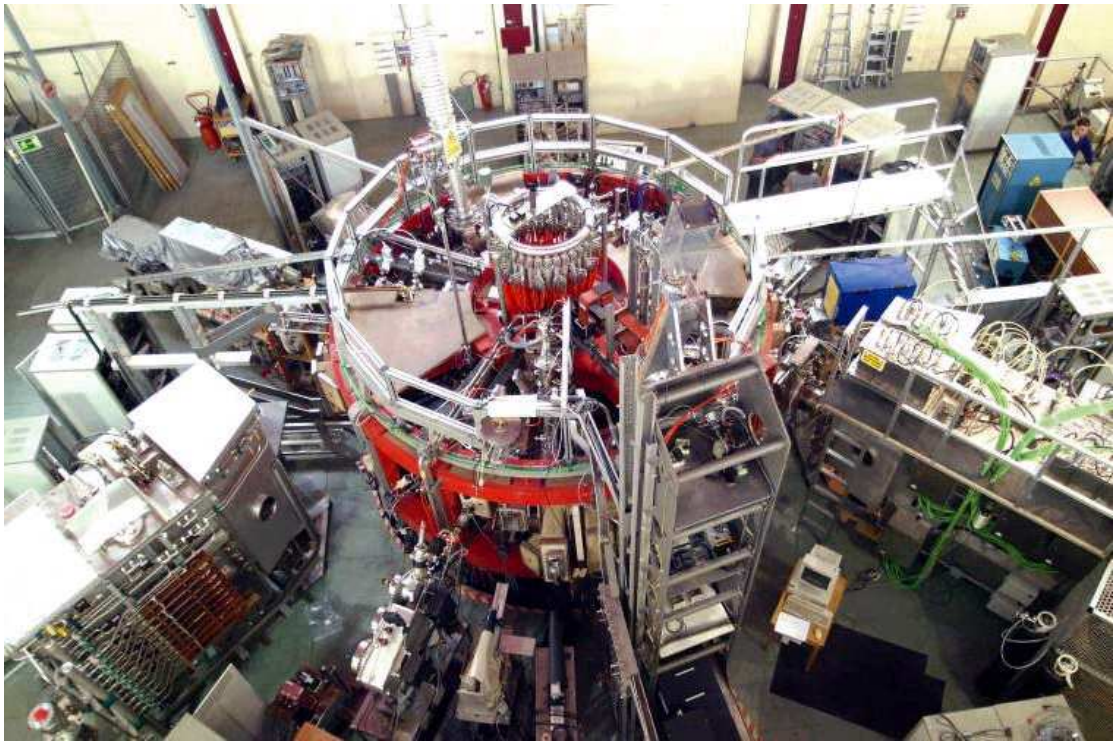


Figura 2.1: TJ-II: Dispositivo de fusión nuclear del tipo stellerator.

vertical. La acción conjunta de estos campos magnéticos genera superficies magnéticas con forma de "judía" que guían las partículas del plasma para que no choquen con las paredes de la cámara de vacío (Alejaldre et al. 1999). La Figura 2.2 muestra un esquema de dos matrices de sensores para adquirir señales bolométricas.

La duración de las descargas en el TJ-II es de alrededor de 150-250ms, con una frecuencia de repetición de aproximadamente 7 minutos. Dependiendo de la tasa de muestreo, el número de muestras podría estar en el rango de 4000 a 16000 por descarga.

En esta Tesis se analizarán dos tipos de señales. Señales temporales (datos unidimensionales) de diferentes diagnósticos, y señales bidimensionales (imágenes) provenientes del diagnóstico Thomson Scattering. Las metodologías aplicadas en este trabajo pueden ser extendidas a otros tipos de dispositivos de fusión para problemas similares de reconocimiento de patrones. De hecho, técnicas similares de minería de datos han sido utilizadas en los dispositivos de fusión DIII-D (General Atomics, San Diego, USA) y JET (EFDA, Oxfordshire, UK). Tal capacidad de los algoritmos de reconocimiento de patrones abre la posibilidad de aplicar trabajos y resultados similares ya realizados en el proyecto ITER.

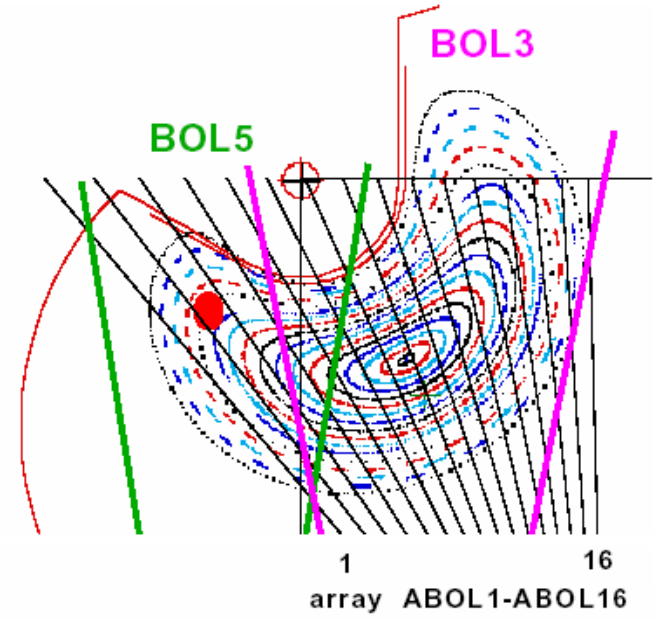


Figura 2.2: Arrays de sensores en la superficie magnética del plasma.

2.1.1 Señales Temporales de Diagnósticos

La Tabla 2.1 muestra algunas señales temporales de la base de datos del TJ-II. Cada una describe una medida particular de una propiedad física del plasma. Su utilidad por supuesto, radica en que por ejemplo una combinación de mediciones de bolometría y rayos X pueden caracterizar la evolución temporal de la densidad del plasma. Los datos que estos sensores proporcionan son del tipo serie temporal, donde una de las coordenadas es el tiempo, y la otra coordenada corresponde a la magnitud de la medida. Se debe tener en cuenta que estas señales pueden estar constituidas por millones de muestras.

Tabla 2.1: Algunas señales temporales adquiridas en el TJ-II.

Clase de señal	Descripción
RX306	rayos X suave
ACTON275	Señal espectroscópica (CV)
HALFAC3	H_α
DENSIDAD2	Densidad de línea media
BOL5	Señal bolométrica
ECE7	emisión ciclotrónica

La Figura 2.3 muestra señales temporales provenientes de los diagnósticos ACTON275, BOL5, Densidad2, y ECE7 para la descarga N^o 10108 en el TJ-II. Nótese que el tiempo esta dado en milisegundos.

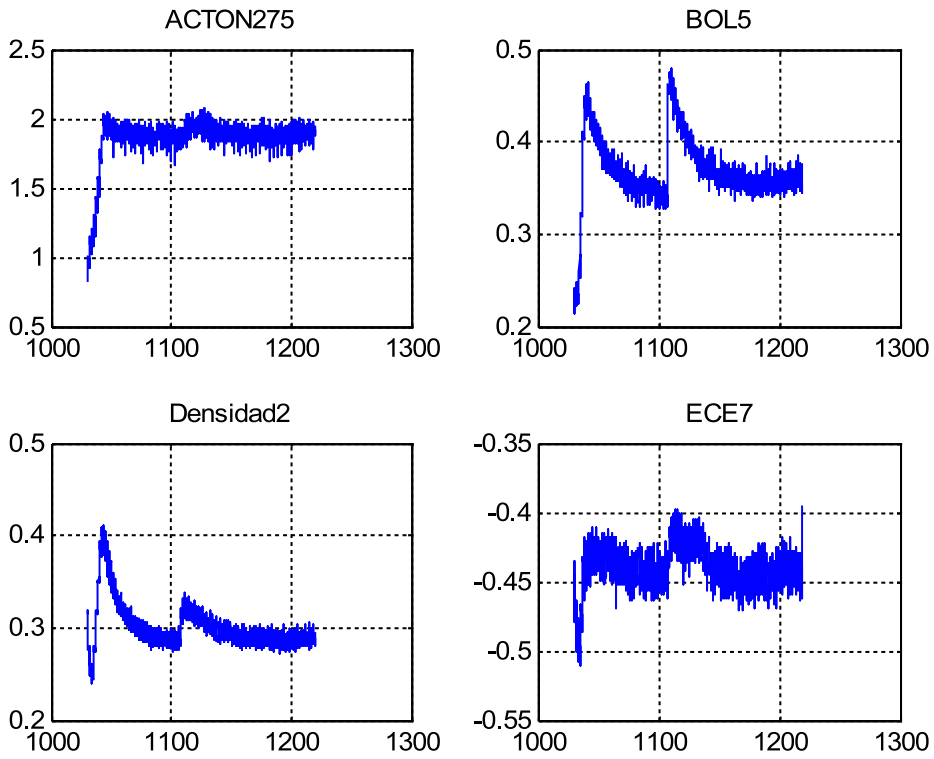


Figura 2.3: Señales temporales para la descarga N^o 10108.

Aunque se puede encontrar patrones representativos de cada clase de señal temporal, las señales de una clase en particular no siempre son similares para diferentes descargas. Esta situación hace necesario el estudio de subclases dentro de cada tipo de señal, lo cual podría en teoría indicar la existencia de un comportamiento físico del plasma distinto. Más detalles acerca del estudio de subclases en los diagnósticos se pueden encontrar en 3.1.3.

2.1.2 Diagnóstico Thomson Scattering en el TJ-II

El diagnóstico Thomson Scattering (TS) del plasma consiste en la re-emisión de radiación incidente (proporcionada por potentes láseres) de electrones libres. La distribución de velocidad electrónica genera un ensanchamiento espectral de la luz dispersada (por efecto Doppler) en relación a la temperatura electrónica. El número total de fotones dispersados es proporcional a la densidad electrónica. La Figura 2.4 muestra un diagrama del diagnostico Thomson Scattering implementado en el TJ-II.

Cada disparo del láser produce una imagen (información espacial o bidimensional) a partir de la cual es posible obtener perfiles radiales de temperatura y densidad. En

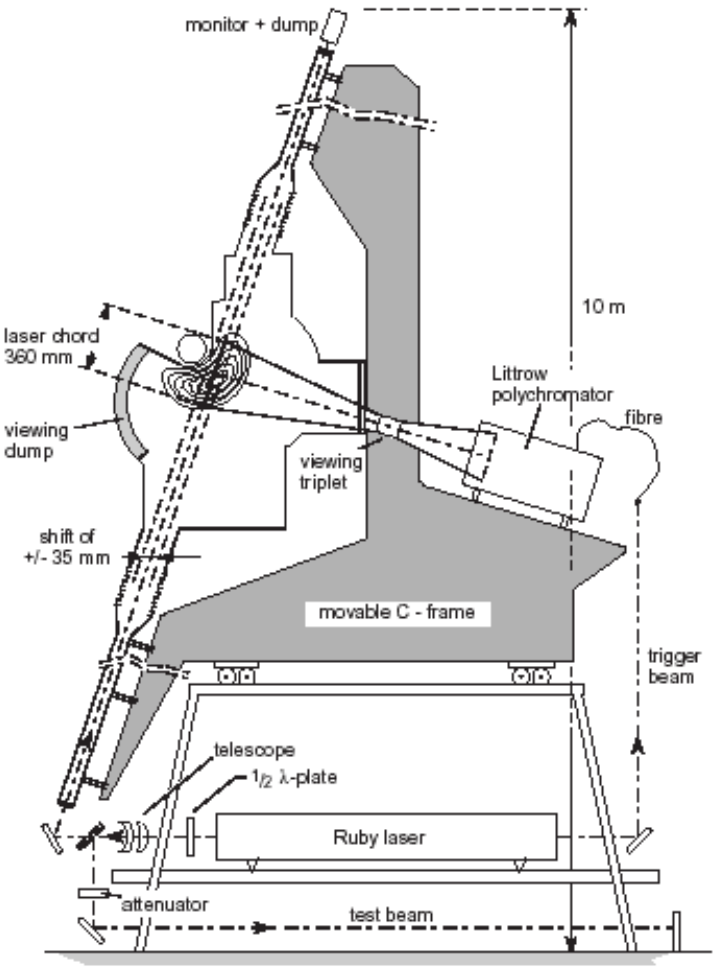


Figura 2.4: Diagrama del diagnóstico Thomson Scattering en el TJ-II.

el TJ-II se adquiere sólo un número conocido de patrones. Cada imagen tiene 385x576 píxeles, *i.e.* 221760 atributos posibles.

Los tipos de imágenes representan diferentes comportamientos físicos relacionados con el modo de calentamiento del plasma o el sistema de calibración. El diagnóstico Thomson Scattering (Farias et al. 2005, Makili et al. 2010) del TJ-II obtiene cinco tipos particulares de patrones (ver Figura 2.5): Fondo de cámara CCD (BKG), medición de luz parásita sin plasma o de una descarga interrumpida (STR), imágenes durante la fase ECRH (ECH), imágenes durante la fase NBI de calentamiento (NBI), y finalmente imágenes obtenidas después del corte de densidad durante la fase ECRH (COFF). Desde el punto de vista de la física del plasma, los patrones más importantes son del tipo ECH y NBI debido a que éstos corresponden a plasma de alta temperatura. En ambos casos, las imágenes son procesadas para obtener los perfiles radiales de temperatura y densidad electrónica.

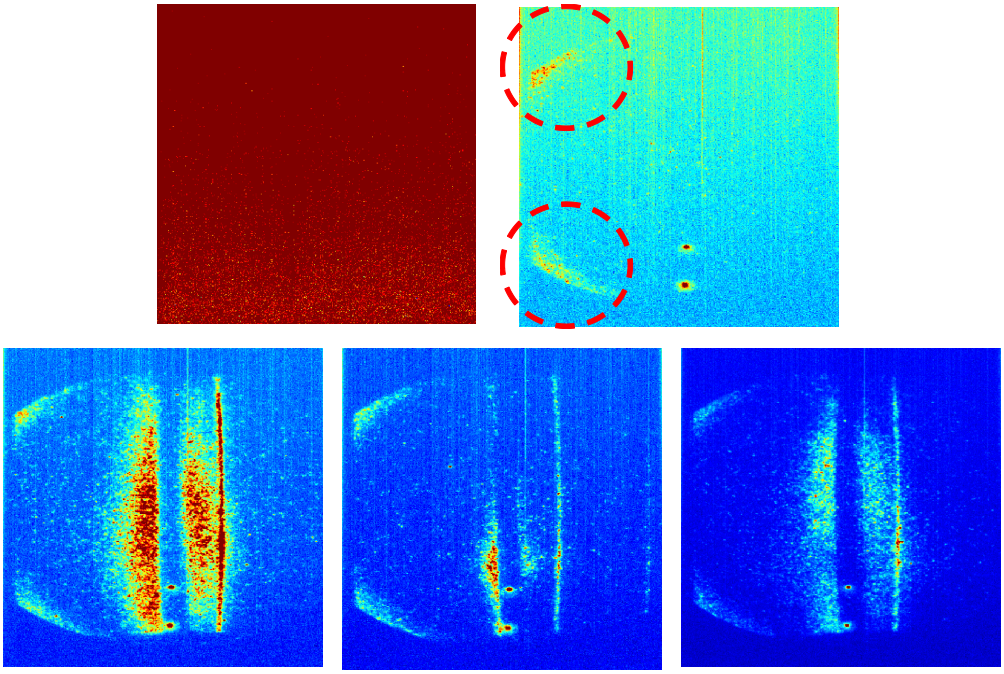


Figura 2.5: Clases de imágenes adquiridas por el diagnóstico Thomson Scattering en el TJ-II: BKG, STR arriba, y NBI, COFF y ECH abajo. Las circunferencias rojas en la imagen STR muestran el ruido debido a la luz parásita. Nótese que este ruido aparece en todas las clases excepto en el tipo BKG.

Luz Parásita en el Diagnóstico Thomson Scattering en el TJ-II

La cámara CCD obtiene frecuentemente imágenes con ruido. En el diagnóstico Thomson Scattering la principal fuente de ruido proviene por lo que se conoce como luz parásita. El control de la luz parásita siempre ha sido un problema óptico importante (Breault 1995). Causada por fenómenos tales como la reflexión de Fresnel de la superficie de las lentes, burbujas de aire en el cristal, polvo, difracción debido a bordes, y varios otros efectos, la presencia de este ruido degrada frecuentemente el contraste de la imagen y la precisión de la medida. En particular, la cámara CCD en el diagnóstico Thomson Scattering en el TJ-II adquiere imágenes corruptas con luz parásita, que en algunos casos, puede producir perfiles imprecisos de temperatura y densidad (ver Figura 2.5). Por tanto, la aplicación de técnicas que permitan reducir este tipo de perturbación incrementará favorablemente la calidad del análisis del diagnóstico Thomson Scattering.

2.2 Eventos Físicos del Plasma

El comportamiento del plasma en los dispositivos de fusión nuclear no es fácilmente predecible; de hecho ha habido un gran esfuerzo por parte de la comunidad científica

por comprender como controlar y estabilizar la actuación del plasma durante la descarga (Schuller 1999). Sin embargo, existen algunos fenómenos o eventos de comportamiento físico del plasma bien conocidos. Esta Tesis ha considerado dos eventos importantes: La transición de modos de confinamiento (denominada transición L-H) y los modos localizados en el borde (denominados ELMs por sus siglas en inglés). La detección y localización de estos fenómenos se trata en ambos casos en este trabajo. La Figura 2.6 presenta los eventos comentados en la señal emisión Dα.

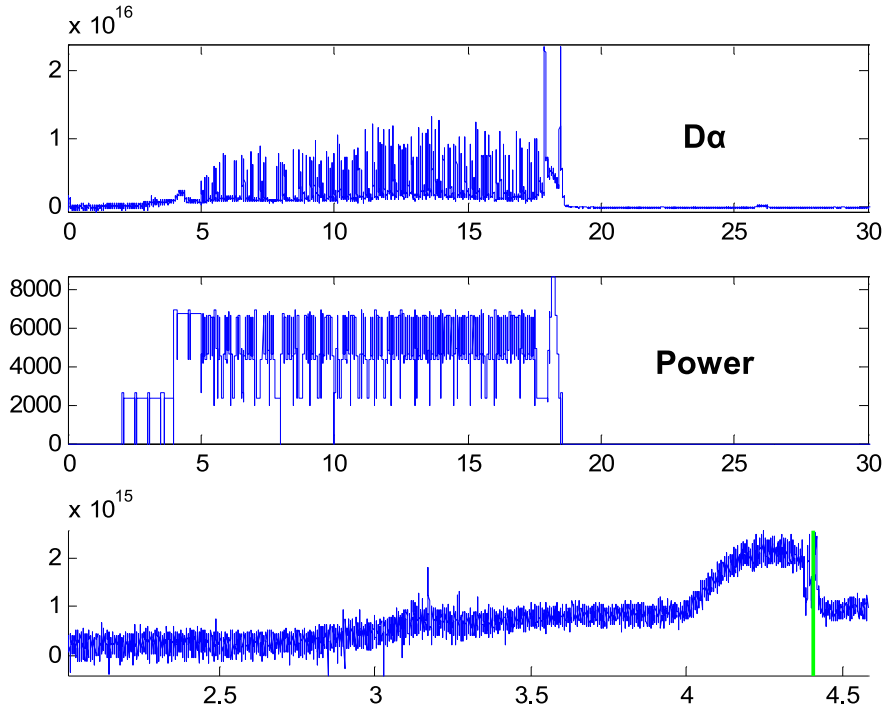


Figura 2.6: Transición L-H y ELMs: En la parte superior se muestra una señal de emisión Dα, la transición L-H ocurre alrededor de los 4.5 segundos y los ELMs aparecen en el rango [5,18]s aproximadamente. En la parte media de la figura se observa una señal de potencia inyectada para calentar el plasma (específicamente PINJ). En la parte inferior de la figura se presenta una imagen ampliada en la zona donde ocurre la transición L-H (indicada por la línea verde).

2.2.1 Transiciones del modo L al modo H

El modo H (modo de alta energía) es uno de los principales regímenes de confinamiento en los dispositivos del tipo tokamaks y stellerators. El modo H fue detectado por primera vez en el tokamak ASDEX (Wagner et al. 1982). La variación repentina de los parámetros del plasma desde el modo L (modo de baja energía) al modo H es conocida como transición L-H. La transición L-H se caracteriza por la creación de una barrera de transporte en el borde del plasma (ETB). Una vez que el fenómeno ETB desaparece, el plasma retorna al modo L. Esta última situación es conocida como transición H-L.

Una transición L-H puede identificarse por una caída rápida de la emisión Dα entre el comienzo del calentamiento por parte del sistema NBI, y la aparición del primer ELM de tipo I. En la parte superior de la Figura 2.6 se muestra la señal de emisión Dα para una descarga particular. La duración del pulso es de aproximadamente 16 segundos. En la parte inferior de la Figura 2.6 se puede observar que la transición L-H (la caída repentina) ocurre alrededor de los 4.5 segundos, y el primer ELM de tipo I esta ubicado aproximadamente en los 5 segundos.

En esta Tesis, diversas técnicas de reconocimiento de patrones se han aplicado para determinar la ubicación de las transiciones L-H en las bases de datos de la máquina tokamak DIII-D. Para tal propósito, se ha desarrollado un sistema multicapa que utiliza datos alrededor de transiciones L-H (previamente localizadas por expertos) para entrenar un sistema capaz de reconocer tales transiciones en descargas nuevas (Farias et al. 2012). Enfoque similares se han utilizado en otros trabajos para estimar transiciones L-H y H-L en el dispositivo JET (González et al. 2012, Vega et al. 2009).

2.2.2 Modos Localizados en el Borde: ELMs

Para producir energía de fusión nuclear de manera eficiente se deben controlar las inestabilidades inherentes del plasma, evitando en todo momento comprometer la vida útil del material que lo rodea. Los modos localizados en el borde representan un tipo de inestabilidad que aún no se comprende completamente, y por ello todavía requiere mayor estudio y análisis. Los ELMs se pueden observar en el borde del plasma como *picos* repetitivos, por ejemplo en la intensidad de luz o en la tensión medida de una sonda eléctrica. La aparición de los modos localizados en el borde plantea un desafío importante en la investigación de los dispositivos de fusión del tipo tokamak, ya que estas inestabilidades pueden dañar componentes de la pared, particularmente el divertor, debido a su extremadamente alta tasa de transferencia de energía.

Una forma de examinar los ELMs es estudiar el comportamiento global del plasma durante la aparición de este fenómeno (Liang 2011, EFDA 2013*b*). Mientras algunas de las características de los ELMs son comunes en todos los picos, existen también diferencias significativas en relación a la frecuencia y amplitud del suceso. Se han encontrado tres tipos de picos. ELMs **Tipo I**: La emisión Dα presenta grandes y aislados vértices o picos, por ello, los ELMs de tipo I son llamados también ELMs *grandes* o incluso

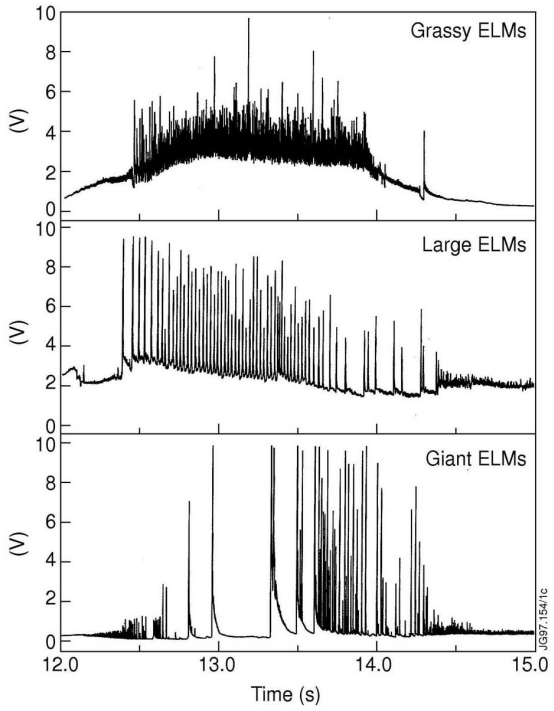


Figura 2.7: Tipos de ELMs: Los ELMs se pueden observar en el borde del plasma como picos repetitivos. La figura muestra los tres tipos de ELMs identificados. Imagen tomada desde (EFDA 2013*b*).

ELMs *gigantes*. ELMs **Tipo II**: este tipo de modo localizado en el borde se observa solamente en plasmas bien formados, es decir, en plasmas cuya sección transversal tiene un pronunciado alargamiento y triangularidad. La magnitud de los picos es inferior a los ELMs de tipo I, aunque su frecuencia es más alta. Los ELMs de tipo II son denominados normalmente ELMs de tipo *herboso* (de grassy en inglés) por su apariencia similar al césped. ELMs **Tipo III**: Los picos son más pequeños y frecuentes. Los ELMs de tipo III son también llamados ELMs *pequeños*. La frecuencia de repetición de estos ELMs disminuye en la medida que la potencia inyectada se incrementa. Más detalles acerca de las características de los modos localizados en el borde pueden encontrarse en (Liang 2011, EFDA 2013*b*, Saibene et al. 2002, Bellizio et al. 2011). La Figura 2.7 muestra la imagen de cada tipo de ELM.

Capítulo 3

Reconocimiento de Patrones en Fusión

3.1 Clasificación y Agrupamiento

El objetivo de la clasificación supervisada es encontrar una regla de decisión, basada en observaciones previas, que permita asignar un determinado objeto a una de las posibles clases. El proceso de clasificación se puede dividir en dos grandes fases: Extracción de características y asignación de clase (Duda et al. 2001, Santos & Farias 2010). La primera etapa involucra la ejecución de algoritmos de procesamiento para intentar extraer información o atributos distintivos de cada muestra. La segunda etapa consiste en asignar las muestras al conjunto de clases definidas.

El clasificador más simple de desarrollar, denominado clasificador binario, permite distinguir dos tipos de clases. Es común construir un clasificador con un mayor número de clases, denominado multi-clasificador, a partir de varios clasificadores binarios. Este enfoque se denomina frecuentemente *uno versus el resto*.

El agrupamiento, al contrario de la clasificación supervisada, intenta agrupar los datos o muestras en grupos sin conocimiento previo de cuantas clases en realidad existen. Este enfoque es bastante útil en reconocimiento de patrones ya que revela la presencia de patrones similares en las bases de datos, lo cual podría indicar que ciertas condiciones o conductas se repiten. El agrupamiento de los datos se realiza con lo que se denomina un criterio de similitud o distancia. En las bases de datos de fusión nuclear el agrupamiento puede utilizarse para descubrir la existencia de un comportamiento similar en la física

del plasma.

3.1.1 Clasificación de Imágenes del Diagnóstico Thomson Scattering

Como se mencionó anteriormente, las imágenes del diagnóstico Thomson Scattering representan diferentes situaciones físicas del plasma relacionadas con el tipo de calentamiento y el sistema de calibración. Dependiendo del patrón obtenido, los imágenes son procesadas de manera diferente. Con el fin de realizar un análisis automático de los datos, se decidió implementar un sistema de clasificación automático que indique el tipo de patrón o clase de una imagen. En esta Tesis se han implementado varios tipos de sistemas de clasificación que utilizan principalmente dos enfoques de aprendizaje automático: Redes Neuronales (Farias et al. 2005) y Máquinas de Vectores Soportes (Makili et al. 2010).

La máquina de vectores soporte (SVM) es un procedimiento de aprendizaje automático basado en la teoría de aprendizaje estadístico (Vapnik 1999, Sebald & Bucklew 2000, Schölkopf & Smola 2001, Hearst et al. 1998, Cherkassky & Mulier 2007). La SVM mapea datos de entrada en un espacio de dimensiones superiores, en donde se utilizan funciones lineales (denominadas comúnmente hiperplanos) para discriminar las muestras de entrada. Los parámetros que definen las funciones se determinan resolviendo un problema de optimización cuadrática. Sin embargo, para un espacio de características de dimensiones superiores, el gran número de parámetros a determinar hacen el problema intratable. Por esta razón, la teoría de optimización dual es utilizada en SVM con el objetivo de estimar los parámetros de una forma computacionalmente abordable. La función de aproximación lineal correspondiente a la solución del problema dual esta dada por la representación de la función nucleo, $k(x, x')$, y es llamada el hiperplano de separación óptimo. La solución representada mediante la función núcleo es escrita como una suma ponderada de los vectores soporte. La Figura 3.1 presenta el método SVM, los datos ubicados en el márgenes (de color gris) son los vectores soporte.

Por su parte, las redes neuronales artificiales (NN) se han utilizado exitosamente en un gran número de problemas de clasificación (Duda et al. 2001, Farias et al. 2010, Farias & Santos 2007). Hay una enorme variedad de NN con diferentes estructuras que pueden ser utilizadas dependiendo de las características del problema.

En todos los casos, una NN se compone de unos elementos básicos de procesamiento

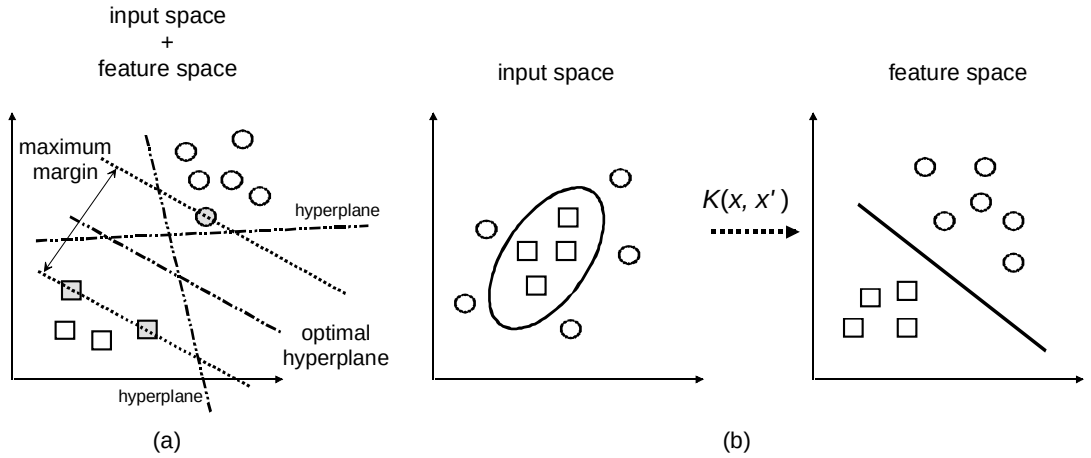


Figura 3.1: La idea de SVM: mapear los datos de entrada en un espacio de características de dimensiones superiores mediante una función núcleo K, y construir un hiperplano de separación con un margen de separación máximo. Esto se traduce en una frontera de decisión no lineal en el espacio de entrada. Mediante el uso de las funciones núcleo, es posible calcular el hiperplano de separación sin llevar a cabo explícitamente el mapeo al espacio de características. (a) caso linealmente separable. (b) caso no linealmente separable.

denominados neuronas, las cuales son agrupadas en capas y conectadas por enlaces o sinapsis que a través de un peso indican el grado de influencia entre las neuronas (Freeman & Skapura 1991, Haykin 2004, Rojas 1996, Hilera & Martínez 1995).

En ambos clasificadores (SVM y NN), la transformada Wavelet (WT) se ha aplicado para la extracción de características. La transformada Wavelet (Daubechies 1992, Mallat 2008, Misiti et al. 2004) es utilizada extensamente en esta Tesis para reducir la dimensionalidad de las señales tratadas sin una pérdida importante de la información. Algunos ejemplos de aplicación de Wavelets en bases de datos de fusión nuclear pueden ser encontrados en (Farias et al. 2004, Dormido-Canto et al. 2004, 2005, Farias & Santos 2005).

El análisis de señales bidimensionales presenta grandes ventajas cuando se utilizan las Wavelets. Debido al hecho de que la descomposición de la WT es multi-escala, las imágenes pueden ser caracterizadas por un conjunto de coeficientes de aproximación y tres conjuntos de coeficientes de detalle (en el sentido horizontal, vertical y diagonal). Los coeficientes de aproximación proporciona el grueso de la información existente en la imagen (los coeficientes de aproximación contienen la mayor parte de la energía de la imagen), mientras que los coeficientes de detalle, con niveles cercanos a cero, contienen información que puede ser relevante en un contexto particular.

En relación a las imágenes del diagnóstico Thomson Scattering, se ha encontrado

en (Farias et al. 2004) que el mejor coeficiente de la WT para extraer características es el detalle vertical, con la Wavelet madre del tipo Haar a nivel 4. Cuando se aplica la WT anterior, las características utilizadas de las imágenes del diagnóstico Thomson Scattering son reducidas desde los 221760 pixeles originales a sólo 900. Así, los 900 atributos obtenidos mediante la transformada Wavelet representan algo menos del 0.4% del tamaño original de la imagen.

Respecto al clasificador implementado con la combinación transformada Wavelet y redes neuronales (WT+NN), la red neuronal utiliza un esquema del tipo *Feed Forward* debido a que el uso de este enfoque en problemas similares se ha mostrado exitoso (Farias & Santos 2007, Farias et al. 2010).

Una de las posibilidades de la red neuronal tipo *Feed Forward* es la utilización de aprendizaje supervisado. Para tal fin, es necesario entrenar la NN indicando a la capa de entrada los atributos de una señal (en este caso la transformada Wavelet de la imagen) y al mismo tiempo proporcionando los valores deseados de la capa de salida (en este caso la clase de la imagen del diagnóstico Thomson Scattering).

La Figura 3.2 muestra la red neuronal que ha generado los mejores resultados. Nótese que la combinación WT+NN tiene una capa de entrada de 900 atributos, los cuales provienen de la etapa de extracción de características debida a la WT. La capa oculta utiliza 90 neuronas, mientras la capa de salida posee sólo 5 neuronas (donde la activación de una neurona de salida indica la clase de la imagen de entrada). En el caso de la capa oculta la función de activación elegida es *Tansig* y en el caso de la capa de salida se seleccionó una función de activación del tipo *Logsig*. Después de la fase de entrenamiento de la red neuronal, cada señal o imagen es asociada con una clase dependiendo de la neurona activada. Este clasificador muestra un porcentaje promedio de acierto del 90.89%.

Respecto al clasificador que combina máquinas de vectores soporte y la transformada Wavelet (WT+SVM), para obtener el hiperplano óptimo, se ha seleccionado una función núcleo lineal con resultados satisfactorios. Nótese sin embargo, que otras funciones núcleo pueden mejorar la tasa de clasificación.

El clasificador WT+SVM presenta un comportamiento bastante robusto alcanzando una tasa del 92.7% en la campaña experimental más reciente del TJ-II (se han conseguido resultados superiores al 98% en campañas anteriores). El clasificador no realizó

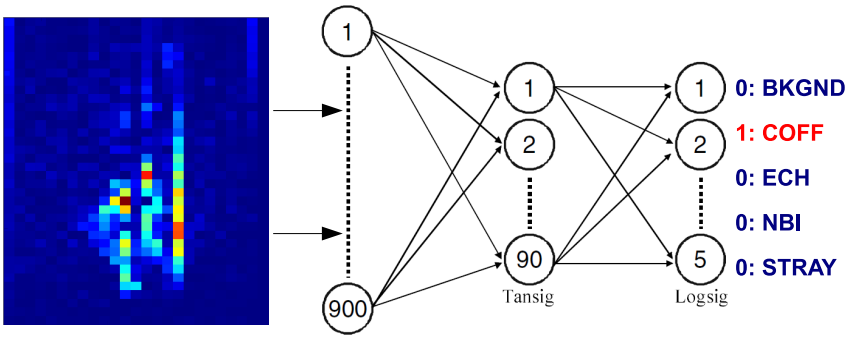


Figura 3.2: Esquema del clasificador WT+NN y estructura de la red neuronal propuesta.

correctamente la asignación de clases al 1.8% de las imágenes, mientras que cerca del 5.5% de los datos se clasifican en más de una clase. Este enfoque ha mostrado una gran robustez en comparación con otras aproximaciones similares, las cuales se basan en características estadísticas de las imágenes del diagnóstico Thomson Scattering.

3.1.2 Clasificación de Diagnósticos y Configuraciones

Debido a que los experimentos de fusión generan cientos de señales, es esencial tener mecanismos automáticos para la búsqueda y recuperación de formas de onda similares en las enormes bases de datos.

De forma similar al diseño de clasificadores para el diagnóstico Thomson Scattering, la transformada Wavelet se utiliza para reducir la dimensionalidad de las señales a tratar. Además tal como se ha comentado previamente, las máquinas de vectores soporte se presentan como un método muy eficiente para generar clasificadores multi-clase en diversos problemas de reconocimiento de patrones. Por estas razones la combinación WT+SVM se ha utilizado para la búsqueda y recuperación de formas de onda similar en las bases de datos del TJ-II (Dormido-Canto et al. 2004, 2005). Por otro lado, la combinación de SVM y el uso de conocimiento *experto* respecto de ciertos parámetros geométricos, se ha utilizado para la identificación automática de configuraciones de plasma en las descargas del JET (Dormido-Canto et al. 2008*a*).

Clasificación de Diagnósticos de Señales Temporales en el TJ-II

Una prueba de la utilidad de estos enfoques fue realizada en la clasificación y reconocimiento de señales de evolución temporal en las bases de datos del TJ-II. Como antes, el problema es tratado en dos fases. Una primera etapa es la encargada de acondi-

cionar la señal (para asegurar el mismo periodo de muestreo, el número de muestras, etc.) y la reducción de la dimensionalidad de la señal mediante la transformada Wavelet (en este caso se utilizó el coeficiente de aproximación, con Wavelet madre del tipo Haar, al nivel 8). La segunda etapa es ejecutada mediante el uso de SVM con diferentes funciones núcleo.

La Figura 3.3 presenta los vectores soporte positivos de 4 clases (ECE7, BOL5, RX306, y Densidad2) utilizando una función núcleo lineal, la señal de entrenamiento original y la señal procesada por la transformada Wavelet.

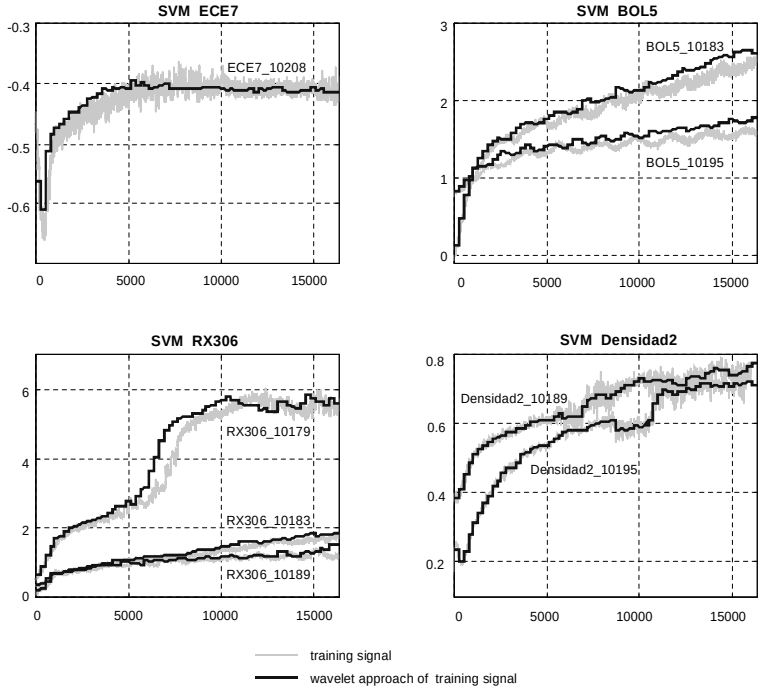


Figura 3.3: Vectores soporte y coeficientes de aproximación de la wavelet para cuatro señales temporales del TJ-II.

Se han construido dos clasificadores multi-clase para 4 y 6 clases diferentes con un 100% (para una función núcleo del tipo RBF) y un 92% (para una función núcleo RBF exponencial) de acierto como mejores resultados, respectivamente (Dormido-Canto et al. 2004, 2005).

Clasificación de Configuraciones de Plasma en JET

La extracción de características de las bases de datos de fusión nuclear no es fácil de implementar debido a que no existe una solución genérica. Normalmente la reducción de la dimensionalidad de una señal, tal como la utilizada por la transformada Wavelet, es un

buen primer intento para ejecutar reconocimiento de patrones en fusión. Sin embargo, la mejor opción siempre será tomar ventaja del conocimiento previo dado por un experto. Este es el caso de la clasificación de configuraciones del plasma en JET, donde los bordes de la superficie externa del plasma pueden ser utilizados para identificar la configuración de diferentes descargas.

La forma de la última superficie es un ingrediente esencial para la determinación de los diversos escenarios de operación del JET que pueden presentarse durante una descarga. Algunos tipos de análisis de datos son sensibles a la configuración del plasma (por ejemplo, a la localización del *punto X* y los puntos de *cuña* o *strike* points) y, por tanto, una identificación o clasificación adecuada de la configuración del plasma es de gran importancia.

Las configuraciones en JET fueron originalmente identificadas mediante el uso de palabras claves que describían, antes de la descarga, el sistema de control del plasma. Esto tiene la desventaja de que es posible encontrar diferentes identificadores que se refieren a la misma configuración, y además de ser incompleto dado que algunas descargas no tienen un identificador, o peor aún, tienen uno erróneamente asignado. Estos problemas han motivado el desarrollo de un clasificador automático.

Se han desarrollado dos sistemas de clasificación exitosos que se basan en el uso de parámetros geométricos de la última superficie (Dormido-Canto et al. 2008*a*).

El primer sistema de clasificación se implementó para discriminar descargas que pertenecen a las siguientes tres clases: VH_3M5_HT, HIXR_GB, y SEPTUM. Un vector de características de dos dimensiones y una función núcleo lineal fue suficiente para clasificar el 100% de las 102 configuraciones probadas. El segundo clasificador fue implementado para el reconocimiento automático de 8 clases. Debido a que este problema es más complicado que el anterior, el vector de características fue ampliado para incorporar un total de 12 parámetros geométricos. Los resultados del clasificador son bastante prometedores, ya que la tasa de acierto de las configuraciones clasificadas alcanzó el 96% en promedio para las 8 clases.

3.1.3 Agrupamiento de Señales Temporales de Diagnósticos

Los diagnósticos proporcionan señales de evolución temporal que traducen las propiedades físicas del plasma. Así, se puede asumir que señales similares corresponden a compor-

tamientos similares del plasma y, por tanto, es posible encontrar patrones representativos para las mismas condiciones físicas.

La física del plasma puede ser descrita por los diferentes tipos de señales adquiridos (densidad, temperatura, rayos X suaves, bolométricas, etc.) de cada pulso. Un método para encontrar formas de onda similares de cada tipo de señal podría ayudar a revelar, de manera automática, un conjunto de descargas que muestran comportamientos equivalentes.

El objetivo de este problema es por tanto clasificar formas de onda en un número de categorías (o grupos) y luego aplicar medidas de proximidad para evaluar la similitud entre las señales adquiridas. El método de agrupamiento es responsable de "revelar" la organización de las señales en grupos "semejantes". Esto es comúnmente conocido como clasificación no supervisada o simplemente agrupamiento.

Agrupamiento de Diagnósticos

El artículo (Duro et al. 2006) muestra la experiencia de aplicar agrupamiento a las bases de datos del TJ-II. En este trabajo se seleccionaron 194 formas de onda de descargas correspondientes a las siguientes señales: *emisión Hα* , *línea media de densidad electrónica*, *bolométricas* y *rayos X suaves*. Todas las señales fueron pre-procesadas con el fin de analizar los datos en la misma ventana de tiempo (258ms) y con el mismo periodo de muestreo (10 μs). Antes de aplicar el agrupamiento, las 4 señales han sido tratadas por dos extractores de características: transformada Wavelet (Haar al nivel 8, y 64 coeficientes de aproximación) y la transformada de Fourier (los 24 primeros coeficientes).

El agrupamiento se ha realizado mediante el uso de cuatro técnicas: *Jerárquica*, *K-means*, *Teoría de resonancia adaptativa (ART)*, y *Gran Tour (GT)*.

La técnica Jerárquica (Johnson 1967) comienza asignando cada ítem o muestra a un grupo, para posteriormente unir pares de grupos similares, con el fin de reducir en uno el número de grupos. El proceso continua hasta que se obtiene un único grupo final.

Respecto a K-means (MacQueen et al. 1967), ésta comienza con un número fijo de grupos, luego se asigna cada ítem al grupo que tiene el centroide más cercano. Cuando todas las muestras han sido asignadas, los centroides son recalculados. El proceso termina cuando ningún centroide es modificado.

La teoría de resonancia adaptativa (Carpenter & Grossberg 1993) es aplicada por

una red neuronal para el desarrollo de un tipo de aprendizaje competitivo. En este caso, cuando la información es proporcionada a la entrada de la red, sólo una neurona de salida es activada. La idea es "resonar" la información de entrada con prototipos de clases que la red reconoce.

El Grand tour es una técnica interactiva de visualización de un conjunto de datos multi-dimensionales que permite examinar la estructura de la información a través de gráficos dinámicos. La idea es proyectar los datos de varias dimensiones a un plano que al ser rotado se puede observar desde diversos ángulos para buscar "estructura" en los datos. Tal estructura se define como la no normalidad (Martinez & Martinez 2001) e incluye aspectos tales como agrupamientos, estructuras lineales, agujeros, etc.

Cada uno de los métodos de agrupamiento (jerárquico, K-means, ART y GT) proporciona un conjunto de grupos. Sin embargo, en este trabajo sólo se consideró grupos que incluyeran al menos un 5% de las formas de onda. Los grupos con un número menor al 5% son agrupados en un grupo misceláneo.

Los resultados analizaron el número de grupos encontrados y el porcentaje de señales incluidas en cada uno (Duro et al. 2006). En primer lugar, se debe observar que se han obtenido resultados equivalentes sin extracción de características. En particular, el uso de las técnicas Jerarquica, ART y GT, muestran porcentajes muy similares no sólo en los grupos principales sino también en el grupo misceláneo. Se puede observar en los resultados, que al menos el 50% de las señales pertenece al mismo agrupamiento. La inspección de los resultados de K-means muestran que esta técnica produce un resultado diferente: se genera un mayor número de grupos. Además, el número de señales en cada grupo es menor. Analizando las señales que constituyen estos grupos se puede concluir que las señales para dos o tres grupos (dependiendo de la experiencia) en el método K-means son integrados en un grupo mayor en el caso de las otras tres técnicas de agrupamiento.

Los diferentes métodos agrupan las misma señales en los mismos grupos, independientemente de las características utilizadas. A grandes rasgos, todas las familias proporcionan dos agrupamientos. En primer lugar, el grupo mayor simboliza que la mayoría de las señales traducen un comportamiento físico promedio de la propiedad medida del plasma. En segundo lugar, el resto de las formas de onda pueden integrarse en un grupo simple. Este último grupo incluye comportamientos no estándar y, por tanto, las señales

que son clasificadas en este grupo revelan propiedades o comportamientos del plasma fuera de lo normal. Este hecho ayuda a quienes utilizan los diagnósticos para poder encontrar, de una forma automática, datos interesantes para ser analizados, en vez de tener que buscarlos de forma manual. Se pueden encontrar más detalles de la aplicación de técnicas de agrupamiento en bases de datos del TJ-II en (Duro et al. 2006).

Agrupamiento y Modelado de Diagnósticos

El artículo (Martín et al. 2009) muestra también la experiencia de aplicar técnicas de agrupamiento y lógica difusa a las bases de datos del TJ-II. El trabajo considera las señales descritas en la Tabla 2.1. La técnica de agrupamiento propuesta esta basada en el método de partición. La estrategia consiste en la generación de una matriz triangular con valores de una medida matemática de similitud, el producto escalar normalizado (NSP), que considera un par de formas de onda y la aplicación de un umbral para generar agrupamientos dinámicos. A partir de otros trabajos que consideran las señales tratadas en el artículo (Dormido-Canto et al. 2004, 2006, Duro et al. 2006, Farias et al. 2006, Martín et al. 2007), se puede concluir que el procedimiento más eficiente para la medida de similitud en tiempo real de las señales de fusión en el TJ-II es NSP. En este trabajo no se realiza ningún proceso de extracción de características.

La información proporcionada por el método de agrupamiento puede ser aplicada también para obtener un modelo representativo de cada clase de señal mediante el uso de diferentes enfoques de modelado. De esta forma, los patrones representativos de cada grupo son obtenidos, lo cual hace posible detectar anomalías o eventos físicos extraños. Para propósitos de modelamiento se han utilizados dos métodos: identificación Neuro-difusa e identificación basada en el tiempo.

El objetivo del modelado neuro-difuso y las estrategias en el dominio del tiempo es identificar grupos naturales de datos de un enorme conjunto de información que proporciona una representación concisa de un tipo de señal.

A partir de los grupos generados, se obtiene un modelo mediante el uso de sistemas de inferencia difusa (Jang 1993). El modelo se utiliza entonces para detectar posibles eventos inesperados o anómalos.

Después de la realización de diversos experimentos, se puede concluir que existe siempre un grupo estable de formas de ondas que incluyen al menos el 75% de las

señales. El resto de datos se agrupa en otros conjuntos de menor tamaño. Es claro que el uso del agrupamiento permite identificar fácilmente si el comportamiento de una descarga es estándar o fuera de lo normal. Además, el uso de modelos puede ser utilizado para reducir el espacio de búsqueda de comportamientos equivalentes, debido a que la similitud de una señal es calculada sobre un conjunto de datos parecidos. Más detalles acerca de los resultados y experimentos llevados a cabo en este trabajo se pueden encontrar en (Martín et al. 2009).

3.1.4 Clasificación y Agrupamiento de ELMs

Los modos localizados en el borde (ELMs) son inestabilidades del plasma que pueden afectar a la cámara interna del dispositivo de fusión. Aunque se han realizado muchos avances desde los puntos de vista teórico y experimental, aún no hay una comprensión total del comportamiento de los ELMs. Con el fin de avanzar en el estudio de la física de los ELMs, el enfoque de reconocimiento de patrones parece ser una poderosa herramienta para extraer conocimiento a partir de las señales experimentales. Este conocimiento podría combinarse con modelos teóricos ya sea para realizar análisis exploratorios o para confirmar hipótesis. En (Duro et al. 2009) se desarrolló un enfoque para la caracterización y clasificación automática de ELMs de tipo I y tipo III (Liang 2011, EFDA 2013*b*, Saibene et al. 2002, Bellizio et al. 2011). Para tal fin, se realizan tres etapas. La primera fase consiste en identificar, aislar y extraer ELMs individuales de señales del JET (cada ELM individual es analizado en vez de considerar un segmento temporal que contenga muchos ELMs). La segunda etapa realiza la extracción de características del proceso para representar los ELMs con un conjunto mínimo de características relevantes. Finalmente, se aplican tres métodos de clasificación (supervisados y no supervisados) para identificar los ELMs.

El reconocimiento y aislamiento de un ELM es llevado a cabo utilizando tres señales: energía diamagnética acumulada (correspondiente a la señal MG3F/WPD de JET), la línea integrada de densidad electrónica (señal KG1V/LID4 de JET), y la Dα (señal S3AD/ AD34 en JET).

Los ELMs son reconocidos por un cambio abrupto en la energía diamagnética y una caída simultanea en la línea integrada de densidad electrónica. Como una consecuencia de la inestabilidad del ELM, una forma típica de pico aparece en Dα. Se pueden observar

formas típicas de picos en las 2.6 y Figuras 2.7.

Respecto al proceso de extracción de características, tres atributos se han considerado a partir de una inspección visual: La caída desde el pico del ELM, el periodo de cada ELM individual, y medida de cresta (la cual es una característica que mide la forma del ELM). Las primeras dos características son calculadas mediante las señales de energía diamagnética, mientras que la medida de cresta es obtenida desde la señal de emisión Dα.

Posteriormente al proceso de extracción de características, se hace uso de las técnicas de clasificación y agrupamiento de los ELMs en dos clases: tipo I y tipo III. El conjunto de entrenamiento esta compuesto por 122 ELMs individuales (97 de tipo I y 25 de tipo III). El conjunto de prueba considera un total de 143 ELMs aislados de las señales del JET. El método de aprendizaje supervisado seleccionado fue implementado mediante máquinas de vectores soporte. Para el caso del agrupamiento (o método de clasificación no supervisado), las técnicas K-means y Jerárquico han sido utilizados.

Independiente del método de clasificación (supervisado o no), los resultados obtenidos son muy similares. En ambos casos el número de clases que proporcionan los mejores resultados son dos. Además, las tasas de acierto (sobre el 93% en la mayoría de los casos) con diferentes técnicas permiten concluir que la extracción de características adoptada es bastante robusta. En particular, utilizando los métodos K-means y Jerárquico, el porcentaje de señales incluidas en cada grupo es el mismo.

Los resultados también muestran que los ELMs de tipo I tienen una caída mayor que la de tipo III. Sin embargo, esta característica no es suficientemente buena para diferenciar completamente ambos tipos de ELMs. Así mismo, el periodo tampoco es una característica distintiva del proceso de clasificación. La que sí parece ser una característica que distingue claramente ambos tipos de ELMs es la medida de la cresta. Más detalles acerca de este enfoque pueden ser encontrados en (Duro et al. 2009).

3.2 Búsqueda y Recuperación de Información

Comportamientos físicos diferentes del plasma son descritos por diferentes señales adquiridas por los diagnósticos. En general, un mapeo lineal puede ser establecido para conectar la evolución temporal de un fenómeno físico con el tipo de señal que éste genera. Por

tanto, es posible hablar acerca de la existencia de patrones en las señales digitalizadas. Para analizar las propiedades del plasma, la búsqueda de patrones puede resultar muy útil. Sin embargo, una base de datos de un dispositivo experimental de fusión contiene miles de señales, de modo que la recuperación automática de la información es una necesidad importante.

3.2.1 Búsqueda de Formas de Onda Completas

En el artículo (Farias et al. 2006), se describe una técnica para la búsqueda automatizada de señales temporales similares a una forma de onda de referencia (señal de entrada). El procedimiento se divide en tres fases. En la primera fase, se realiza un proceso de extracción de características. El paso posterior consiste en la ejecución de un sistema de clasificación que reduce el espacio de búsqueda. La técnica descrita finaliza con métodos de consulta basadas en medidas de similitud para encontrar formas de onda similares a una señal de entrada. La técnica es aplicada a series temporales provenientes de los diagnósticos del TJ-II.

Como se ha comentado anteriormente, la transformada Wavelet hace posible alcanzar un nivel de descomposición deseado preservando la información de la señal. La información redundante es minimizada y la carga computacional es sustancialmente reducida. Así la primera etapa hace uso de WT para disminuir el ruido de cada diagnóstico. Además, las señales procesadas mediante WT son reducidas desde 16384 muestras a tan solo 64. Algunos trabajos previos, ver (Rafiei & Mendelzon 1998, Nakanishi et al. 2004, 2006), han considerado la aplicación de la transformada discreta de Fourier, pero dado que las formas de onda de fusión nuclear no presentan un comportamiento estacionario, el uso de WT parece ser una mejor opción para la caracterización de los datos.

Después de extraer características, el proceso de búsqueda comienza. Este procedimiento implica el desarrollo de las siguientes dos etapas. La primera etapa se utiliza para reducir el espacio de búsqueda, es decir, para limitar la búsqueda de las señales a un subconjunto adecuado de la base de datos. Esto es llevado a cabo mediante un clasificador SVM, el cual ha sido previamente entrenado para distinguir entre los diferentes tipos de señales.

Una vez que el espacio de búsqueda es reducido por el uso de un sistema de clasificación, la segunda etapa consiste en encontrar las señales más parecidas a la señal de

referencia. Para este propósito dos técnicas pueden ser aplicadas: Distancia euclidiana y envolvente. La distancia euclidiana simplemente calcula la distancia muestra a muestra de la señal de entrada versus el resto de la base de datos. El método de la envolvente está basado en la construcción de dos bandas alrededor de una señal (banda superior e inferior). Una medida de distancia puede ser hecha mediante el conteo del número de muestras que esta fuera de las bandas. Ambos métodos de búsqueda de similitud pueden utilizarse para obtener la distancia mínima entre la señal de referencia y la base de datos, y por tanto para encontrar las señales más parecidas.

Los experimentos realizados mostraron que el método de la envolvente es más robusto que la técnica de la distancia euclidiana. Esto se debe a que la distancia euclidiana tiende a acumular error, mientras que el método de la envolvente considera por igual los puntos que están fuera de las bandas superior e inferior, no importando su distancia a las bandas. Más detalles pueden ser encontrados en (Farias et al. 2006).

Otro enfoque para buscar señales similares se ha desarrollado en el artículo (Vega et al. 2008). Este trabajo describe una técnica mediante la cual, dada una forma de onda, es posible encontrar señales similares en una gran base de datos de una forma rápida y automatizada. El método utiliza una señal de entrada para buscar las formas más similares dentro de una base de datos. La técnica esta basada en el desarrollo de un sistema de clasificación de estructura arborescente que agrupa las formas de onda en grupos de acuerdo a ciertas reglas. El agrupamiento de las formas de onda es el elemento esencial para acelerar la búsqueda y para reducir las necesidades de cómputo. El proceso de búsqueda llevado a cabo es un método de comparación uno a uno pero sólo de aquellas formas de onda al interior de un grupo, en vez de realizarlo entre todas las señales de la base de datos.

El proceso de búsqueda de las señales más similares en este trabajo se divide en cuatro pasos. Dada una forma de onda, el primer paso consiste en la extracción de características. A continuación el vector de características es clasificado en uno de los agrupamientos existentes. El tercer paso es el cálculo del factor de similitud entre el vector de características de entrada y el resto de vectores de características del grupo. Finalmente, las formas de onda son ordenadas de acuerdo a una medida de distancia en orden descendente.

El valor absoluto del producto interno normalizado ha sido utilizado como función

de similitud. Nótese que las formas de onda similares no implica que sean casi iguales (similitud cerca a 1). Así, algunas ventajas de este método sobre la distancia euclidiana o la envolvente provienen del hecho de que el método no depende de la amplificación de la ganancia (es decir, formas de onda con diferente factor de ganancia son reconocidas como similares) o de la polaridad de la señal (es decir, las formas de onda invertidas son consideras iguales). El método encuentra las señales más similares.

Los resultados de este enfoque muestran que el sistema es capaz de encontrar formas de onda (bolométricas y rayos X suaves) en un par de segundos de entre cientos de descargas. Más detalles pueden ser encontrados en (Vega et al. 2008).

3.2.2 Búsqueda de Patrones dentro de Formas de Ondas

El análisis visual de los datos es una herramienta crucial en la física del plasma. Una simple inspección visual de las señales puede ser suficiente para reconocer una evolución típica del plasma o para distinguir la presencia de eventos físicos interesantes. Un investigador identifica el comportamiento del plasma mediante la detección de patrones dentro de las formas de onda: saltos, cambios inesperados de amplitud, picos abruptos, o componentes sinusoidales. Por ello un desafío importante es la creación de medios de identificación de patrones *dentro de formas de onda*.

Existen algunos trabajos previos de técnicas de reconocimiento de patrones para la búsqueda en bases de datos de fusión. En los enfoques mostrados anteriormente, los esfuerzos se han centrado en la búsqueda de formas de ondas completas, es decir en donde las señales cubren la vida completa del plasma. El trabajo pionero descrito en (Nakanishi et al. 2006) se orienta a la búsqueda de patrones dentro de las formas de onda, para lo cual utiliza la componente de frecuencia más importante de la señal. Sin embargo, dado que las señales en fusión son principalmente no estacionarias, se requiere realizar la búsqueda de patrones de manera más general.

La búsqueda de patrones en el interior de la forma de onda se ha considerado en los artículos (Dormido-Canto et al. 2006, 2008*b*, Rattá et al. 2008, Vega et al. 2007). Para este propósito se ha seleccionado el enfoque sintáctico. El método sintáctico considera que los patrones están compuestos por subpatrones más simples (Fu & Albus 1982). Los subpatrones más elementales son conocidos como primitivas. Un patrón complejo es por tanto expresado en términos de relaciones entre las primitivas. Se pueden establecer los

fundamentos del reconocimiento de patrones sintáctico mediante una analogía entre las estructuras de patrones y la teoría de lenguajes formales. Los patrones representan las oraciones en un lenguaje, mientras que las primitivas consituyen el alfabeto del lenguaje. Una gramática de un lenguaje genera e identifica las oraciones pertenecientes al lenguaje que emplea sus reglas. Sin embargo, en algunos casos el uso de gramáticas no es adecuado debido a que los patrones carecen de regularidades. En tal caso, el enfoque estructural para el reconocimiento de patrones se puede adoptar.

En el reconocimiento estructural de patrones, las primitivas son representadas por cadenas de texto. Consecuentemente, el problema de reconocimiento se transforma en un problema de emparejamiento de patrones. Por ejemplo, dado un patrón descompuesto en primitivas (conjunto de caracteres de texto), el objetivo final es encontrar los patrones más similares de una base de datos de cadenas de texto.

La Figura 3.4 muestra las etiquetas de las primitivas utilizadas para codificar una forma de onda. La clasificación de los ángulos proporcionan toda la información estructural elemental requerida para construir subpatrones más complejos en las formas de onda. Nótese que las formas de onda son divididas en segmentos de longitud fija.

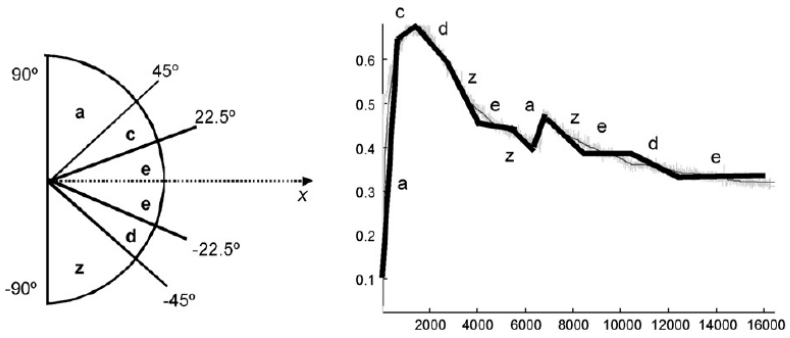


Figura 3.4: Codificación de una forma de onda con primitivas: En la izquierda de la figura se muestran las primitivas y etiquetas utilizadas para clasificar los ángulos. A la derecha se proporciona una forma de onda codificada con las primitivas.

Los tipos de búsqueda mencionados anteriormente pueden ser aplicados con un administrador de bases de datos relacionales mediante el uso del lenguaje de consulta (SQL). Para propósitos experimentales, las bases de datos fueron implementadas con la base de datos MICROSOFT ACCESS^{TM} (Feddema 2001). El algoritmo de recuperación de información se ejecuta como sigue: Inicialmente el usuario selecciona una descarga, y luego escoge una sección de la señal (patrón), y solicita a la aplicación un tipo de búsqueda. La aplicación, desarrollada en MATLAB, lleva a cabo el pre-procesamiento

y cómputo de las primitivas del patrón. Posteriormente, se realiza una consulta SQL de acuerdo al tipo de búsqueda seleccionado. Finalmente, ACCESS envía de vuelta a MATLAB los resultados SQL, y la aplicación muestra todos los emparejamientos encontrados en las señales devueltas.

La Figura 3.5 muestra un ejemplo de la técnica desarrollada para buscar patrones en el interior de formas de onda.

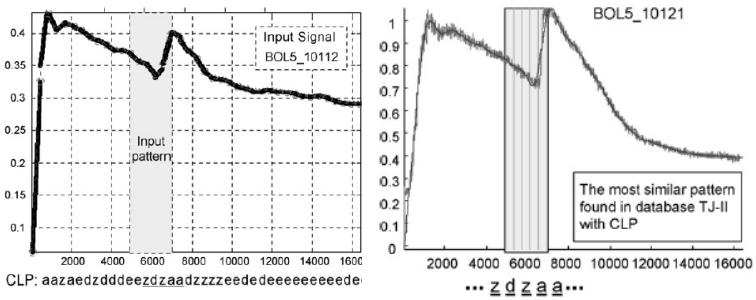


Figura 3.5: Ejemplo de recuperación de información mediante reconocimiento estructural de patrones.

Más detalles y variantes de este enfoque aplicado a bases de datos del TJ-II y del JET pueden ser encontrados en (Dormido-Canto et al. 2006, 2008*b*, Vega et al. 2007, Rattá et al. 2008).

3.2.3 Detección de Tiempos de Transición L-H

Para determinar de manera automatizada los instantes en que ocurre la transición L-H en el dispositivo de fusión nuclear DIII-D, se han desarrollado algunos métodos de reconocimiento de patrones y aprendizaje automático (Farias et al. 2012). Un conjunto de entrenamiento se utiliza para generar un modelo no paramétrico que distingue los modos de confinamiento L y H en cualquier instante de tiempo de vida de la descarga. El único requerimiento para crear el modelo es asumir que todas las muestras son independientes y que son idénticamente distribuidas de acuerdo a una función de distribución fija pero desconocida. El modelo también proporciona una incertidumbre (barra de error) en la predicción del tiempo de transición. Para este fin se utilizan los predictores conformales. Los predictores conformales proporcionan sus predicciones acompañados de valores que indican la confianza y credibilidad de la estimación, la cual proporciona información acerca de cuan preciso y fiable son las predicciones (Vovk et al. 2005).

El sistema es implementado en dos pasos dentro de un sistema distribuido de com-

putación. En primer lugar, un modelo SVM multi-capa es creado mediante un conjunto de datos de entrenamiento. El modelo SVM utiliza una combinación de varias señales que determinan la transición L-H. La selección del conjunto de datos se logra de forma automatizada mediante el uso de una técnica de reconocimiento de patrones morfológica (MPR) sobre la señal emisión Dα. El algoritmo morfológico busca una caída rápida en la señal simplemente utilizando información estructural de la forma de onda (González et al. 2010). La transformada Wavelet y la técnica de regresión de máquinas de vectores soporte son las técnicas empleadas en el proceso completo del algoritmo MPR.

Posteriormente, el modelo SVM y el algoritmo MPR son combinados para predecir de manera separada los tiempos de transición L-H para nuevas descargas. La Figura 3.6 describe el uso del clasificador SVM para estimar los tiempos de transición L-H.

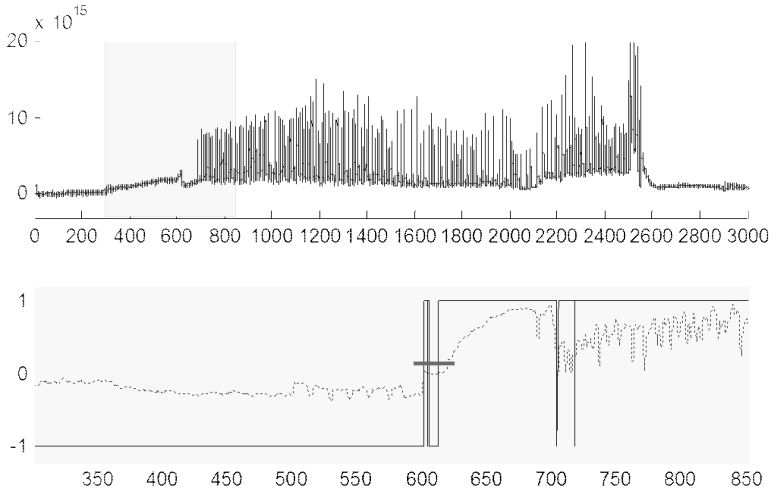


Figura 3.6: Predicción de transición L-H con SVM: En la parte superior de la figura se muestra un intervalo de la señal de emisión Dα donde la transición L-H ocurre aproximadamente a los 600ms. En la parte inferior de la figura se presenta la distancia de cada muestra al hiperplano óptimo del modelo SVM, y la clasificación de las mismas muestras mediante el modelo SVM para el intervalo de interés.

Con el fin de predecir la transición L-H, las predicciones del algoritmo MPR y el modelo SVM multi-capa han sido combinados. Por una parte, el algoritmo morfológico considera la señal emisión Dα y la señal de potencia inyectada para predecir la transición. Por otro lado, el modelo SVM multi-capa ejecuta la predicción utilizando las señales o características seleccionadas. Nótese que el modelo SVM se focaliza en la búsqueda de la transición solamente en la zona anterior a la aparición de los ELMs. Si la diferencia de ambas predicciones (MPR y SVM) son menores a los 100ms entonces la predicción final será dada por el algoritmo MPR, de lo contrario la predicción utilizada está determinada

por el modelo SVM. La Figura 3.7 muestra el esquema adoptado para predecir las transiciones L-H en el dispositivo de fusión DIII-D.

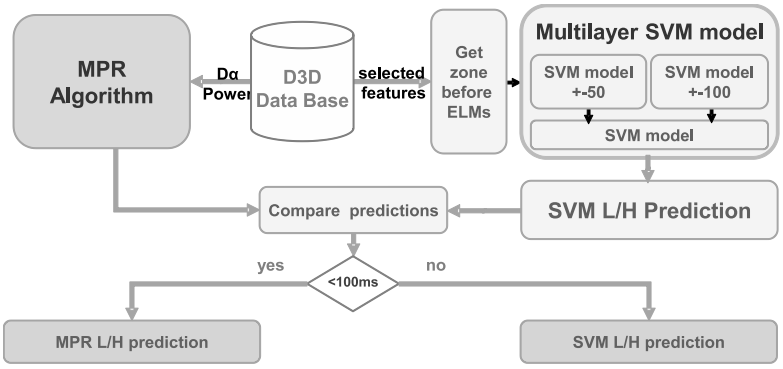


Figura 3.7: Combinación del algoritmo MPR y el modelo SVM para predecir los tiempos de transición L-H.

El predictor ha sido probado con un conjunto inicial de 354 descargas. El predictor combinado tiene un promedio de error de 6 ms y una desviación estándar de 49 ms. La tasa de éxito es del 95.6%. La Figura 3.8 muestra el histograma con la frecuencia del error de predicción.

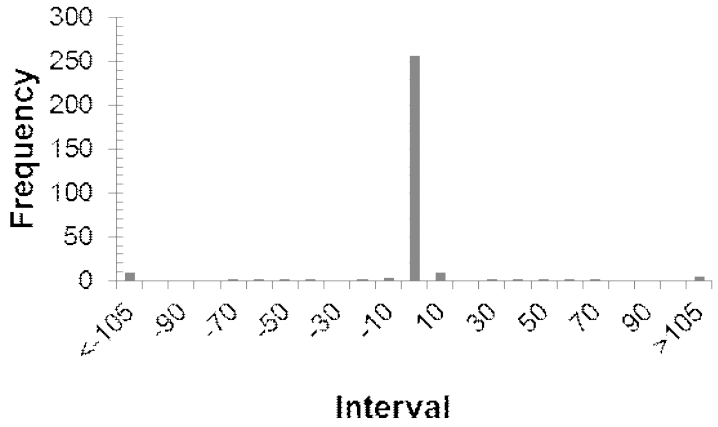


Figura 3.8: Histograma de predicción del error del sistema MPR + SVM.

3.3 Reducción de Ruido

En esta Tesis también se ha considerado la reducción de ruido en el diagnóstico Thomson Scattering del TJ-II. Como se mencionó anteriormente, la cámara CCD del diagnóstico Thomson Scattering adquiere imágenes (espectros de la luz laser dispersada por el plasma) corrompidas por la existencia de luz parásita que, en algunos casos, puede producir perfiles de temperatura y densidad no fiables. Un ejemplo es la luz proveniente

del laser de rubí que alcanza el espectrómetro y que no es posible distinguirla de la luz dispersada por los electrones. Hasta ahora, diferentes técnicas que utilizan hardware se han implementado para remover o disminuir la contribución de la luz parásita, pero sólo con un éxito parcial. Entre otros se pueden mencionar en este sentido el uso del filtro *notch* en frente del espectrómetro.

El ruido en las imágenes puede ser reducido mediante la aplicación de diversas técnicas clásicas y avanzadas tales como filtros pasa bajo o la transformada Wavelet. Sin embargo, en algunos casos la presencia del ruido no es global, sino más bien localizada en zonas particulares de la imagen. Ante esta situación, la aplicación de filtros con efecto global sobre la imagen completa no resultan adecuados debido a que el ruido y la información son reducidos en la misma proporción. La alternativa a las técnicas *globales* provienen de la teoría de segmentación de regiones o imágenes. En los últimos años, existe un gran interés en el uso de resultados de los algoritmos de segmentación de regiones que ayudan a obtener segmentación de objetos con un nivel de precisión de píxeles (Fulkerson et al. 2009, Gu et al. 2009). La segmentación es utilizada para subdividir una imagen en un conjunto de regiones. Estas regiones no tienen una forma predefinida como bloques rectangulares, más bien sus bordes son irregulares dentro de la imagen, de manera que algunas propiedades de forma y borde se pueden utilizar como método de extracción de características. Otra ventaja del uso de regiones es la escalabilidad y las potenciales eficiencias desde el punto computacional que se pueden obtener. Las regiones usualmente proporcionan un conjunto de hipótesis mucho más pequeño para analizar que el obtenido mediante el enfoque clásico de ventana deslizante. Así, este trabajo emplea el enfoque de extracción de regiones con componente conectados (ERCC) con el fin de eliminar algunas regiones de la imagen asociadas al ruido de luz parásita.

El método ERCC esta basado en la teoría de segmentación. La segmentación propiamente se refiere al proceso de particionar una imagen digital en múltiples segmentos (conjuntos de pixeles). De forma más precisa, la segmentación de una imagen es el proceso de asignar una etiqueta a cada pixel de modo que los pixeles con la misma etiqueta comparten características visuales comunes. En general, una segmentación de una imagen es una partición en subimagenes (regiones) conectadas R_1, R_2,..., R_n tal que todas las regiones son disjuntas, y la unión de todas ellas reconstruyen la imagen. Cada

subimagen satisface un predicado de forma que todos los pixeles en una subimagen R_i no deben diferir a lo más que un Δ_x de niveles de gris. Todos los pixeles en cualquier subimagen R_i deben estar unidos por un factor de conectividad.

El procedimiento propuesto para remover el ruido mediante el uso de componentes conectadas se muestra en la Figura 3.9.

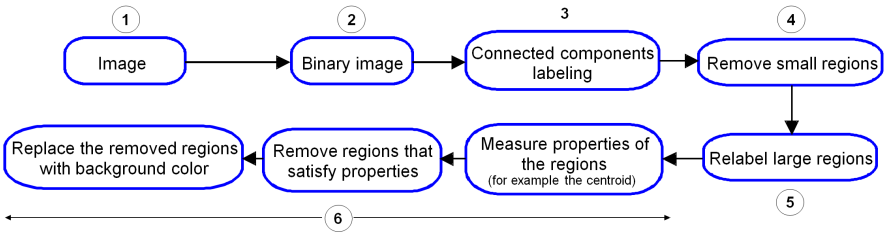


Figura 3.9: Diagrama de flujo para la extracción de regiones con componentes conectadas.

El enfoque se inicia con la conversión de la imagen original en una versión binarizada (pixels de 1s y 0s) a través del uso de un umbral. Luego, las regiones de la imagen binaria son etiquetadas como regiones, donde los 1s pertenecen a la misma región si éstos están conectados (es decir, son vecinos) en cualquier dirección. El próximo paso descarta aquellas regiones que tienen un número menor a una determinada cantidad de píxeles (regiones pequeñas), a continuación las regiones restantes son re-etiquetadas. Posteriormente, todas las regiones cuyos centroides sean inferiores a 100 (es decir, se ubiquen en la parte izquierda de la imagen) son eliminadas. Las regiones eliminadas son reemplazadas por los valores de los píxeles de la parte simétrica derecha de la imagen. Por último, las regiones pequeñas descartadas son restauradas e incorporadas a la imagen final.

La Figura 3.10 muestra el resultado para cada etapa del procedimiento descrito anteriormente para una imagen del tipo NBI. Más información puede encontrarse en (Dormido-Canto et al. 2012).

El método ERCC puede reducir significativamente la luz parásita en las imágenes del diagnóstico Thomson Scattering, sin emargo el predicado de conexión para un pixel es en ocasiones muy fuerte. Por ejemplo, píxeles muy cercanos, pero no conectados a la región detectada como luz parásita, no son considerados como ruido en este enfoque. Aquí es donde el método de crecimiento de región (RG) es muy útil. Esta técnica permite añadir píxeles a una región mediante un predicado personalizado. Así, un pixel podría ser considerado parte de una región aunque la localización del pixel sea cercana pero no

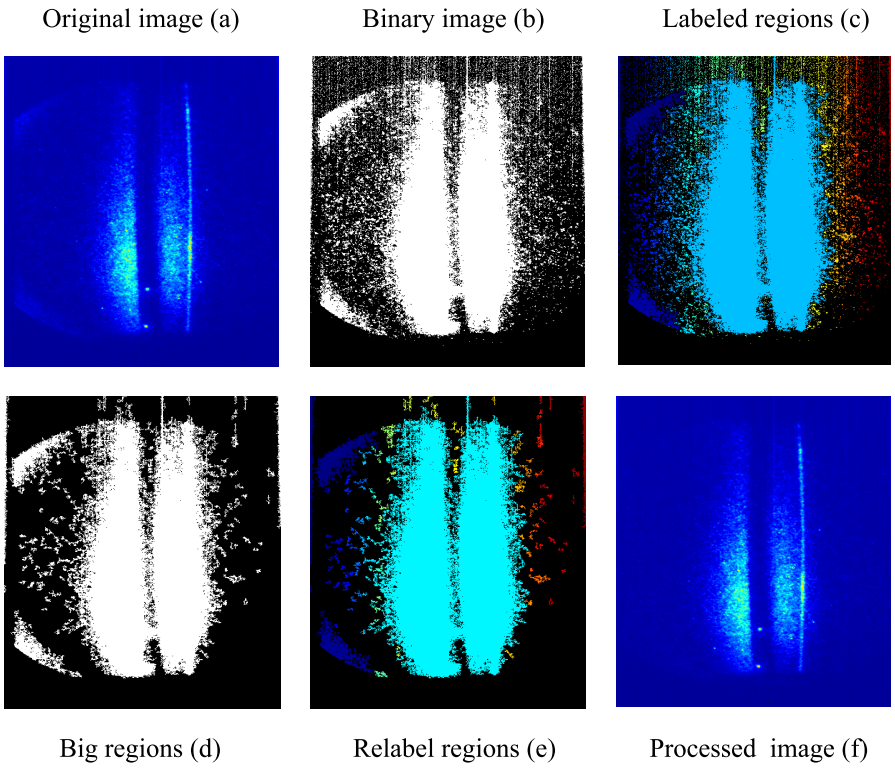


Figura 3.10: Las imágenes muestran las etapas del algoritmo ERCC: (a) imagen original, (b) imagen binarizada, (c) imagen etiquetada, (d) imágenes con regiones más grandes que P pixeles, (e) imagen re-etiquetada, y (f) finalmente la imagen sin ruido.

"conectada" a la región.

Con el fin de validar la efectividad de la reducción del ruido en una imagen, se ha definido una función de eliminación de ruido. Los resultados de ambos enfoque, ERCC y RG, se muestran en la Tabla 3.1.

Tabla 3.1: Tasas de éxito de eliminación de ruido para los algoritmos ERCC y RG.

	BKG	STR	NBI	COFF	ECRH
ERCC	98%	96%	91%	97%	94%
RG	98%	97%	95%	97%	96%

El proceso de validación ha sido probado para 1146 imágenes del diagnóstico Thomson Scattering. Nótese que los resultados del método RG son ligeramente mejores al algoritmo ERCC. En el caso de las imágenes NBI, la diferencia se debe principalmente a que ERCC no es capaz de reducir la luz parásita que está "conectada" a la parte central de la imagen (la que se considera como información significativa y no se puede eliminar). Más detalles de la evaluación de ambos métodos de reducción de ruido se pueden encontrar en (Farias et al. 2013).

Capítulo 4

Conclusiones y Trabajos Futuros

4.1 Conclusiones

Hay muchos dispositivos experimentales de fusión orientados al estudio del proceso de fusión nuclear. Cada experimento produce miles de señales, con un enorme conjunto de datos. Por ejemplo en JET cada descarga, de unas cuantas decenas de segundos, genera alrededor de 10GB de datos. Más aún, en ITER la cifra podría llegar hasta 1 TB. Sin embargo, no toda la información es procesada. Es más, se estima que sólo el 10% de los datos son procesados.

Debido a que los experimentos en fusión generan una ingente cantidad de información, es esencial tener mecanismos automáticos para el análisis y procesamiento de las bases de datos.

Las primeras aplicaciones realizadas en esta Tesis involucraron la clasificación y agrupamiento de señales temporales e imágenes.

El desarrollo del clasificador Thomson Scattering mostró el potencial del enfoque que combina la transformada Wavelet y las máquinas de vectores soporte sin la necesidad de conocimiento experto. Por una parte, las Wavelets permiten reducir la alta dimensionalidad de las imágenes Thomson Scattering. Por otro lado, el uso de SVM permite obtener un sistema de clasificación robusto con una elevada tasa de acierto. Aunque se obtuvieron resultados similares con una combinación de Wavelets y redes neuronales, la característica de SVM respecto a la minimización del riesgo estructural es un particularidad interesante. Este atributo de SVM permite obtener reglas de decisión genéricas, sin sobreajuste, lo cual es una propiedad deseable de cualquier algoritmo de aprendizaje.

Además, los tiempos de entrenamiento y prueba son normalmente inferiores en las SVM comparados con los de NN. Una desventaja de SVM con respecto a NN podría ser la dificultad de encontrar una función núcleo adecuada.

La combinación de la transformada Wavelet y las máquinas de vectores soporte también se ha aplicado a la clasificación de señales temporales. A partir de la observación de varios experimentos, el método WT+SVM se presenta como una opción bastante fiable y eficiente, donde además los resultados presentaron unas tasas de acierto cercanas al 100% para un clasificador de 4 clases en los diagnósticos del TJ-II. Un enfoque similar fue desarrollado para clasificar configuraciones de plasma en JET. Aunque una simple inspección visual de las características geométricas de las configuraciones podría ser utilizada para discriminar las diferentes clases, el análisis se complica si las categorías consideradas se elevan. Aquí se vuelve necesario el uso de SVM. El clasificador así entrenado es capaz de obtener tasas superiores al 90% cuando se tienen 8 clases.

La localización de eventos físicos del plasma también ha sido tratada en esta Tesis. Un ejemplo de esto es la detección automática de transiciones L-H en el DIII-D. El sistema fue entrenado en los supercomputadores del CIEMAT y posteriormente operado en los servidores del DIII-D. El predictor de transiciones fue probado para un conjunto inicial de 354 descargas. El predictor desarrollado, una combinación de un algoritmo morfológico y un clasificador SVM, logró alcanzar un error promedio de estimación de 6ms y una desviación estándar de 49ms. La tasa de éxito es del 95.5%.

Otro tema tratado es el desarrollo de clasificadores de modos localizados en el borde. Concretamente se construyó un clasificador de ELMs individuales. Aunque no se han establecido las bases físicas del clasificador, el método parece funcionar para una base de datos de 300 ELMs. Además, la base de datos de prueba se ha clasificado con una elevada tasa de acierto con una muy baja dimensionalidad del vector de características. Aunque los resultados son prometedores, se debe notar que la base de datos no es genérica y una conclusión definitiva debe realizarse con una base de datos mayor.

La clasificación no supervisada también se ha aplicado en este trabajo. El agrupamiento puede utilizarse de forma muy eficiente para encontrar formas de onda similares a una señal de referencia o entrada. Un enfoque de agrupamiento aplicado a las bases de datos del TJ-II muestra que, típicamente, la mayoría de las formas de onda son agrupadas en un gran grupo. Las restantes señales son asignadas en grupos menores.

Los diferentes métodos implementados que ejecutan el agrupamiento muestran resultados similares, agrupando las mismas señales en los mismos grupos, independientemente de las características utilizadas. Los grupos con un número pequeño de señales pueden integrarse un único grupo misceláneo que podría permitir la búsqueda por parte de un científico de aquellos comportamientos del plasma fuera de lo normal de una manera mucho más directa y reducida, que hacerlo en la base de datos completa.

La búsqueda y recuperación de información automatizada es una pieza fundamental para el análisis rápido de comportamientos de plasma similares. La Tesis considera algunas estrategias para tratar este asunto. Una aproximación muy utilizada es el reconocimiento estructural de patrones. Este método reduce la búsqueda de patrones similares a una búsqueda de cadenas de texto similares, lo cual permite por una parte realizar recuperación de información de patrones dentro de una forma de onda completa, y por otro lado, es posible utilizar toda la potencia de los lenguajes de consulta de las bases de datos relacionales.

Un tema no menor considerado en esta Tesis, es la disminución de ruido de imágenes del diagnóstico Thomson Scattering. La presencia de ruido de luz parásita en el diagnóstico podría generar la obtención de perfiles de temperatura y densidad electrónica no fiables o corruptos. Se han presentados dos aproximaciones para tratar este asunto. Ambos métodos están basados en los que se conoce como la teoría de segmentación de imágenes. Su uso es fundamental, dado que el ruido a pesar de estar localizado no es completamente regular. La aplicación de los dos métodos muestra que es posible eliminar en gran parte la distorsión generada por el ruido de luz parásita en el diagnóstico.

4.2 Trabajos Futuros

Todos los problemas presentados en este trabajo pueden ser extendidos en muchas dimensiones. La aplicación de los enfoques implementados en otros dispositivos de fusión es uno de los primeros trabajos futuros disponibles. Sin embargo, muchos otros problemas podrían tener un interés desde el punto de vista de la investigación. Tres asuntos importantes podrían ser considerados. En primer lugar sería interesante aplicar los métodos de búsqueda de patrones en las bases de datos del diagnóstico Thomson Scattering. La idea clave sería traducir los píxeles a primitivas que codifiquen la imagen

como una cadena de texto. El procedimiento de búsqueda sería muy similar entonces a la forma de la aproximación implementada para las señales temporales.

Un segundo trabajo futuro que tiene gran interés es la construcción de sistemas de clasificación multi-capa para la determinación de algún evento físico del plasma. Una posibilidad interesante es la predicción de la aparición de disrupciones en una descarga. Hasta ahora existen trabajos tendientes al desarrollo de tales sistemas predictores (Cannas et al. 2003, 2006, Murari et al. 2008, 2009, Rattá 2012, Ruiz et al. 2010), pero aún con los avances actuales no esta claro como construir un sistema multi-capa específico para este problema. Además las predicciones obtenidas hasta ahora de las disrupciones no consideran el momento en que se estima que el evento ocurrirá, algo que un sistema multi-capa específicamente entrenado para ello podría estimarlo. Junto a esta línea existe otro problema que ocurre antes, y es la selección automática de características (Weston et al. 2001) para un sistema multi-capa, asunto que también se plantea como posible trabajo a futuro a abordar.

Un tercer problema interesante es la construcción de sistemas de aprendizaje automático a partir de bases de datos desbalanceadas. Esta situación ocurre cuando se intenta por ejemplo crear un clasificador, pero no se cuenta con la suficiente cantidad de datos para obtener altas tasas de acierto. Existen técnicas en la literatura que aceleran el proceso de aprendizaje de un sistema de clasificador (Chi et al. 2008, Forman & Cohen 2004), y por tanto sería interesante aplicarlo a algún problema relacionado con las bases de datos de fusión nuclear. Nótese que tener bases de datos desbalanceadas o con un número reducido de datos, puede ocurrir cuando no se tiene un historial de datos, cuando se instala un nuevo diagnóstico, o cuando se ha modificado algún elemento clave del dispositivo de fusión.

III

Published Articles

Article 1

Application and validation of image algorithms on TJ-II TS diagnostic

1.1 Bibliographic Description

Title

Application and validation of image processing algorithms to reduce the stray light on the TJ-II Thomson Scattering diagnostic.

Citation

G. Farias, S. Dormido-Canto, J. Vega, I. Pastor, M. Santos (2013) Application and Validation of Image Processing Algorithms to Reduce the Stray Light on the TJ-II Thomson Scattering Diagnostic, *Fusion Science and Technology*, ISSN 1536-1055, Volume 63, Number 1, Pages 20-25.

Abstract

Stray light is the main source of noise on the Thomson scattering diagnostic images of the TJ-II stellarator. The diagnostic provides temperature and density profiles of the plasma. A charge-coupled-device camera acquires images that are disturbed by noise, which, in some cases, can produce unreliable profiles. In this paper we describe three

different approaches to reduce or mitigate the stray light on these images: exhaustive detection, extraction of regions with connected components, and extraction of regions with the approach of region growing. The performance of the two most interesting techniques is evaluated by a validation process. This process quantifies the noise eliminated by each method.

References

G. Farias et al. (2005); L. Makili et al.(2010); C. Alejaldre et al.(1999); R. P. Breault (1995); H. Rowley et al. (1995); V. Ferrari et al.(2008); C. Gu et al.(2009); B. Fulkerson et al. (2009).

Impact Factor

Fusion Science and Technology has an impact factor of 1.12 according to Thomson Reuters Journal Citation Reports (2011).

APPLICATION AND VALIDATION OF IMAGE PROCESSING ALGORITHMS TO REDUCE THE STRAY LIGHT ON THE TJ-II THOMSON SCATTERING DIAGNOSTIC

GONZALO FARIAS,[a]* SEBASTIÁN DORMIDO-CANTO,[b] JESÚS VEGA,[c] IGNACIO PASTOR,[c] and MATILDE SANTOS[d]

[a]*Pontificia Universidad Católica de Valparaíso, Av. Brasil 2147, Valparaíso, Chile*
[b]*Universidad Nacional de Educación a Distancia, Madrid, Spain*
[c]*Asociación EURATOM/CIEMAT para Fusión, Madrid, Spain*
[d]*Universidad Complutense de Madrid, Madrid, Spain*

Received April 27, 2012
Accepted for Publication August 1, 2012

Stray light is the main source of noise on the Thomson scattering diagnostic images of the TJ-II stellarator. The diagnostic provides temperature and density profiles of the plasma. A charge-coupled-device camera acquires images that are disturbed by noise, which, in some cases, can produce unreliable profiles. In this paper we describe three different approaches to reduce or mitigate the stray light on these images: exhaustive detection, extraction of regions with connected components, and extraction of regions with the approach of region growing. The performance of the two most interesting techniques is evaluated by a validation process. This process quantifies the noise eliminated by each method.

KEYWORDS: *stray light, image processing, Thomson scattering diagnostic*

I. INTRODUCTION

An automatic image classification system has been in operation for years in the TJ-II Thomson scattering (TS) diagnostic.[1,2] It recognizes five different types of images: charge-coupled-device (CCD) camera background (BKG), measurement of stray light without plasma or in a collapsed discharge (STR), images during the electron cyclotron resonance heating (ECRH) phase, images during the neutral beam injection (NBI) phase, and images after reaching the cutoff density during ECRH heating (COFF).

Since the first implementation of the classifier,[1,2] a relevant improvement has been accomplished in the diagnostic: A new notch filter is in operation, having a larger stray-light rejection at the ruby wavelength than the previous filter. However, the unfiltered stray light acts as a disturbance (noise) in the images, and total elimination is required.

*E-mail: gonzalo.farias@ucv.cl

The noise on images can be reduced by applying many classical and advanced techniques such as low-pass filters or wavelets. However, in some cases the presence of noise is not global but located only in particular regions of the image. In these situations the application of global filters over the entire image is not a suitable option since the noise and the information are equally reduced.

In this paper, three different approaches are presented to reduce or mitigate the stray light (noise) on the images taking into account the particularities (especially the regular location of the anomalies) of the noise: exhaustive detection, region extraction, and region growing. Exhaustive detection removes the noise in specific zones of the image, and region extraction and region growing detect and then eliminate regions of the image with the same features as stray light.

The goal of this work is to eliminate the regions (noise) due to stray light without eliminating significant information. The TS diagnostic of the TJ-II stellarator is briefly described in Sec. II. The problem formulation and

the approaches implemented to remove stray light in an image are analyzed with a simple example in Sec. III. Results and details about a specific implementation of the image processing methods for noise reduction in the TJ-II TS diagnostic are given in Sec. III as well. A validation mechanism has also been implemented to evaluate the performance of each algorithm and is described in Sec. IV. Finally, Sec. V summarizes the main conclusions and possible future works where the research can be addressed to improve the results.

II. TJ-II TS DIAGNOSTIC

The TJ-II is a medium-size stellarator (heliac type)[3] located at CIEMAT (Spain). The TS in plasmas consists of the reemission of incident radiation (from very powerful lasers) by free electrons. Electron velocity distribution generates a spectral broadening of the scattered light (by the Doppler effect) related to the electronic temperature. The total number of scattered photons is proportional to the electronic density.

Every laser shot produces a bidimensional image from which it is possible to obtain radial profiles of temperature and density. Only a restricted number of pattern images appear in the TJ-II. They represent different physical situations related to either the plasma heating or the system calibration. Depending on the pattern obtained, the data are processed in different ways. Figure 1 shows the five types of images found in the TJ-II TS diagnostic.

In the TS diagnostic the main source of noise is so-called stray light. Controlling stray light has always been important in optical design.[4] Caused by phenomena such as Fresnel reflection from lens surfaces, air bubbles in glass, dust, diffraction at aperture edges, and numerous other effects, its presence frequently degrades both image contrast and measurement accuracy. In particular, the CCD camera in the TS diagnostic acquires images corrupted with stray light, which, in some cases, can produce unreliable profiles of temperature and density. This is light from the ruby laser that reaches the spectrometer, which is impossible to separate from the real scattered signal. According to experts, the stray light (noise) always appears on the left side of TS images. So far, different hardware techniques have been tried to remove or decrease the stray-light contribution but with only partial success. Note the location of the stray light on the TS images in Fig. 1.

III. IMAGE PROCESSING ALGORITHMS

In this work, three different approaches to reduce stray light on TS images are described: (a) exhaustive detection, (b) connected regions, and (c) region growing.

To explain these approaches, let us consider a toy example (Fig. 2). The image is 10×12 pixels. The image can be represented by a matrix, where the term $I[r, c]$ denotes the gray level of the pixel located at row r, column c of the image. A matrix has $\boldsymbol{m}$ rows and $\boldsymbol{n}$ columns.

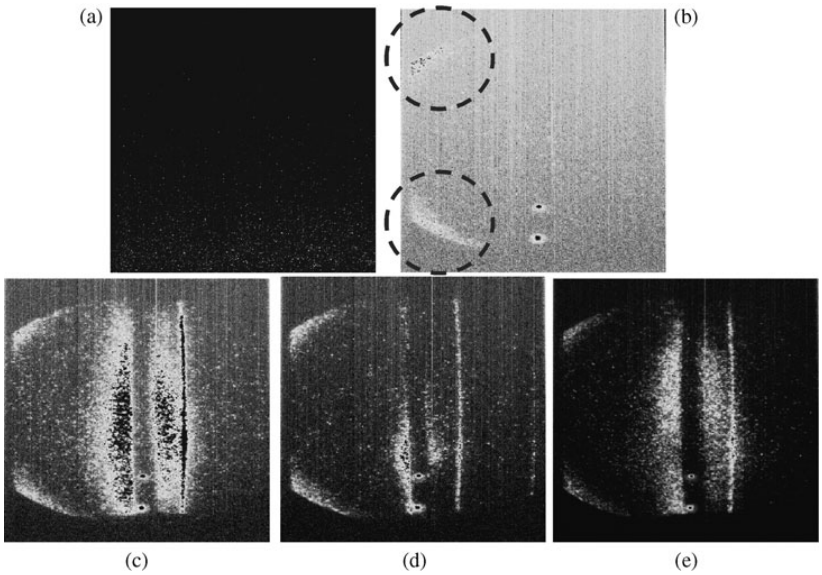


Fig. 1. Classes of images acquired by the TJ-II TS: (a) BKG, (b) STR, (c) NBI, (d) COFF, and (e) ECRH. The circles in (b) show the noise of stray light. Note that noise appears on all classes except BKG.

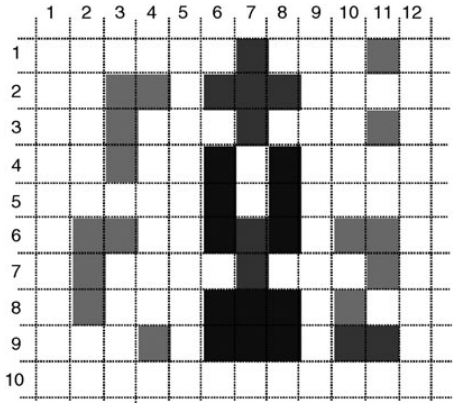


Fig. 2. Original image for toy example.

Thus, $I[1,1]$ refers to the gray level of the upper-leftmost pixel and $I[\boldsymbol{m},\boldsymbol{n}]$ refers to the gray level of the lower-rightmost pixel. The goal in this work is to remove or reduce as much as possible the noise present on the image. In the case of Fig. 2, assume that noise is located on the following sets of pixels: [(2,3), (2,4), (3,3), (4,3)] and [(6,2), (6,3), (7,2), (8,2)]. Assume also that the rest of the pixels are considered as background [e.g., pixels (1,1), (2,5), etc.] or as information [e.g., pixels (2,7), (1,11), etc.]. Note that normally the background color is represented by white color (i.e., gray level equal to zero).

III.A. Exhaustive Noise Detection

This method is based on the sliding window scanning throughout the image.[5,6] The technique requires describing the noise to be removed by using a template, which corresponds to a predefined set of pixels. In the toy example, the template should be represented by the following pattern of pixel locations: $[(r,c), (r+1,c), (r+2,c), (r,c+1)]$. Once the template is defined, it is moved throughout the image to find matches. Obviously, all coincidences should be replaced by background color. It is easy to see some problems that occur with this technique: What happens if the noise borders are not regular? What happens if there are connected pixels that lie outside the template? What happens when the pixels within the template have different intensities or gray levels? What determines whether pixels within the template are noise or not? The technique of extraction of regions with connected components (ERCC) tries to answer all these questions.

III.B. Extraction of Regions with Connected Components

The ERCC method is based on segmentation theory.[7,8] Segmentation refers to the process of partitioning a digital image into multiple segments (sets of pixels). More precisely, image segmentation is the process of assigning a label to every pixel in an image so pixels with the same label share certain visual characteristics. In general, a segmentation of an image is a partition into connected subimages (regions) $R_1, R_2, \ldots, R_n$ such that all regions are disjointed, and the union of all of them makes up the image. Each subimage satisfies a predicate such as all pixels in any subimage R_i must not differ by more than Δx gray levels, all pixels in any subimage R_i must be joined by a connectivity factor, etc.

The procedure proposed for noise removal by using connected components is shown in Fig. 3. Each step of the process is explained below.

A binary image can be obtained from a grayscale or color image by thresholding, which selects a subset of the image pixels as foreground pixels and leaves the rest as background. In this work, the threshold is obtained by calculating the mean of the pixel values on the right side of the image because the noise is not present there. The pixels of a binary image are 0s and 1s; the 1s stand for foreground pixels and the 0s stand for background pixels. After that, segmentation of the image can be obtained by using a predicate of connectivity. A connected-components labeling of a binary image B is a labeled image LB in which the value of each pixel is the label of its connected component. The connected-components labeling is associated to a

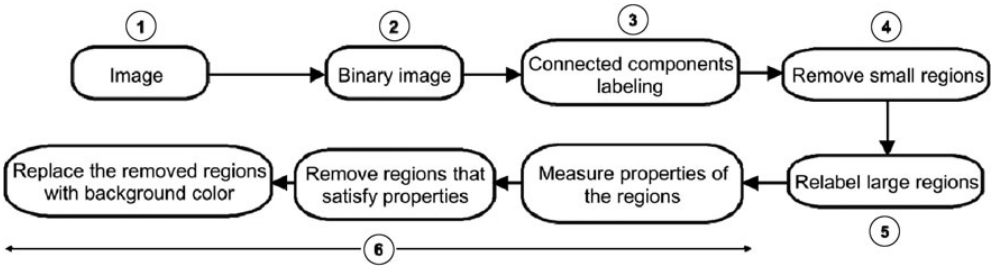


Fig. 3. Flowchart for ERCC.

connectivity factor between neighbors. The two most common areas of neighbors are the four-neighbors and the eight-neighbors of a pixel. The four-neighborhood $N_4(r,c)$ of pixel (r,c) includes pixels $(r-1,c)$, $(r+1,c)$, $(r,c-1)$, and $(r,c+1)$, which are often referred to as its north, south, west, and east neighbors, respectively. The eight-neighborhood $N_8(r,c)$ of pixel (r,c) includes each pixel of the four-neighborhood plus the diagonal neighbor pixels $(r-1,c-1)$, $(r-1,c+1)$, $(r+1,c-1)$, and $(r+1,c+1)$.

The next step is to remove from *LB* all regions that have fewer than *P* pixels (small regions), and then the remaining regions are relabeled. After that, for each labeled region in the label matrix (*LB*), it is possible to calculate certain properties of the regions such as the area, centroid, circularity, elongation, etc. The regions that meet the desired properties are replaced by background color, and small regions that were deleted are restored. In our example the parameters, properties, and conditions were the following: N_8, $P = 2$, and elimination of regions whose centroid is on the left half-plane.

Figure 4 shows the result for each stage of the procedure described with an image of type NBI. The parameters, properties, and conditions in the procedure applied to the TS diagnostic were the following: N_8, $P = 20$, and elimination of regions for which the *c* coordinate of the centroid is <100.

III.C. Extraction of Regions with Region Growing

As can be observed in Fig. 4, there are very good results on the noise reduction in the TS diagnostic with connected components, but the predicate of connection for a pixel is sometimes too strong. For instance, pixels quite near, but not connected, to the region are not considered as noise in this approach. This is where region growing comes in handy. Region growing allows addition of pixels to a region by using a custom predicate. Thus, a pixel could be considered part of a region although the location of the pixel is near but not connected to the region. Besides, pixels "connected" to a region can be considered as a different region if the intensity of the pixel is quite different to the mean intensity of the region. Region growing requires an adequate selection process of seeds. In this work, the seeds are selected from an initial set of candidate pixels. Candidate pixels must hold higher values than a threshold (the same threshold used in the binarization process). Since the position of noise is commonly regular, the seeds normally match inside the noise. The procedure proposed for noise removal by using region growing is shown in Fig. 5.

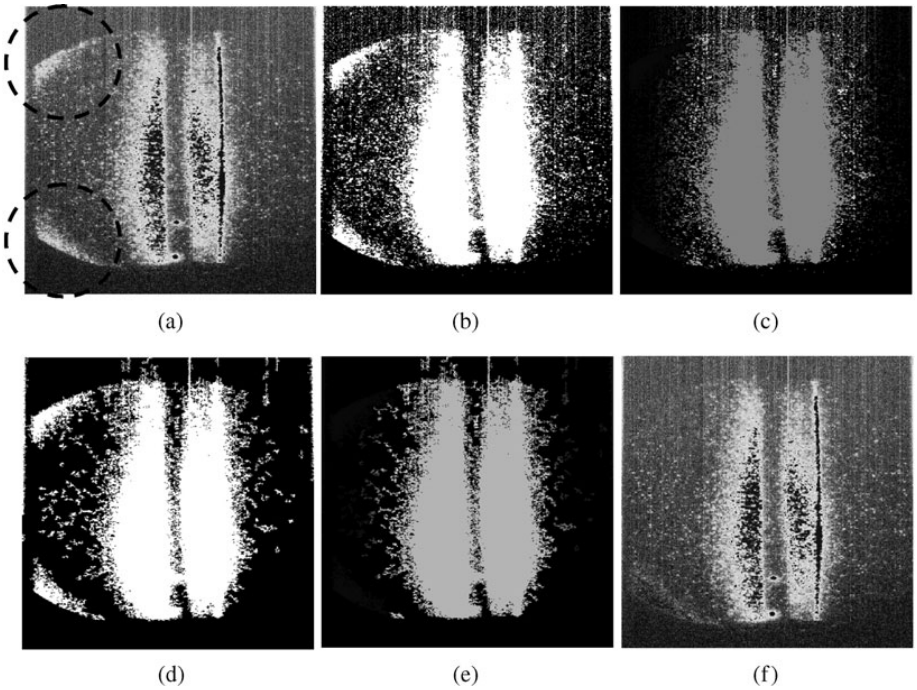


Fig. 4. Images showing the steps of the ERCC algorithm: (a) original image, (b) binarized image, (c) labeled image, (d) image with regions greater than *P* pixels, (e) relabeled image, and (f) image with noise removed.

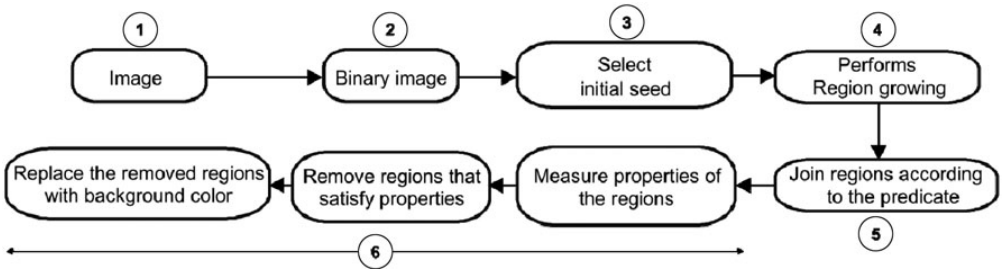


Fig. 5. Flowchart for region extraction with the region growing approach.

IV. VALIDATING THE APPROACHES

To validate both algorithms, a denoised function has been defined, which is equal to 1 if the stray light has been removed successfully from the image; otherwise, the function value is 0. The key idea behind the denoised function is based on the premise that the left and right extremes of any image after applying a denoised method should be similar. Thus, the mean of the pixel values of each side of an image is computed as Fig. 6 shows, and the means are denoted as μ_{left} and μ_{right}, respectively.

Considering $\delta\mu_f = \mu_{left} - \mu_{right}$ as the difference between the mean pixel values of the left side and the mean pixel values of the right side of an image f, the denoised $D(f)$ function can be defined as

$$D(f) = \begin{cases} 1\ , & \delta\mu_{\mathrm{BKG}} - 3\sigma\mu_{\mathrm{BKG}} < \delta\mu_f < \delta\mu_{\mathrm{BKG}} \\ & \quad + 3\sigma\mu_{\mathrm{BKG}} \\ 0\ , & \text{other case}\ , \end{cases}$$

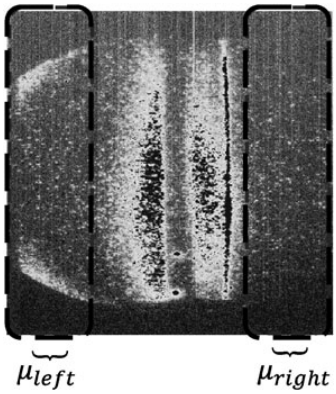


Fig. 6. Image areas considered to compute the mean pixel values on the left and right extremes of any image.

TABLE I

Success Rate of the Denoised Function for Both Algorithms

	BKG	STR	NBI	COFF	ECRH
ERCC	98%	96%	91%	97%	94%
Region growing	98%	97%	95%	97%	96%

where $\delta\mu_{\mathrm{BKG}}$ and $\sigma\mu_{\mathrm{BKG}}$ are the mean and the standard deviation, respectively, of $\delta\mu_f$ for all BKG images. Thus, the $D(f)$ function equals 1 when the stray light is reduced; otherwise, it equals 0.

The validation process has been tested for 1146 TS images. The $D(f)$ function is computed for all images per each TS class. Table I shows the percentage of denoised images for each algorithm.

Note that the results for region growing are slightly better than for the ERCC algorithm. In the case of NBI images, the difference is mainly because ERCC is not able to reduce the stray light when the noise is "connected" to the central part (the significant information).

V. CONCLUSIONS

Segmentation-based methods have proven to be useful for removing the stray light in the TJ-II TS diagnostic without eliminating significant information. ERCC shows a great performance even though the predicate, the connectivity, which defines when a pixel belongs to a region, is very simple. Region growing improves the region extraction approach of ERCC, involving a more complex predicate that allows distinguishing in a better way when a pixel belongs to a region. The performance of both algorithms has been tested by a validation method that allows us to quantify the removed information. Further

work can focus on using the validation mechanism to automatically adjust the algorithm's parameters to eliminate stray light.

ACKNOWLEDGMENTS

This work was partially funded by the Spanish Ministry of Science and Innovation under project ENE2008-02894/FTN.

REFERENCES

1. G. FARIAS et al., "Image Classifier for the TJ-II Thomson Scattering Diagnostic: Evaluation with a Feed Forward Neural Network," *Lect. Notes Comput. Sci*., **3562**, *2*, 604 (2005).

2. L. MAKILI et al., "Upgrade of the Automatic Analysis System in the TJ-II Thomson Scattering Diagnostic: New Image Recognition Classifier and Fault Condition Detection," *Fusion Eng. Des*., **85**, 415 (2010).

3. C. ALEJALDRE et al., *Plasma Phys. Control. Fusion*, **41**, *1*, A539 (1999).

4. R. P. BREAULT, "Control of Stray Light," *Handbook of Optics*, Vol. 1, McGraw-Hill, Inc., New York (1995).

5. H. ROWLEY, S. BALUJA, and T. KANADE, "Human Face Detection in Visual Scenes," *Proc. Advances in Neural Information Processing Systems*, Denver, Colorado, November 27–30, 1995, p. 875.

6. V. FERRARI et al., "Groups of Adjacent Contour Segments for Object Detection," *IEEE Trans. Pattern Anal. Mach. Intell*., **30**, 36 (2008).

7. C. GU et al., "Recognition Using Regions," *Proc. Computer Society Conf. Computer Vision and Pattern Recognition (CVPR)*, Miami, Florida, June 20–25, 2009, p. 1030, IEEE (2009).

8. B. FULKERSON, A. VEDALDI, and S. SOATTO, "Class Segmentation and Object Localization with Superpixel Neighborhoods," *Proc. 12th Int. Conf. Computer Vision (ICCV)*, Kyoto, Japan, September 27–October 4, 2009, p. 670, IEEE (2009).

Article 2

Automatic determination of L/H transition times in DIII-D

2.1 Bibliographic Description

Title

Automatic determination of L/H transition times in DIII-D through a collaborative distributed environment.

Citation

G. Farias, J. Vega, S. González, A. Pereira, X. Lee, D. Schissel, P. Gohil (2012) Automatic determination of L/H transition times in DIII-D through a collaborative distributed environment, *Fusion Engineering and Design*, ISSN 0920-3796, Volume 87, Issue 12, Pages 2081-2083.

Abstract

An automatic predictor of L/H transition times has been implemented for the DIII-D tokamak. The system predicts the transition combining two techniques: A morphological pattern recognition algorithm, which estimates the transition based on the waveform of a Dα emission signal, and a support vector machines multi-layer model, which predicts the L/H transition using a non-parametric model. The predictor is employed within a collaborative distributed computing environment. The system is trained remotely in the

Ciemat computer cluster and operated on the DIII-D site.

References

V. Vapnik (2000); R.O. Duda, P.E. Hart, D.G. Stork(2001); S. González et al. (2010); J. Vega, A. Murari, G. Vagliasindi, G.A. Ratta, JET-EFDA Contributors (2009).

Impact Factor

Fusion Engineering and Design has an impact factor of 1.49 according to Thomson Reuters Journal Citation Reports (2011).

G Model
FUSION-6203; No. of Pages 3
ARTICLE IN PRESS

Fusion Engineering and Design xxx (2012) xxx–xxx

Contents lists available at SciVerse ScienceDirect

Fusion Engineering and Design

journal homepage: www.elsevier.com/locate/fusengdes

Automatic determination of L/H transition times in DIII-D through a collaborative distributed environment

G. Farias [a,*], J. Vega [a], S. González [a], A. Pereira [a], X. Lee [b], D. Schissel [b], P. Gohil [b]

[a] Asociación EURATOM/CIEMAT para fusión, Avd. Complutense 22, 28040 Madrid, Spain
[b] General Atomics, P.O. Box 85608, San Diego, CA 92186-5608, USA

ARTICLE INFO

Article history:
Available online xxx

Keywords:
L/H transition time
Conformal predictor
Distributed environment

ABSTRACT

An automatic predictor of L/H transition times has been implemented for the DIII-D tokamak. The system predicts the transition combining two techniques: A morphological pattern recognition algorithm, which estimates the transition based on the waveform of a Dα emission signal, and a support vector machines multi-layer model, which predicts the L/H transition using a non-parametric model. The predictor is employed within a collaborative distributed computing environment. The system is trained remotely in the Ciemat computer cluster and operated on the DIII-D site.

1. Introduction

Machine learning methods have been developed to automatically determine the time instants of L/H transitions in the DIII-D tokamak. A training dataset is used to generate a non-parametric model to distinguish between the L and H confinement modes at any time of a discharge. The only requirement to create the model is to assume that all samples are independent and they are identically distributed according to a fixed but unknown distribution. The model also provides the uncertainty (error bar) in the prediction of the transition time. To this end, conformal predictors are used. Conformal predictors qualify their predictions with a couple of values, confidence and credibility, that provide information about how accurate and reliable the predictions are. The system implementation is carried out in two steps within a distributed computing environment. Firstly, a support vector machine (SVM) [1] multi-layer model is created by using a training dataset. The SVM model uses a combination of several features (signals) to determine the L/H transitions. The selection of the dataset is accomplished in an automatic way by means of a morphological pattern recognition (MPR) technique in the Dα emission signal. Only the discharges that show a clear pattern of an L/H transition constitute the training dataset. Secondly, the SVM multi-layer model and the MPR algorithm are combined to predict separately the L/H transition time of new discharges. The first step, which is called training and validation mode, requires high performance computing due to the possibility of using a large number of training discharges with hundreds or thousands of samples per discharge. This step is executed remotely in a CIEMAT computer cluster. The second step, called operation mode, is executed on the DIII-D site. It is important to note that this step can be executed in an unattended manner after each discharge and therefore can be added to the existing automatic between pulse data processing. The training and validation process, i.e. first step, can be also executed at any time to update the model or to test new models. Results of this two-step process will be presented along with initial results obtained during between pulse data analysis at DIII-D.

The article is organized as follows: Section 2 describes briefly the two techniques used to predict L/H transitions. Section 3 shows the process to build the MPR+SVM predictor. Finally Section 4 presents the main results and conclusion of this work.

2. Approaches to predict L/H transition times

2.1. Morphological pattern recognition algorithm

A L/H transition can be located by a fast drop of the Dα emission between the start of the NBI heating system and the first type I ELM of the pulse. Top plot of Fig. 1 shows a Dα emission signal for a particular discharge. Note that the L/H transition (fast drop) occurs about 600 ms, the duration of the discharge is approximately 3000 ms, and the first type I ELM is located around 700 ms.

The MPR algorithm looks for the fast drop by using only structural information of the waveform of the Dα signal. The process is divided in 5 steps. The first three steps are mainly devoted to reducing the searching interval. This interval is located detecting the beginning of the power injection and the identification of the ELMs region (gray zone on the top plot of Fig. 1). The last two steps try to identify the time when the L/H transition (fast drop) occurs. Wavelet [2] and SVM regression techniques are used in the

* Corresponding author. Tel.: +34 913987147.
E-mail address: gfarias@bec.uned.es (G. Farias).

0920-3796/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.fusengdes.2012.02.038

Please cite this article in press as: G. Farias, et al., Automatic determination of L/H transition times in DIII-D through a collaborative distributed environment, Fusion Eng. Des. (2012), doi:10.1016/j.fusengdes.2012.02.038

G Model
FUSION-6203; No. of Pages 3
ARTICLE IN PRESS

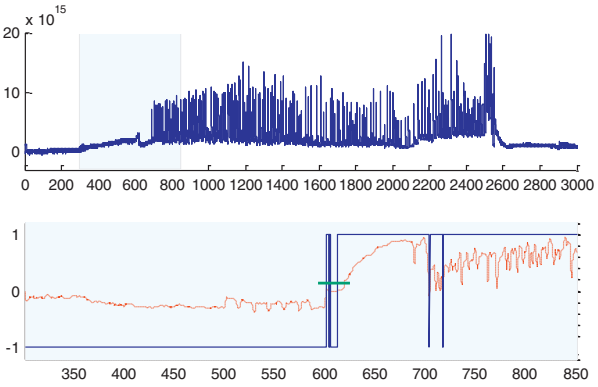


Fig. 1. Upper plot shows an interval of a Dα emission signal where the L/H transition takes place at 600 ms approx. Lower plot depicts the distance to the SVM separating hyperplane and the classification of a SVM model for the interval.

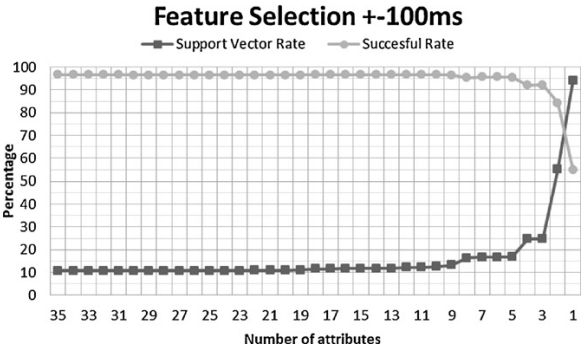


Fig. 2. Feature selection process for ±100 interval. Note that when the number of features is lower than 5, the successful rate decreases enormously.

whole process mainly to reduce the dimensionality of the signal, and to detect ELMs respectively. Apart of the prediction, the MPR also classifies the type of transitions as *sharp* and *no sharp* in order to distinguish the type of drop from a morphological point of view. Sharp predictions are normally much more accurate than no sharp ones.

2.2. Predicting L/H transitions with the SVM model

A SVM bi-classifier has been used to predict the L/H transitions. The basic idea is to train a SVM model using suitable signals per each sample around the previously known L/H transition times. The two classes will be then −1 and +1 corresponding to the L and H modes respectively. Note that the time information should not be used in order to get a generalized predictor because; using only signal's amplitude will produce a SVM model useful to detect the L/H transitions at any time of the discharge.

The bottom plot of Fig. 1 shows the prediction of the L/H transition using a SVM classifier for the highlighted interval time (300–850 ms) in the upper plot. The dotted and solid lines represent the distance to the SVM separating *hyperplane* and the predicted class respectively. The horizontal line at 600 ms corresponds to the error bar given by a conformal predictor.

Observing the predicted class of the SVM classifier, the L/H transition seems to be near to 600 ms. Since sometimes the classification could be disturbed by noise, such as is shown near 600 ms or 700 ms, a criterion is needed in order to distinguish these disturbances from the actual transition. In our case, a L/H transition is predicted by the classifier if there is a L interval followed by a separation interval (where the transition is normally switching between L and H modes) and then by a H interval. The L and H intervals should have at least 20 ms of longitude while the separation between them should be no longer than 200 ms. Under this situation, the transition is predicted at half of the separation interval, otherwise there is no prediction.

Conformal predictors are used to qualify the SVM predictions with confidence and credibility, which provide information about how accurate and reliable the predictions are. The extension of the error bar, depicted in the lower plot of Fig. 1, is computed considering 30% of the credibility around the L/H transition predicted by the SVM model.

3. Building the L/H predictor

The SVM multi-layer model is trained by using the output of two SVM models. The two SVM models are, in turn, trained with samples of ±50 ms and ±100 ms around the previously known L/H transitions. The process to train a SVM model consists of two steps: *Feature Selection* and *Automatic Determination of Training Set*.

The aim of these two steps is to define a set of training signals and discharges in order to get a generalized predictor with a high success rate.

After the training process, the L/H transition time was predicted by using a combination of the MPR algorithm and the SVM models.

3.1. Feature selection process

A large set of 35 candidate features were defined based on the previous experience in JET [3,4]. The set of candidates is composed of Dα signals, density and power signals, among others. The use of such a set of features implies a large amount of data requiring high performance computing. So instead of using all features of the original set, it was decided to perform a selection process in order to get a suitable subset of signals.

The feature selection process evaluates a subset of signals as a group for suitability. The evaluation of a subset is given by creating a linear SVM model and measuring its success rate (this kind of evaluation is normally known as *wrapper* subset selection). Observing the weight of each feature in the linear model, it is reasonable to discard the signal with the lowest weight.

The selection process starts with the set of 35 candidate features, discarding a signal at each evaluation, and finishes when there is only 1 feature. The final subset can be chosen considering the success rate, the number of features, and the percentage of the samples considered as support vectors (known as *support vector rate*).

The described selection process was carried out on the two intervals of interest: ±50 ms and ±100 ms. Fig. 2 depicts the result of the process for the ±100 interval.

Two subsets with 6 and 5 features for the intervals ±50 and ±100 respectively were selected. Table 1 shows the *pointnames* and description of each feature selected.

Table 1
Pointnames of the selected signals for ±50 ms and ±100 ms.

±50 ms	±100 ms
fs04da (*Dα signal*)	fs04da (*Dα signal*)
density (*density*)	density (*density*)
echpwrc (*ECH power*)	poh (*ohmic power*)
ip (*plasma current*)	prad_tot (*radiated power*)
prad_tot (*radiated power*)	totalpower
totalpower	

Please cite this article in press as: G. Farias, et al., Automatic determination of L/H transition times in DIII-D through a collaborative distributed environment, Fusion Eng. Des. (2012), doi:10.1016/j.fusengdes.2012.02.038

G Model
FUSION-6203; No. of Pages 3
ARTICLE IN PRESS
G. Farias et al. / Fusion Engineering and Design xxx (2012) xxx–xxx

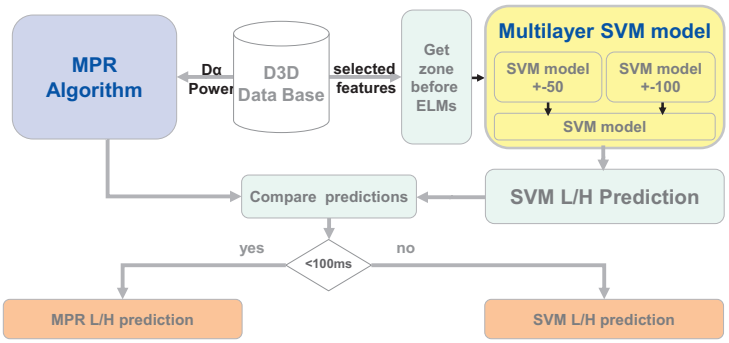


Fig. 3. Combination of MPR algorithm and SVM models to predict L/H transition times.

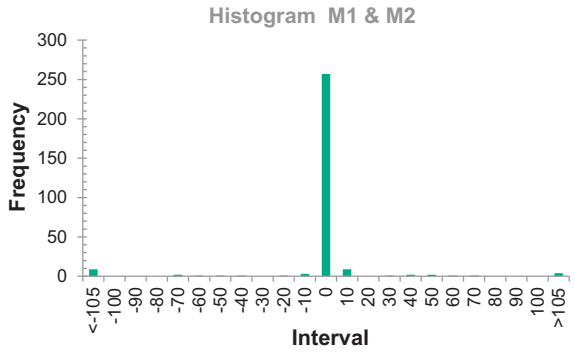


Fig. 4. Histogram of the prediction error of the MPR + SVM system.

3.2. Automatic determination of the training set

After the feature selection process, an automatic process was performed to select a set of shots or discharges in order to train a SVM model. The process is divided in three stages.

First the morphologic algorithm is executed to predict L/H transitions in 354 shots. The algorithm predicted 291 transitions as sharp transitions. A simple outlier removal process is then used reducing the number of shots to 279.

Secondly, the shots with sharp transitions were used to train a linear SVM model with the subset of signals selected for the ±100 interval. The model is used to predict L/H transitions for the same training shots. Those shots which have not been predicted with L/H transition by the SVM model are discarded, and a new linear SVM model is trained using the rest of shots. When no shot is discarded, the same procedure is executed but now using the subset of signals selected for the ±50 interval.

Finally, a linear SVM multi-layer model is built by using the output (i.e. the distance to the hyperplane) of the ±50 and ±100 SVM models previously obtained. The prediction of the SVM multi-layer uses the criteria described in Section 2.2.

3.3. Predicting L/H transition times

In order to predict the L/H transition the predictions of the MPR algorithm and the multi-layer SVM model were combined.

On one hand, the morphologic algorithm takes into account a Dα emission and power signals to predict the transition. On the other hand, the multilayer SVM model uses the selected features of Table 1 to perform the prediction. Note that the SVM models put the focus of the searching only on the zone before ELMs start.

If the difference of both predictions is less than 100 ms then the final prediction is given by MPR, otherwise it is given by SVM. This criterion was obtained after several experiments. Fig. 3 shows the scheme used to predict L/H transitions in this work.

4. Main results and conclusions

An automatic predictor of L/H transition times has been implemented to be used in DIII-D. The system is trained in Ciemat and the operation is carried out in DIII-D site.

The predictor has been tested with the initial 354 discharges. The combined predictor has an average of error prediction of 6 ms and a standard deviation of 49 ms. The successful rate is 95.6%. Fig. 4 shows the histogram with the frequency of the prediction error.

Acknowledgments

This work was supported in part by The Spanish Ministry of Science and Innovation under the Project No. ENE2008-02894/FTN, and in part by The US Department of Energy under DE-FC02-04ER54698.

References

[1] V. Vapnik, The Nature of Statistical Learning Theory, 2nd ed., Springer, 2000.
[2] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, 2nd ed., John Wiley & Sons, Inc., 2001.
[3] S. González, J. Vega, A. Murari, A. Pereira, J.M. Ramírez, S. Dormido-Canto, et al., Review of Scientific Instruments 81 (2010) 10E123 (3 p).
[4] J. Vega, A. Murari, G. Vagliasindi, G.A. Ratta, JET-EFDA Contributors, Nuclear Fusion 49 (2009) 085023 (11 p).

Please cite this article in press as: G. Farias, et al., Automatic determination of L/H transition times in DIII-D through a collaborative distributed environment, Fusion Eng. Des. (2012), doi:10.1016/j.fusengdes.2012.02.038

Article 3

Image processing methods for noise reduction on TJ-II TS diagnostic

3.1 Bibliographic Description

Title

Image processing methods for noise reduction in the TJ-II Thomson Scattering diagnostic.

Citation

S. Dormido-Canto, G. Farias, J. Vega, I. Pastor (2012) Image processing methods for noise reduction in the TJ-II Thomson Scattering diagnostic, *Fusion Engineering and Design*, ISSN 0920-3796, Volume 87, Issue 12, Pages 21702173.

Abstract

The Thomsom Scattering diagnostic of the TJ-II stellarator provides temperature and density profiles. The CCD camera acquires images corrupted with noise that, in some cases, can produce unreliable profiles. The main source of noise is the so-called stray-light. In this paper we describe an approach that allows mitigation of the effects that stray-light has on the images: extraction regions with connected components. In addi-

tion, the robustness and effectiveness of the noise reduction technique is validated in two ways: (1) supervised classification and (2) comparison of electron temperature profiles.

References

H. Rowley, S. Baluja, T. Kanade(1995); V. Ferrari et al. (2008); C. Gu et al. (2009); B. Fulkerson, A. Vedaldi, S. Soatto (2009); R.P. Breault (1995); The Mathworks (2011); L. Makili et al. (2010).

Impact Factor

Fusion Engineering and Design has an impact factor of 1.49 according to Thomson Reuters Journal Citation Reports (2011).

Fusion Engineering and Design 87 (2012) 2170–2173

Contents lists available at SciVerse ScienceDirect

Fusion Engineering and Design

journal homepage: www.elsevier.com/locate/fusengdes

Image processing methods for noise reduction in the TJ-II Thomson Scattering diagnostic

S. Dormido-Canto[a,*], G. Farias[c], J. Vega[b], I. Pastor[b]

[a] Departamento de Informática y Automática, UNED, Madrid 28040, Spain
[b] Asociación EURATOM/CIEMAT para Fusión, Madrid 28040, Spain
[c] Pontificia Universidad Catolica de Valparaiso, Valparaiso, Chile

ARTICLE INFO

Article history:
Available online 7 May 2012

Keywords:
Image processing
Thomson Scattering
Segmentation

ABSTRACT

The Thomsom Scattering diagnostic of the TJ-II stellarator provides temperature and density profiles. The CCD camera acquires images corrupted with noise that, in some cases, can produce unreliable profiles. The main source of noise is the so-called stray-light. In this paper we describe an approach that allows mitigation of the effects that stray-light has on the images: extraction regions with connected-components. In addition, the robustness and effectiveness of the noise reduction technique is validated in two ways: (1) supervised classification and (2) comparison of electron temperature profiles.

1. Introduction

Over the past couple of decades, most object recognition approaches were based on the sliding window scanning [1,2], localizing the bounding box of an object. While very successful for some applications, such as face and pedestrian detection, there are several well-known issues with the sliding window approach. First, it is mostly appropriate for object classes that are well approximated by a rectangle, but not for object classes with irregular parts. Another problem is that there is no precise pixel-level segmentation of the detected object. Also, sliding window approach is computationally expensive, since a large number of sub-windows at different scales need to be processed for a single image.

In recent years, there has been a lot of interest in using results of a generic image segmentation algorithm to help obtain pixel-precise object segmentation [3,4]. Segmentation is used to subdivide an image into a set of regions. These regions do not have predefined shapes like rectangular patches and their boundaries are irregulars in the image. Therefore the shape and boundary properties of regions can be used for feature extraction. Another advantage of using image regions is scalability and potential savings in computational efficiency. Image regions usually provide a much smaller set of hypothesis to examine compared to the sliding window approach, and at "natural" scales that are obtained through segmentation.

This paper has been structured to mainly focus the attention on the noise reduction in the TJ-II Thomson Scattering (TS) diagnostic. In TS diagnostic the main source of noise is the so-called stray-light. Controlling stray light has always been important in optical design [5]. Caused by phenomena such as Fresnel reflection from lens surfaces, air bubbles in glass, dust, diffraction at aperture edges, and numerous other effects, its presence frequently degrades both image contrast and measurement accuracy. In particular, the CCD camera in the TS diagnostic acquires images (spectra of laser light scattered by plasma) corrupted with stray light that, in some cases, can produce unreliable profiles. One example is the light from the ruby laser which reaches the spectrometer and there is no possibility to distinguish it from the light scattered by the electrons. So far, different hardware techniques have been tried to remove/decrease the stray light contribution but only with partial success. Such as, to place a notch filter in front of the spectrometer or inside the spectrometer or to carry out a correct alignment of the system.

The objective of this work is to eliminate the regions due to stray light without eliminating significant information. The problem formulation and the approach implemented in order to remove undesired regions in an image are analyzed with a simplified example in Section 2. Results and details about a specific implementation of the image processing methods for noise reduction in the TJ-II Thomson Scattering diagnostic are given in Section 3. Finally, Section 4 summarizes the main conclusions.

2. Problem formulation

In our case, we define an approach based on extraction regions with connected-components (ERCC) [6] in order to remove some specifics regions associated to the noise.

To explain it let us consider a simplified example (Fig. 1a). The image has 10×12 pixels. The goal is to remove the noise which at sets of pixels [(2,3), (2,4), (3,3), (4,3)] and [(6,2), (6,3), (7,2), (8,2)]. The background color in the image is white, for example pixel (1,1).

* Corresponding author. Tel.: +34 913987194; fax: +34 913987690.
E-mail address: sebas@dia.uned.es (S. Dormido-Canto).

0920-3796/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.fusengdes.2012.04.004

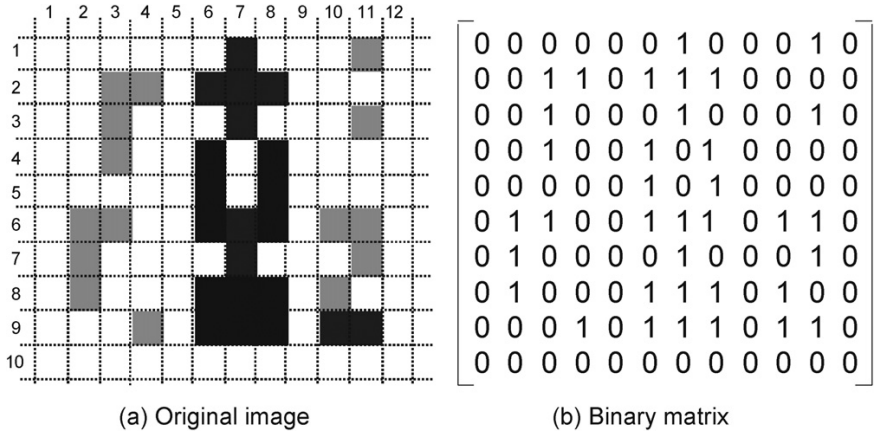


Fig. 1. Simplified example.

The ERCC is based on the segmentation theory in image processing.

Segmentation refers to the process of partitioning a digital image into multiple segments (sets of pixels). The goal of segmentation is to simplify and/or change the representation of an image into something that is more meaningful and easier to analyze. More precisely, image segmentation is the process of assigning a label to every pixel in an image such that pixels with the same label share certain characteristics.

In general, a segmentation of an image $f(x, y)$ is a partition of $f(x, y)$ into connected sub-images $R_1, R_2, \ldots, R_n$ such that the following constraints are satisfied: (1) $\bigcup_{i=1}^{n} R_i = f(x, y)$, (2) $R_i \bigcap R_j = \varnothing$ and (3) each sub-image satisfies a predicate or set of predicates.

Where some examples of predicates could be: all pixels in any sub-image R_i must have the same gray level or all pixels in any sub-image R_i must be joined by a connectivity factor.

Specifically, the procedure has been implemented as depicted in the flowchart of Fig. 2.

A binary image can be obtained from a gray scale or color image through an operation that selects a subset of the image pixels as foreground pixels, the pixels of interest in an image analysis task, leaving the rest as background pixels to be ignored. The selection operation can be as simple as the thresholding operator that chooses pixels in a certain range of gray-tones or subspace of color space, or it may be a complex classification algorithm. In this example the thresholding operator has been computed from a histogram with the intensity of the all pixels. The intensity of the pixels in the image with the highest frequency is the threshold. Thus, Fig. 1b shows the binary matrix corresponding to Fig. 1a.

A connected-components labeling of a binary image B is a labeled image LB in which the value of each pixel is the label of its connected-component. The pixels of a binary image are 0's and 1's; the 1's will be used to denote foreground pixels and the 0's background pixels. The term $B[r,c]$ denotes the value of the pixel

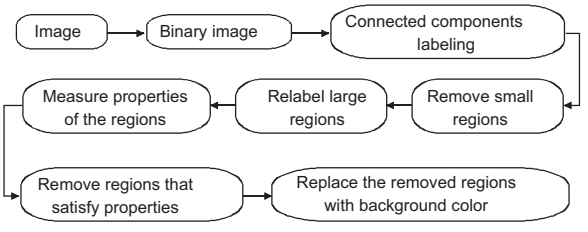


Fig. 2. The flowchart for ERCC.

located at row r, column c of the binary matrix of the image. A binary matrix has M rows and N columns. Thus $B[1,1]$ refers to the value of the upper leftmost pixel and $B[M,N]$ refers to the value of the lower rightmost pixel. The connected-components labeling is associated to a connectivity factor between neighbors. The two most common definitions for neighbors are the four-neighbors and the eight-neighbors of a pixel. The four-neighborhood $N_4(r,c)$ of pixel (r,c) includes pixels $(r-1,c)$, $(r+1,c)$, $(r,c\text{-}1)$ and $(r,c+1)$. The eight-neighborhood $N_8(r,c)$ of pixel (r,c) includes each pixel of the four-neighborhood plus the diagonal neighbor pixels $(r-1,c-1)$, $(r-1,c+1)$, $(r+1,c-1)$ and $(r+1,c+1)$. Fig. 3 shows the labeled matrix (LB) with respect to either the N_4 and the N_8. The elements of LB are integer values greater than or equal to 0. The pixels labeled 0 are the background. The pixels labeled 1 make up one object; the pixels labeled 2 make up a second object; and so on. There are nine and seven objects (regions) for N_4 and N_8 respectively.

Then all regions from LB that have fewer than P pixels are removed, and the remaining regions are relabeled. Once a set of regions has been identified, for each labeled region in the LB it is possible to calculate certain properties. Common properties include geometric properties such as the area of the region and the centroid; shape properties such as measures of the circularity and elongation; and intensity properties such as mean gray tone.

In the discussion that follows, we denote the set of pixels in a region by R. Assuming square pixels, we define the area as:

$$A = \sum_{(r,c)\in R} 1 \tag{1}$$

which means that the area is just a count of the pixels in the region R. The centroid $(\bar{r}, \bar{c})$ is the "average" location of the pixels in the set R.

$$\bar{r} = \frac{1}{A}\sum_{(r,c)\in R} r \quad \text{and} \quad \bar{c} = \frac{1}{A}\sum_{(r,c)\in R} c \tag{2}$$

Finally, the regions which meet the desired conditions in the calculated properties are replaced by background color and small regions that were deleted are restored. In our example the parameters, properties and conditions have been the following: N_8, $P=2$ and elimination of regions which centroid is in the left half plane. The final result is exactly Fig. 1a with the pixels [(2,3), (2,4), (3,3), (4,3)] and [(6,2), (6,3), (7,2), (8,2)] substituted by background color.

$$\begin{bmatrix} 0&0&0&0&0&0&4&0&0&0&8&0\\ 0&0&2&2&0&4&4&4&0&0&0&0\\ 0&0&2&0&0&0&4&0&0&0&9&0\\ 0&0&2&0&0&5&0&5&0&0&0&0\\ 0&0&0&0&0&5&0&5&0&0&0&0\\ 0&1&1&0&0&5&5&5&0&6&6&0\\ 0&1&0&0&0&0&5&0&0&0&6&0\\ 0&1&0&0&0&5&5&5&0&7&0&0\\ 0&0&0&3&0&5&5&5&0&7&7&0\\ 0&0&0&0&0&0&0&0&0&0&0&0 \end{bmatrix} \quad \begin{bmatrix} 0&0&0&0&0&0&4&0&0&0&6&0\\ 0&0&2&2&0&4&4&4&0&0&0&0\\ 0&0&2&0&0&0&4&0&0&0&7&0\\ 0&0&2&0&0&4&0&4&0&0&0&0\\ 0&0&0&0&0&4&0&4&0&0&0&0\\ 0&1&1&0&0&4&4&4&0&5&5&0\\ 0&1&0&0&0&0&4&0&0&0&5&0\\ 0&1&0&0&0&4&4&4&0&5&0&0\\ 0&0&0&3&0&4&4&4&0&5&5&0\\ 0&0&0&0&0&0&0&0&0&0&0&0 \end{bmatrix}$$

(a) N_4 (b) N_8

Fig. 3. Connected-component labeling.

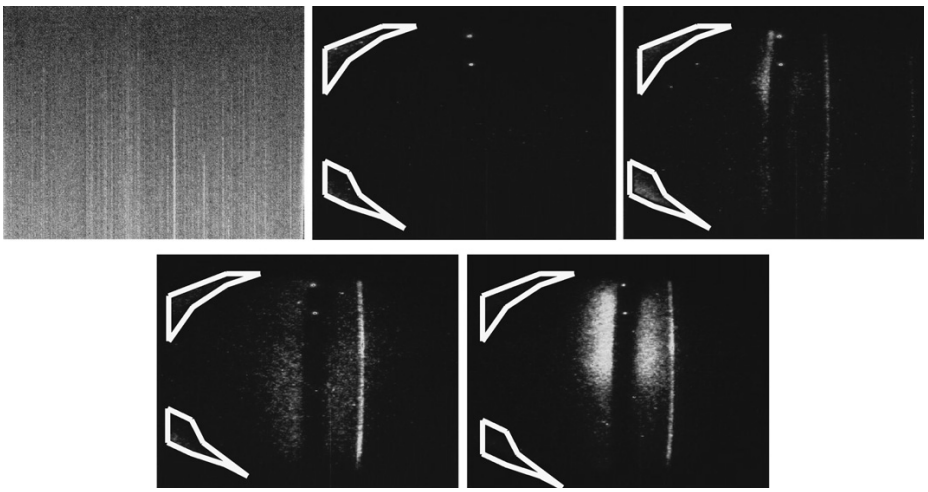


Fig. 4. Classes of images acquired by the TJ-II Thomson Scattering: BKG, STR and COFF at the top and ECH and NBI at the bottom.

3. ERCC for TJ-II TS diagnostic

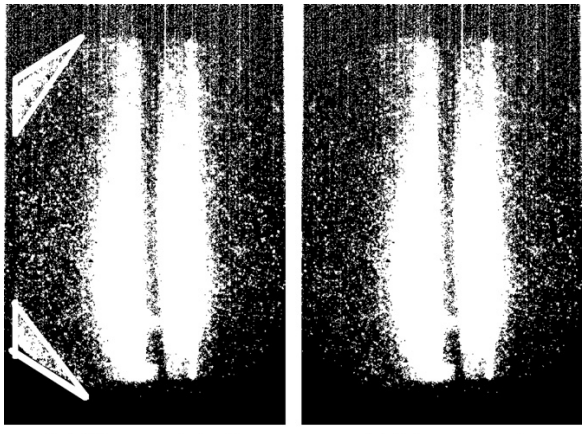
Throughout the life-cycle of a plasma shot, the TJ-II TS diagnostics captures five different kinds of images (576×385 pixels) in gray tones, depending on the plasma state (see Fig. 4): CCD camera background (BKG), measurement of stray light without plasma or in a collapsed discharge (STR), images during ECH phase (ECH, Electron Cyclotron Resonant Heating), during NBI phase (NBI, Neutral Beam Injectors) and after reaching the cut-off density during ECH heating (COFF). From the point of view of plasma physics, the most important images are ECH and NBI because they correspond to high temperature plasmas. In both cases, the image is processed to obtain the radial profiles of the electron density and the electron temperature. The implementation has been programmed in Matlab.

The objective is to eliminate the regions due to stray light without eliminating significant information. These regions have been delimited in Fig. 4 with white polygons.

Fig. 5a shows the binary image corresponding to discharge with NBI.

Fig. 5b shows the binary result image where the stray light regions have been removed.

The parameters, properties and conditions in the procedure applied to TS diagnostic have been the following: the thresholding



(a) Binary original (b) Binary result

Fig. 5. Example of an image with NBI heating.

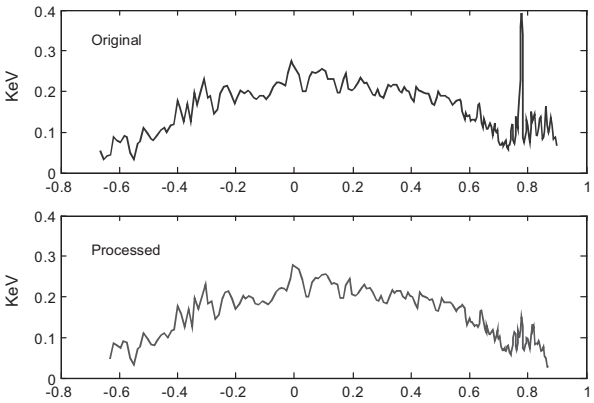


Fig. 6. Radial profiles of the electron temperature.

operator has been computed from a histogram with the intensity of all pixels in the BKG images, N_8, $P = 20$ and elimination of regions which x coordinate of the centroid is lower than 100.

In order to show the validation the overall procedure of ERCC for TJ-II TS diagnostic, the obtained results have been divided in two parts. The first consists in to use a classification system based on support vector machine (SVM) [7]. In the experiments a total of 242 images were used. These images belong to one of the following classes: BKG (50), COFF (42), ECH (50), NBI (50) and STR (50). The 100% STR signals once the method explained above has been applied are classified as BKG signals. Secondly, radial profiles of the electron temperature are compared. Fig. 6 shows the important reduction of the processed profile in the area belonging to the stray light for an image with NBI heating.

4. Conclusions

ERCC method has proven to be useful removing the stray light in the TJ-II TS diagnostic without eliminating significant information. In a future work could be interesting to quantify the removed information.

Acknowledgment

This work was partially funded by the Spanish Ministry of Science and Innovation under the Project No. ENE2008-02894/FTN.

References

[1] H. Rowley, S. Baluja, T. Kanade, Human face detection in visual scenes, Advances in Neural Information Processing Systems 8 (1995) 875–881.
[2] V. Ferrari, L. Fevrier, F. Jurie, C. Schmid, Groups of adjacent contour segments for object detection, IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (2008) 36–51.
[3] C. Gu, J. Lim, P. Arbelaez, J. Malik, Recognition using regions, in: Proceedings of the 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2009, pp. 1030–1037.
[4] B. Fulkerson, A. Vedaldi, S. Soatto, Class segmentation and object localization with superpixel neighbourhoods, IEEE International Conference on Computer Vision (2009) 670–677.
[5] R.P. Breault, Control of stray light, in: Handbook of Optics, vol. I, McGraw-Hill, 1995.
[6] The Mathworks, Image Processing Toolbox, Matlab, R2011.
[7] L. Makili, J. Vega, S. Dormido-Canto, I. Pastor, A. Pereira, G. Farias, et al., Upgrade of the automatic analysis system in the TJ-II Thomson Scattering diagnostic: new image recognition classifier and fault condition detection, Fusion Engineering and Design 85 (2010) 415–418.

Article 4

Making decisions on brain tumor diagnosis by soft computing techniques

4.1 Bibliographic Description

Title

Making decisions on brain tumor diagnosis by soft computing techniques.

Citation

G. Farias, M. Santos, V. Loópez (2010) Making decisions on brain tumor diagnosis by soft computing techniques, *Soft Computing*, ISSN 1432-7643, Volumen 14, Number 12, Pages 1287-1296.

Abstract

In this paper, a synergy of advanced signal processing and soft computing strategies is applied in order to identify different types of human brain tumors, as a help to confirm the histological diagnosis of experts and consequently to facilitate the decision about the correct treatment or the necessity of an operation. A computational tool has been developed that merges, on the one hand, wavelet transform to reduce the size of the biomedical spectra and to extract the main features, and on the other hand, Support

Vector Machine and Neural Networks to classify them. The influence of some of the configuration parameters of each of those soft computing techniques on the clustering is analyzed. These two methods and another one based on medical knowledge are compared. The classification results obtained by these computational tools are promising specially taking into account that medical knowledge has not been considered and that the number of samples of each class is very low in some cases.

References

I. Daubechies(1992); H. Demuth, M. Beale(1998); R. Duda, P. Hart, D. Stork(2001); G. Farias, M. Santos(2005); G. Farias, M. Santos, V. López (2008); ML. García-Martín et al. (2001); G. Hagberg (1998); S. Haykin(1999); P. Hore(1983); SL Howells, R. Maxwell, JR Griffiths(1992); I. Kim, J. Watada, I. Shigaki(2008); Y. Kinoshita Y, A. Yokota (1997); P. Laguna et al. (1999); SG Mallat(2001); I. Martínez-Pérez et al. (1995); MathWorks (1989); RJ Maxwell et al. (1998); M. Misiti et al. (1997); J. Pascual et atl. (1998); J. Peeling, G. Sutherland(1992); D. Rafiei D, A Mendelzon(1998); JM Roda et al. (2000); R. Rojas(1995); JC Ruiz-Molina, J. Navarro-Moreno, A. Oya(2001); RL Somorjai et al. (1996); AR Tate (1997); AR Tate et al. (1998); CW Therrien(1992); M. Unser, A. Aldroubi(1995); VN Vapnik(2000); X. Wen et al. (2009); P. Yin et al. (2008).

Impact Factor

Soft Computing has an impact factor of 1.88 according to Thomson Reuters Journal Citation Reports (2011).

Soft Comput (2010) 14:1287–1296
DOI 10.1007/s00500-009-0495-0

FOCUS

Making decisions on brain tumor diagnosis by soft computing techniques

G. Farias · M. Santos · V. López

Published online: 16 September 2009

Abstract In this paper, a synergy of advanced signal processing and soft computing strategies is applied in order to identify different types of human brain tumors, as a help to confirm the histological diagnosis of experts and consequently to facilitate the decision about the correct treatment or the necessity of an operation. A computational tool has been developed that merges, on the one hand, wavelet transform to reduce the size of the biomedical spectra and to extract the main features, and on the other hand, Support Vector Machine and Neural Networks to classify them. The influence of some of the configuration parameters of each of those soft computing techniques on the clustering is analyzed. These two methods and another one based on medical knowledge are compared. The classification results obtained by these computational tools are promising specially taking into account that medical knowledge has not been considered and that the number of samples of each class is very low in some cases.

Keywords Decision making · Soft computing · Neural networks · Support vector machines · Wavelets · Classification · Medical diagnosis

G. Farias
Dpto. Informática y Automática, Escuela Superior de Ingeniería Informática, UNED, 28040 Madrid, Spain
e-mail: gfarias@bec.uned.es

M. Santos · V. López (✉)
Dpto. Arquitectura de Computadores y Automática, Facultad Informática, UCM, 28040 Madrid, Spain
e-mail: vlopez@fdi.ucm.es

M. Santos
e-mail: msantos@dacya.ucm.es

1 Introduction

A main concern in the medical environment is the development of non-histological methods of diagnosis based on in vitro ^{1}H magnetic resonance spectroscopy (MRS) biopsies of human brain tumors. Histological procedures remain mandatory for tumor diagnosis. However, pathologist may find computational alternatives useful in cases where a confirmation of the histological diagnosis by an independent method is advisable or in situations in which an adequate anatopathological examination cannot be performed. Furthermore, there are only a few specialists who are able to analyze, in a right way, the relevant biochemical symptoms of MRS (Kinoshita and Yokota 1997), and the training to get this knowledge is slow and laborious. It must be mentioned that medical systems are complex systems involving inexact, uncertain, imprecise, and ambiguous information (Kim et al. 2008).

The progress in statistical techniques and in pattern recognition suggests the development of automatic procedures that could be implemented in surgical spectrometers. But it would be necessary to incorporate some heuristics, and soft computing can provide with useful tools to deal with this information. They are not intrusive methods and at the same time, they incorporate the knowledge of the experts if it is available.

Different strategies have been applied to deal with this medical problem of pattern recognition such as linear discriminant analysis, statistical correlation, different neural networks, clusters analysis, etc. (Hagberg 1998; Howells et al. 1992; Martínez-Pérez et al. 1995; Maxwell et al. 1998; Tate 1997; Pascual et al. 1998; Tate et al. 1998; Roda et al. 2000; Farias et al. 2008). Most of those papers are based on specialized medical knowledge, which is not always available. Furthermore, the final classification is

strongly influenced by the measurement conditions (Somorjai et al. 1996) and by the classifier parameters.

In this work, the two stages of the classification process of these signals are described. First, a processing phase is applied, where the signals are bounded to the range of interest, the noise is reduced, the waveforms are normalized, etc. At the same time, Wavelet transform is applied in order to extract the relevant information. The row spectra are large and complex and the application of this data compression technique helps to reduce the size of the spectra and to obtain the main features while filtering the noise. On the other hand, wavelets are suitable for the analysis of non-stationary signals (Daubechies 1992).

Then, soft computing techniques such as neural networks (NN) and support vector machines (SVM) are applied to match the spectra to the type of tumor. The results of the diagnosis obtained by those strategies and some of the proposed in other papers are compared.

The novelty of the proposal is that the characterization of these signals is not based on the medical knowledge of the molecular or metabolic profiles. The method uses just computational information in an automatic procedure.

Therefore, a hybrid computational tool that merges preprocessing and soft computing techniques for clustering is developed: an intelligent classifier that applies wavelets and either neural networks or support vector machines. The results are encouraging taking into account that medical information is not considered. This tool can help to make a decision about the right medical treatment according to the diagnosis.

The paper is organized as follows. In Sect. 2, the brain tumor diagnosis problem is introduced. Section 3 presents the wavelet transform as a suitable tool for processing medical spectra. Section 4 shows the implementation of the SVM and NN classifier. Results of human brain tumors classification by those computational tools are discussed in Sect. 5. Conclusions end the paper.

2 Medical signals processing

Nuclear magnetic resonance (NMR) has provided a great help in the knowledge of different pathologies and also to establish the relations between different lesions. This particular aspect has a great importance in tumor pathology (Pascual et al. 1998; Roda et al. 2000; García-Martín et al. 2001). It is not an invasive method and it is able to determine, both in a qualitative and in a quantitative way, a great variety of metabolites in each tissue, giving significant information. Nevertheless, its diagnostic application has been limited due to the fact that the in vivo ^{1}H NMR spectra only give small accuracy and because of the difficulties in the quantification of the metabolites.

Most of the limitations could be overcome if extracts of tumor biopsies are available, by applying the in vitro technique (Peeling and Sutherland 1992). This method has been used to obtain the spectra that are going to be analyzed in this paper. Samples from normal brain and different tumor tissues were obtained after craniotomy. The preparation and characterization of biopsies are described in Roda et al. (2000). The main drawback of this method is the difficulty to obtain a large database of biopsies as each of them comes from a patient that has suffered an operation.

The spectra that have been obtained in this way are made up of thousands of data. In fact, each spectrum has 16,384 samples taking into account only the real part. The spectrum of each tumor (Fig. 1) represents the intensity—proportional to the concentration of protons in the tissue—(*y* axis) versus the distance in ppm (part per million) (*x* axis), i.e., at a particular resonance frequency. This frequency depends

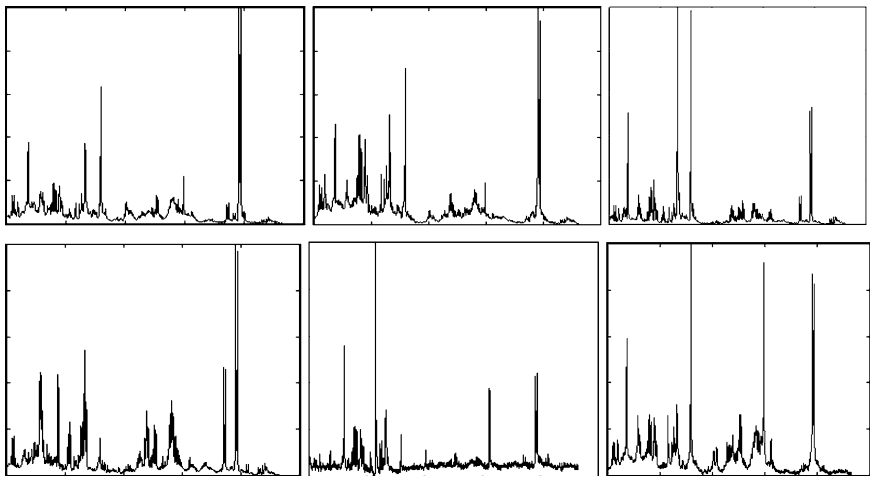


Fig. 1 Patterns of different human brain tumors: from *left to right*, high-grade astrocytoma, low-grade astrocytoma, medulloblastoma, meningioma, neurinoma, normal brain

on the magnetic field. It is not an absolute value but a ratio: the chemical displacement, δ, defined as,

$$\delta = \frac{\Delta v}{f} \times 10^6 [\text{ppm}] \quad (1)$$

where f is the frequency of resonance in the magnetic field (8 T) and Δv is the difference between,

1. the point we are on and the reference point (0 ppm) in Hz, and
2. the frequency of observation.

Before the signal processing, some of the spectra presented different resolution (8, 16, and 32 kb). Therefore, the first step was to adjust all the signals to the same size (16,384), applying zero padding when the number of samples needed to be enlarged. Due to this, the interval between points is different for each signal, in spite of the fact that the interval is constant for every particular signal. The size of the mentioned interval ΔP (in ppm) was computed by the absolute value of the difference between the final P_f and the initial P_i points of the spectrum, dividing it by the total number of points minus one, as it is shown in Eq. 2.

$$\Delta P = \frac{|P_f - P_i|}{16,384 - 1} [\text{ppm}] \quad (2)$$

Most of the spectra had values for P_i and P_f about 10 and 1 ppm, respectively. Once the size of the interval between samples of each signal has been obtained, we proceed to extract the amplitudes in the region of interest, which covers the range from 0.8 ppm up to 4.2 ppm, according to the information provided by the medical experts. The number of points of each spectrum is now 4,655.

The calibration of the intensity for the signals was also different depending on how they were obtained. For that reason, a normalization procedure is applied. As it is shown in Eq. 3, the normalized amplitude is obtained by dividing the value of the intensity $I(\delta)$ of every point of the spectrum by the maximum of the intensity I_{Max} (Roda et al. 2000).

$$I_{\text{norm}}(\delta) = \frac{I(\delta)}{I_{\text{Max}}} \quad (3)$$

Therefore, the signals have been normalized in both, the resonance intensity and the number of samples of the spectra. The data are thus prepared to be classified.

The one-dimensional ^{1}H MRS can be classified in eight different groups that correspond to normal brain and seven different classes of human brain tumors, as it has been stated by the World Health Organization (WHO).

Logically, there is a great difficulty to get these biopsies, as each of them requires a cranial operation. Although the Hospital "La Paz" (Madrid, Spain) has a large database of human brain tumors, the experimental data set is made up of 112 biopsies so far. This means that there are classes with few examples and that will make difficult the training of the classifier. In Table 1, the different classes of these tumors and the number of samples of each of them are listed. Figure 1 shows some tumor patterns.

Table 1 Tumor classes and number of samples

Tumor	No. of samples	Class
High grade astrocytoma HGA	12	1
Low grade astrocytoma LGA	16	2
Normal brain NB	16	3
Medulloblastoma MB	4	4
Meningioma MG	31	5
Metastasis MT	14	6
Neurinoma NN	9	7
Oligodendroglioma OD	10	8

Another extra difficulty when identifying the tumors is that, as it is possible to see in Fig. 1, there are only few differences between the profiles of the different tissue classes, at least without applying specialized medical knowledge. Furthermore, the biopsies have been obtained in different experimental conditions and they appear to be different even though they belong to the same class.

For these reasons, the classification is not an easy task. Finally, as we are not applying histological knowledge, the lack of meaning of the peaks in the spectra may complicate the diagnosis.

3 Feature extraction with wavelets

Once the brain signals are normalized according to the procedures of Sect. 2 and the ^{1}H spectra is now concentrated in the range of 0.8–4.22 ppm (4,655 points), several techniques can be used in different domains to obtain the main features that characterize each tumor tissue. Besides, because of the large size of each spectrum, compression methods are required in order to reduce the number of attributes of each signal.

One of these methods is the Karhunen–Loève (KL) representation, which works with the autovectors of the correlation matrix (Therrien 1992). Although it is an efficient representation, some studies have shown some drawbacks. For example, in most of the cases, the correlation matrix is unknown, the computation of the autovectors has a complexity that is a cubic function of the dimension, and finally it depends on the input data (Laguna et al. 1999; Ruiz-Molina et al. 2001). In this sense, the Fast Fourier Transform is independent of the set of data, and there are many fast algorithms for its implementation. However, it is not able to eliminate the correlation for big data set (Hore 1983).

Another method is based on finding similar time sequences by the discrete Fourier transform (DFT) (Rafiei

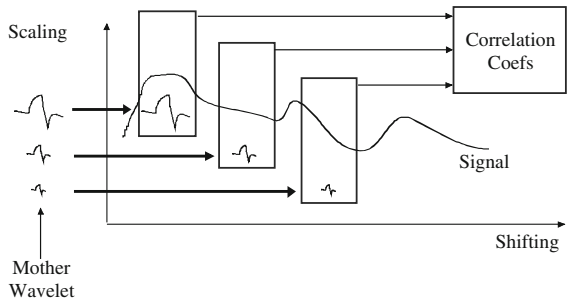


Fig. 2 Wavelet transform processing

and Mendelzon 1998). This method uses the DFT to extract the main characteristics of the signal and to reduce the dimensionality of the feature vectors. The small order coefficients are enough to describe signals that vary slowly. Besides, by applying the Fourier coefficients, it is possible to establish a correspondence between each signal and a point in the Fourier series multidimensional space. However, the DFT has difficulties when it is used with fast varying waveforms; time information may be lost when is transformed into the frequency domain and besides, transitory characteristics cannot be then detected. As medical spectra change abruptly, this technique is not suitable.

On the other hand, wavelet transform (WT) is a very powerful computational tool (Unser and Aldroubi 1995; Mallat 2001; Daubechies 1992). It is based on the selection of an optimum database to work with. Its application allows a high level of compression without losing information. The redundant information is minimized and so the computational load is substantially cut down (Mallat 2001). This point is especially important as the consideration of redundant and irrelevant features has negative effects on classification task (Yin et al. 2008). Some of the properties that make wavelets so useful in pattern recognition are the capacity of noise reduction, the signal enhancement, and the detection of similarities. They allow analyzing periodic and non-periodic signals. It is a time-scale approach, which allows understanding the results in the time–frequency plane.

The discrete wavelet transform (DWT) works with the so-called Mother Wavelet which is a prototype function. The mother wavelet is scaled and shifted to be compared with the original signal (Fig. 2). Because of this comparison, the DWT computes a few correlation coefficients at different decomposition levels for each signal in a fast way. From these coefficients, it is possible to reconstruct the original signal by using the inverse Wavelet Transform.

The coefficient matrixes are obtained by filtering and down sampling. These types of coefficients are:

- Approximation (A): a smoothed and sub-sampled version of the original signal. It represents the coarse approximation.
- Detail Coefficients (D): it represents the high frequency components, i.e., the discontinuities or abrupt changes in the signals.

For example, Fig. 3 shows four different decomposition levels of the normal brain spectrum when the Haar mother wavelet is applied.

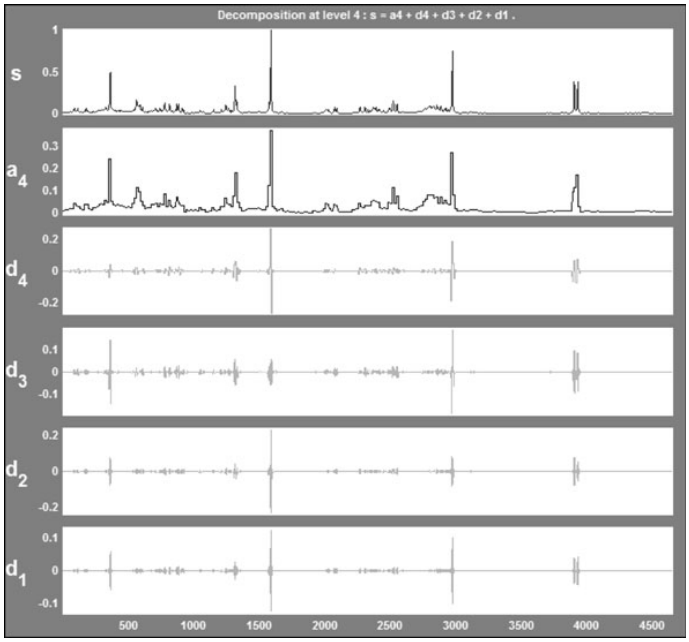


Fig. 3 Wavelet transform of the normal brain spectrum. Signal (s), approximation (a4), and details from level 1–4

After applying wavelets, the number of data is reduced in an exponential way while the decomposition level increases. In this case, each spectrum has been reduced from 4,655 samples to 291 attributes at decomposition level 4. Selecting the most suitable family of mother wavelets and the best scale for particular signals is a difficult task (Farias and Santos 2005).

4 The classifier

Our purpose is to develop a computational tool that fulfills the following requirements:

- Accuracy in the diagnosis (high percentage of correct classifications).
- Easy to integrate in a surgery environment.
- Friendly use and easy to apply, i.e., it does not require specialized skills.
- Open to modifications and improvements.

The initial hypothesis is that there should be a pattern for each type of neoplasm. That is, there will be a pattern that represents all the signals of every tissue class. The similarity between signals means they share some characteristics or properties.

The second assumption is that the characterization of each type of tumor is possible without the histological interpretation of the profile, only with computational processing.

The classifier has been implemented in MATLAB (MathWorks 1989; Demuth and Beale 1998; Misiti et al. 1997). The developed computational application allows not only to classify biomedical spectra and any other kind of signals or even images, but also to evaluate the performance of different classifier structures. These classifier configurations can be easily obtained by modifying some of the parameters such as:

- In the pre-processing step: the wavelet mother, the decomposition level, the coefficient A or D, etc.
- In the SVM classifier: the kernel, the order of the polynomial, coefficients of the base, etc.
- In the Neural classifier: the number of layers and neurons, the activation functions, epochs, error goal, etc.

The spectra can be displayed in the image window at the left side (Figs. 4, 6). Once the type and parameters of the WT have been chosen, the application displays the Wavelet Transform of the previously selected signal. The View option allows showing either the original spectrum or its wavelet transform at any stage (Figs. 4, 6).

The classification process starts when pressing the Generate button: two sets of signals are then randomly

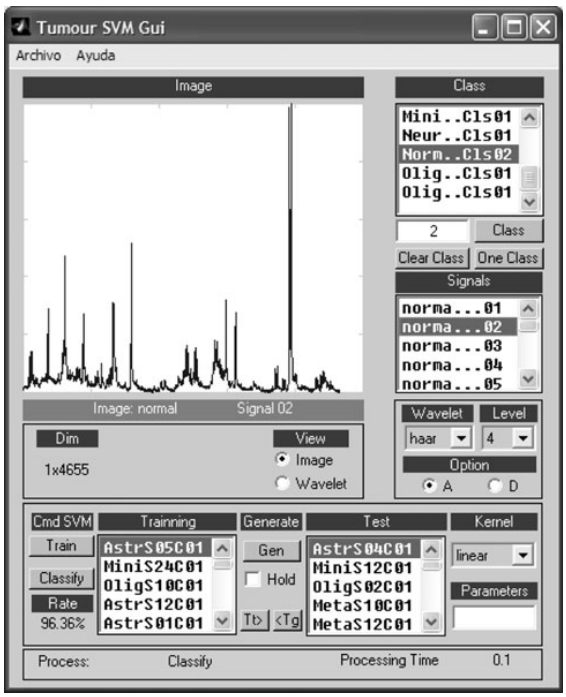


Fig. 4 Graphical user interface of the Wavelet-SVM classifier

obtained for training and testing purposes. The proportion of signals of each set can be defined by the user. In addition, the leave-one-out (LOO) method is available. As there are few examples of each class, this method has been used to train the system in some cases.

After pressing the Train button, the classifier starts working. To evaluate the results, it is necessary to press the Classify button. The classification process ends when the classifier decides whether a spectrum belongs to a class or to any other. Correct diagnosis is obtained when the class selected by the computational tool matches the histological diagnosis. Automatically, the classifier compares the obtained results with the labeled classes, giving the percentage of correct classifications (success) and showing the processing time.

Between the different strategies to tackle the multi-class problem, there are two classical ones. One is to break the multi-class problem into a series of binary classifiers, e.g., one-versus-one and one-versus-all. The other strategy is to consider all classes in one optimization formulation to form a single machine. The three of them have been applied.

4.1 The wavelet-SVM classifier

Support vector machine (SVM) is a very effective method for pattern recognition (Duda et al. 2001; Vapnik 2000). In

brief, given a set of input vectors, which belong to different classes, the SVM maps the inputs into a high-dimensional feature space through some non-linear mapping, where an optimal separating hyperplane is generated in order to minimize the risk of misclassification. The hyperplane is determined by a subset of points of the classes, named support vectors (SV). This technique has been applied for different clustering application.

Figure 4 shows an example where the 96.36% of the signals was rightly classified (hits) and the computational time was 0.1 s. In this case, the kernel was linear.

In this Fig. 4, it is possible to see how the user can chose different kernels and tune the parameters associated to each of them.

4.2 The wavelet-NN classifier

Neuronal networks (NN) have been successfully applied in a great number of classification problems (Duda et al. 2001). There are many types of NN with different structure that can be applied depending on the application characteristics. In any case, a NN consists of some basic processing elements called neurons, which are grouped in layers and connected by synapses connections that are weighted by a factor (Haykin 1999; Rojas 1995). Combinations of wavelets and NN have been applied in different fields (Wen et al. 2009).

In this work, a multilayer perceptron (MLP) with supervised learning is used. Figure 5 shows the neural network structure that has been used after trying different configurations. The NN has an input layer for the 291 attributes that have been generated by the WT. The number of inputs has been notably reduced due to the compression made by the wavelets. Two hidden layers with 140 and 70 nodes are implemented, with *Tansig* activation function (Eq. 4). The two output neurons for binary classification have function of activation *Logsig* (Eq. 4). After training the NN, every signal is associated to its corresponding class through the activation function.

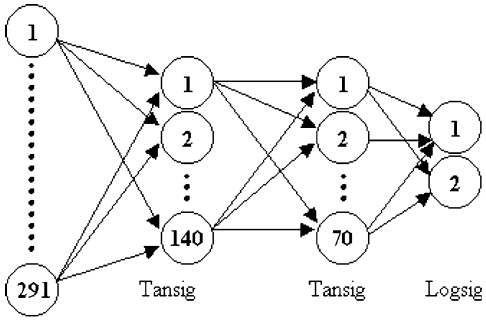


Fig. 5 Structure of the proposed NN

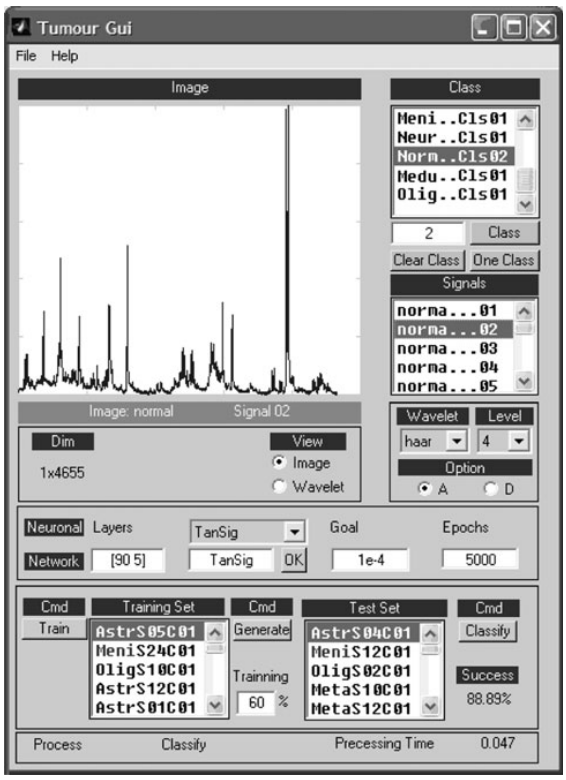


Fig. 6 Graphical user interface of the wavelet-neural classifier

$$Tansig(n) = \frac{2}{1+e^{-2n}} - 1; \quad Logsig(n) = \frac{1}{1+e^{-n}} \tag{4}$$

After pressing the Train button, the NN works until it reaches the error goal. The training process results a set of weights through the Back-Propagation algorithm, which is derived by computing the error between the values of the output layer and the desired values. The classification process ends when the trained NN decides whether a spectrum belongs to a class or to any other.

To evaluate the results, it is necessary to press the Classify button. Automatically the classifier compares the obtained results with the labelled classes, giving the percentage of hits and the processing time (Fig. 6).

5 Results

After trying different configurations for the processing, the Mother Wavelet Haar at decomposition level 4 with approximation coefficient has been selected.

To test the classifier, many experiments were carried out and the results presented in this paper are the average. It is needed to remark that the number of available spectra is only 112.

Correct diagnosis (hits) is obtained when the class selected by the computational tool for a specific signal matches the right diagnosis given by the histologist.

5.1 WT-SVM results

First, a binary classification was carried out. Using this strategy, it is possible to classify the complete data set into only two groups, normal brain or any other tumor pathology. The best score provided for correct classification was 96.36% with computational time 0.1 (Fig. 4).

After that, binary comparisons were performed between every tissue class and each one of the remaining ones (Table 2). Classification between two classes may yield different scores for each class, depending on the number of elements and the number of correct classifications in each class. For instance, the first row depicts the comparison of high-grade astrocytoma (HGA) versus any other class of tumor. When compared with medulloblastoma (MB), the correct score was 91.66%. That is, 11 extracts of the 12 biopsies of high-grade astrocytoma class were correctly classified when compared to this other class of tumor. Nevertheless, when comparing medulloblastoma versus high-grade astrocytoma, the percentage of hits was the 25%. This means that only one out of 4 samples of MB were rightly classified when compared with the 12 samples of HGA. Similar interpretations are applicable to the rest of the rows.

The best results were obtained for normal brain (NB) versus medulloblastoma (MB), and for meningioma (MG) versus normal brain (NB). As it is possible to see in Table 2, the scores reached 100%. That may be because the number of samples of these classes, normal brain (16) and meningioma (31), is the largest. The worst case was the comparison between medulloblastoma and high-grade astrocytoma, because of the scant number of samples of medulloblastoma, only 4, and may be because the profiles of both tissue classes are quite similar and so it is difficult to distinguish between them without extra knowledge.

Table 2 Average percentage of correct classifications for binary SVM classification

	HGA	LGA	NB	MB	MG	MT	NN	OD
HGA		58.33	83.33	91.66	75	66.66	91.67	66.66
LGA	81.25		87.5	81.25	87.5	81.25	68.75	75
NB	87.5	93.75		100	93.75	93.75	93.75	93.75
MB	25	75	75		75	75	75	75
MG	90.32	90.32	100	90.32		87.09	90.32	93.54
MT	50	85.71	85.71	78.57	78.57		64.28	71.42
NN	77.77	77.77	88.88	88.88	77.77	66.66		77.77
OD	40	50	90	90	60	70	60	

Table 3 Average percentage of correct classifications for one vs. the rest WT-SVM

	HGA	LGA	NB	MB	MG	MT	NN	OD
WT-SVM	87	86	94	89	94	88	95	95

In any case, as it possible to see in Table 2, the results of the fourth row, corresponding to MB, are quite low and they depend on how many of the four samples of MB are rightly classified.

The one-versus-the-rest method has also been applied, with encouraging results. In this case, the Wavelet-SVM classifier gives good scores for all the classes, as it is possible to see in Table 3.

Nevertheless, when we tried multi-class classification, the scores were much lower, although they present higher accuracy in the classification process than when considering the effects of chance taking into account the eight different classes (12.5%). The Medulloblastoma class was discarded because of the small number of available samples of this tumor. These final values were obtained by carrying out more than 30 experiments to calculate average values (Fig. 7). As it is possible to see, NB (class 3) reaches the 100% of hits. The worst values are obtained for LGA (class 2) and OD (class 8), maybe because of their profiles and the small number of samples of these classes.

The number of samples for training and testing was randomly selected. It is possible that using the LOO strategy the results would improve.

5.2 WT-NN results

When binary classification (normal brain vs. any other class) was carried out using the neural classifier, the best scores provided for correct classification were 95.7%.

Then binary comparisons were carried out between every tissue class and each one of the remaining ones (Table 4).

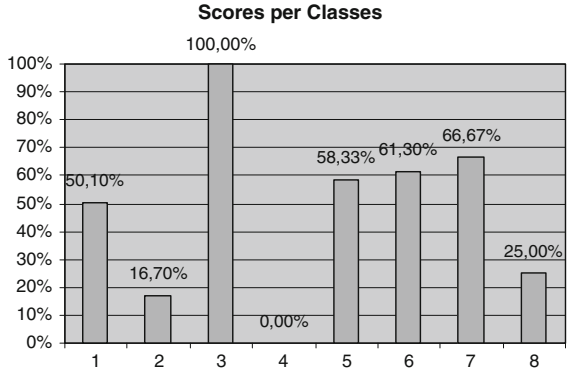


Fig. 7 Results for each one of the classes when applying WT-SVM multi-classification

Table 4 Average percentage of correct classifications for binary NN classification

	HGA	LGA	NB	MB	MG	MT	NN	OD
HGA		67	75	92	84	58	100	59
LGA	75		94	81	82	88	81	82
NB	88	94		94	94	88	94	94
MB	50	50	75		75	50	75	75
MG	87	90	90	97		87	84	90
MT	57	93	93	86	79		86	93
NN	67	56	89	56	78	67		89
OD	50	60	80	80	50	70	80	

The first row depicts the comparison of high-grade astrocytoma and any other of the rest. For instance, when compared with low-grade astrocytoma, the correct score was 66.7%. That is, 8 extracts of the total of 12 biopsies of high-grade astrocytoma class were correctly classified.

Again, the best results were obtained for meningioma (31 samples) versus medulloblastoma (four samples). As it is possible to see in Table 4, the scores were 97% of hits in this case. In addition, the row corresponding to normal brain shows good percentage of hits, that is, normal brain versus any other class. The worse case was the comparison between medulloblastoma and other classes, as it was when the SVM classifier was applied, for the same reasons that were stated in Sect. 5.1.

The one-versus-the-rest method has also been applied, with promising results. In this case, the Wavelet-NN classifier gives good scores for all the classes, as it is possible to see in Table 5.

Multi-class classification was also applied. The output layer of the NN (Fig. 5) was set to eight neurons, involving the eight possible classes considered. The multiple-choice comparisons allow the classification of any arbitrarily chosen sample of the database into any of the eight human brain tumors. For training and testing, the number of elements of each set was randomly selected. The final values were obtained by carrying out more than 30 experiments to calculate the average values.

The scores obtained after the multiclassification procedure represent higher accuracy in the classification process than when considering the effects of chance (12.5%) but they are not good enough (Fig. 8). As it is possible in that figure, the best scores correspond to NB (92%) and MG (79%), the classes with higher number of samples.

Table 5 Average percentage of correct classifications for one vs. the rest WT-NN

	HGA	LGA	NB	MB	MG	MT	NN	OD
WT-NN	83	85	96	96	92	87	91	88

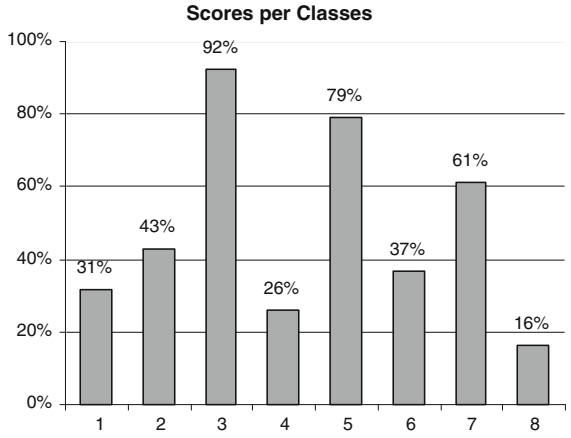


Fig. 8 Results for each one of the classes when applying WT-NN multi-classification

Again, the results could possibly improve if LOO strategy were applied.

5.3 Comparison between classification methods

The results of applying linear discriminant (LD) to the same signals for binary classification are also available and are shown in Table 6 (Roda et al. 2000).

As it is possible to see in some cases, the scores are better than the ones obtained by SVM or NN, although the results are quite similar and good enough in any case. It is needed to remark that when LD is applied in this problem, the medical knowledge about the discriminant characteristics is used.

In Table 7, the best row of the three methods: LD, SVM and NN, for binary classification are shown. This row corresponds to the comparison of normal brain versus any of the rest of the tumors. As it is possible to see, the best results are obtained by LD that gets the 100% of hits in some cases, but SVM and NN reach quite good scores. It may be explained because the medical description of the

Table 6 Average percentage of correct classifications for binary LD classification

	HGA	LGA	NB	MB	MG	MT	NN	OD
HGA		84	100	100	100	100	100	89
LGA	90		90	90	90	100	100	100
NB	89	100		100	100	100	100	100
MB	75	75	100		50	100	100	100
MG	89	94	94	94		89	94	89
MT	86	86	71	100	100		71	100
NN	100	86	100	86	100	100	100	
OD	75	100	75	100				

Table 7 Average percentage of hits of NB class for binary LD, SVM and NN classification

	HGA	LGA	NB	MB	MG	MT	NN	OD
LD	89	100	–	100	100	100	100	100
SVM	88	94	–	100	94	94	94	94
NN	88	94	–	94	94	88	94	94

Table 8 Comparison of the three methods for one-versus-the-rest classification

	HGA	LGA	NB	MB	MG	MT	NN	OD
WT-SVM	87	86	94	89	94	88	95	95
WT-NN	83	85	96	96	92	87	91	88
LD	NA	NA	100	NA	95	86	NA	75

normal brain profile that is introduced when working with LD helps to identify this class.

Finally, although there are no data available for all the classes with LD method, in Table 8 the results of applying the one-versus-the rest-method are presented. In this case, the SVM gives the best scores of hits in most of the cases although it does not reach the 100% as LD does for NB class.

6 Conclusions

In this paper, an approach that combines advanced pre-processing techniques and soft computing clustering for the classification of biomedical spectra is presented. In the first stage, wavelet transform is applied to reduce the dimension of the features vector. After that, soft computing techniques as neural networks and support vector machines are used to classify those brain tumors using the pre-processed signals as input space.

Therefore, two computation tools that merge wavelets and SVM, and wavelets and Neural Networks have been developed. Those intelligent classifiers allow to observe the influence of the design parameters of each technique on the clustering, and therefore to reduce the classification time and to improve the results.

A relevant conclusion drawn from the comparison of the scores obtained with these computational tools and the others that come from alternative procedures based on medical knowledge (Hagberg 1998; Howells et al. 1992; Martínez-Pérez et al. 1995; Tate 1997; Roda et al. 2000; Hore 1983) is that these soft-computing-based strategies work reasonably well, bearing in mind that qualitative information is not used. The percentage of correct classification with these soft computing methods may be a bit low, although it reaches the 100% in some cases. However, a significant advantage of the proposed tools is that they allow non-specialists to classify any sample of the database without applying medical knowledge. It is also necessary to emphasize that the training data set of the classifier was very limited. Therefore, if the number of available spectra increases, it will be possible to improve the scores of hits.

Another way of improving the results could be to try different configuration parameters of the classifiers, especially the ones related to the training.

Finally, it is possible to consider the incorporation of medical knowledge in the pre-processing phase in order to eliminate the signals that do not add anything and make more difficult the classification task, and to confirm the labels of the spectra.

These tools could help the histologists to make a decision and to confirm the diagnostic, and constitute an alternative for automated classification of biomedical spectra.

Acknowledgments We appreciate the collaboration of the members of the Neurosurgery Service at University Hospital of La Paz (Madrid, Spain) and the Biomedical Research Institute (CSIC), who have provided us with the signals and data used in this work.

References

Daubechies I (1992) Ten lectures on wavelets: CBMS lectures series. SIAM, Philadelphia

Demuth H, Beale M (1998) Neural network toolbox user's guide. The MathWorks Inc, USA

Duda RO, Hart PE, Stork D (2001) Pattern classification. Wiley, New York

Farias G, Santos M (2005) Analysis of the wavelet transform parameters in images processing. LNCCS 2:51–54

Farias G, Santos M, López V (2008) Brain tumour diagnosis with wavelets and support vector machines. Proceedings of the 3rd IEEE international conference on intelligent systems and knowledge engineering, pp 1453–1459

García-Martín ML, Hérigault G, Rémy CH, Farion R, Ballesteros P, Coles J, Cerdán S (2001) Mapping extracellular pH in rat brain gliomas "in vivo" by H magnetic resonance spectroscopic imaging: comparison with maps of metabolites. Cancer Res 61:6524–6531

Hagberg G (1998) From magnetic resonance spectroscopy to classification tumors: a review of pattern recognition methods. NMR Biomed 11:148–156

Haykin S (1999) Neural networks: a comprehensive foundation, 2nd. edn. Prentice-Hall, Upper Saddle River, NJ

Hore P (1983) Solvent suppression in Fourier Transform nuclear magn. reson. J Magn Reson 55:283–300

Howells SL, Maxwell R, Griffiths JR (1992) An investigation of tumor ^{1}H NMR spectra by pattern recognition. NMR Biomed 5:59–64

Kim I, Watada J, Shigaki I (2008) Complementary case-based reasoning and competitive fuzzy cognitive maps for advanced medical decision. Soft Comput 12:191–199

Kinoshita Y, Yokota A (1997) Absolute concentrations of metabolites in human brain tumours using in vitro proton magn. reson. spectroscopy. NMR Biomed 10:2–12

Laguna P, Moody GB, García J, Goldberger AL, Mark RG (1999) Analysis of the ST-T complex using the KL transform: adaptative monitoring and alternant detection. Med Biol Eng Comput 37(2):175–189

Mallat SG (2001) A wavelet tour of signal processing, 2nd edn. Academic Press, San Diego

Martínez-Pérez I, Maxwell RJ, Howells SL, van der Bogaart A, Mazucco R, Griffiths JR, Arús C (1995) Pattern recognition analysis of ^{1}H NMR spectra from human brain tumours biopsies. Proceedings Soc magnetic resonance. 3rd annual meeting Abstract P1709

MathWorks Inc (1989) MATLAB®, MA, USA

Maxwell RJ, Martínez-P. I, Cerdán S, Cabañas ME, Arús C, Moreno A, Capdevila A, Ferrer E, Bartomeus F, Aparicio A, Conesa G, Roda JM, Carceller F, Pascual JM, Howells SL, Mazzuco R, Griffiths J (1998) Pattern recognition analysis oh ^{1}H NMR spectra from perchloric acid extracts of human brain tumor biopsies. Magn Reson Med 39:869–877

Misiti M, Misiti Y, Oppenheim G, Poggi J-M (1997) Users Guide: Wavelet Toolbox for use with MATLAB. The MathWorks, Inc, Natick

Pascual J, Carceller F, Cerdán S, Roda JM (1998) Diagnóstico diferencial de tumores cerebrales "in vitro" por espectroscopia de resonancia magnética de protón. Método de los cocientes espectrales. Neurocirugía 9:4–10

Peeling J, Sutherland G (1992) High-resolution ^{1}H NMR spectroscopy studies of extracts of human cerebral neoplasm. Magn Reson Med 24:123–136

Rafiei D, Mendelzon A (1998) Efficient retrieval of similar time sequences using DFT. Proceedings of the 5th international conference on foundations of data organization, pp 249–257

Roda JM, Pascual JM, Carceller F, Gonzalez-Llanos F (2000) Nonhistological diagnosis of human cerebral tumors by ^{1}H magnetic resonance spectroscopy and amino acid analysis. Clin Cancer Res 6:3983–3993

Rojas R (1995) Neural networks: a systematic introduction. Springer, New York

Ruiz-Molina JC, Navarro-Moreno J, Oya A (2001) Signal detection using approximate Karhunen-Loève expansions. IEEE Trans Inf Theory 47(4):1672–1680

Somorjai RL, Dolenko B, Nikulin AK, Pizzi N, Scarth G, Zhilkin P, Halliday W, Fewer D, Hill N, Ross I, West M, Smith ICP, Donnelly SM, Kuesel AC, Bière KM (1996) Classification of ^{1}H MR spectra of human brain neoplasms: the influence of preprocessing and computerized consensus diagnosis on classification accuracy. J Magn Reson Imaging 6:437–444

Tate AR (1997) Statistical pattern recognition for the analysis of biomedical magnetic resonance spectra. J Magn Reson Anal 3:63–78

Tate AR, Griffiths JR, Martínez PI, Moreno A, Barba I, Cabañas M, Watson D, Alonso J, Bartumeus F, Isamat F, Ferrer I, Vila F, Ferrer E, Capdevilla A, Arús C (1998) Towards a method for automated classification of ^{1}H MRS spectra from brain tumours. NMR Biomed 11:177–191

Therrien CW (1992) Discrete random signals and statistical signal processing. Prentice-Hall, Upper Saddle River, NJ

Unser M, Aldroubi A (1995) A review of wavelets in biomedical applications. Proc IEEE 48:626–638

Vapnik VN (2000) The nature of statistical learning theory, 2nd edn. Springer, New York

Wen X, Zhang H, Xu X, Quan J (2009) A new watermarking approach based on probabilistic neural network in wavelet domain. Soft Comput 13:355–360

Yin P, Sun F, Wang C, Liu H (2008) An adaptive feature fusion framework for multi-class classification based on SVM. Soft Comput 12:685–691

Article 5

Upgrade of the automatic analysis system in the TJ-II TS diagnostic

5.1 Bibliographic Description

Title

Upgrade of the automatic analysis system in the TJ-II Thomson Scattering diagnostic: New image recognition classifier and fault condition detection.

Citation

L. Makili, J. Vega, S. Dormido-Canto, I. Pastor, A. Pereira, G. Farias, A. Portas, D. Pérez-Risco, M.C. Rodríguez-Fernández, P. Busch (2010) Upgrade of the automatic analysis system in the TJ-II Thomson Scattering diagnostic: New image recognition classifier and fault condition detection, *Fusion Engineering and Design*, ISSN 0920-3796, Volume 85, Issues 34, Pages 415-418.

Abstract

An automatic image classification system based on support vector machines (SVM) has been in operation for years in the TJ-II Thomson Scattering diagnostic. It recognizes five different types of images: CCD camera background, measurement of stray light without plasma or in a collapsed discharge, image during ECH phase, image during NBI phase and image after reaching the cut off density during ECH heating. Each

kind of image implies the execution of different application software. Due to the fact that the recognition system is based on a learning system and major modifications have been carried out in both the diagnostic (optics) and TJ-II plasmas (injected power), the classifier model is no longer valid. A new SVM model has been developed with the current conditions. Also, specific error conditions in the data acquisition process can automatically be detected and managed now. The recovering process has been automated, thereby avoiding the loss of data in ensuing discharges.

References

J. Vega et al. (2005); V. Vapnik (1998); A. Cherkasski, F. Mullier (2007); I. Daubechies (1992); S. Mallat (2001); J. Weston, C. Watkins (1999).

Impact Factor

Fusion Engineering and Design has an impact factor of 1.49 according to Thomson Reuters Journal Citation Reports (2011).

Fusion Engineering and Design 85 (2010) 415–418

Contents lists available at ScienceDirect

Fusion Engineering and Design

journal homepage: www.elsevier.com/locate/fusengdes

Upgrade of the automatic analysis system in the TJ-II Thomson Scattering diagnostic: New image recognition classifier and fault condition detection

L. Makili[a], J. Vega[b], S. Dormido-Canto[a,*], I. Pastor[b], A. Pereira[b], G. Farias[a], A. Portas[b], D. Pérez-Risco[b], M.C. Rodríguez-Fernández[b], P. Busch[c]

[a] Dpto. Informática y Automática - UNED, Madrid, Spain
[b] Asociación EURATOM/CIEMAT para Fusión, Madrid, Spain
[c] FOM Institut voor PlasmaFysica Rijnhuizen, Nieuwegein, The Netherlands

ARTICLE INFO

Article history:
Available online 6 December 2009

Keywords:
Support vector machines
Multi-class
Wavelet
Classifier

ABSTRACT

An automatic image classification system based on support vector machines (SVM) has been in operation for years in the TJ-II Thomson Scattering diagnostic. It recognizes five different types of images: CCD camera background, measurement of stray light without plasma or in a collapsed discharge, image during ECH phase, image during NBI phase and image after reaching the cut off density during ECH heating. Each kind of image implies the execution of different application software. Due to the fact that the recognition system is based on a learning system and major modifications have been carried out in both the diagnostic (optics) and TJ-II plasmas (injected power), the classifier model is no longer valid. A new SVM model has been developed with the current conditions. Also, specific error conditions in the data acquisition process can automatically be detected and managed now. The recovering process has been automated, thereby avoiding the loss of data in ensuing discharges.

1. Introduction

An automatic image classification system has been in operation for years in the TJ-II Thomson Scattering Diagnostic (TSD) [1]. The method to recognize the several classes was based on a learning system, in particular support vector machines (SVM) [2].

Since the first implementation of the classifier some important improvements has been accomplished both in the diagnostic and TJ II plasmas. A new notch filter is in operation, having a larger stray-light rejection at the ruby wavelength than the previous filter, for example. On the other hand, its location in the optical system has been modified. As a consequence, the stray-light pattern in the CCD image is located in a different position. In addition to these transformations, the power of neutral beams injected in the TJ-II plasma has been increased about a factor of 2.

Consequently, the creation of a new model under the present conditions has been necessary.

In this paper we present the fundamental options about the classifier's design. A couple of strategies emerge from the design: the use of the wavelet transform in the pre-processing stage and support vector machines in the classification stage. In Section 2 we describe the fundamentals of wavelet transforms and SVM and their uses in pre-processing, training and classifying the TSD images. Section 3 presents the programmed tool and the results regarding the TSD images. Finally, in Section 4, some conclusions are shown.

2. Classification process

The aim of patterns classification is to find a rule, based on external observations or training elements, that allows assigning each pattern to anyone of several possible classes. There are two big stages to implement in a classification process: features extraction and pattern classification [3]. The first one consists of performing some pre-processing on the patterns trying to extract specific differentiating features. The second stage groups the patterns into a set of classes.

In the TJ-II Thomson Scattering case, the patterns to classify are images. Each of them belongs to one of the five following classes: CCD camera background (BKGND), measurement of stray light without plasma or in a collapsed discharge (STRAY), image during ECH phase (ECH), image during NBI phase (NBI) and image after reaching the cut off density during ECH heating (COFF).

Each image has (385×576) pixels, *i.e.* 221,760 possible attributes. Firstly, we took into consideration a general classification scheme based on a wavelet transform (WT) to reduce the characteristics set, and secondly SVM has been used for pattern recognition.

* Corresponding author. Tel.: +34 91 3987194; fax: +34 91 3987690.
E-mail address: sebas@dia.uned.es (S. Dormido-Canto).

0920-3796/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.fusengdes.2009.10.004

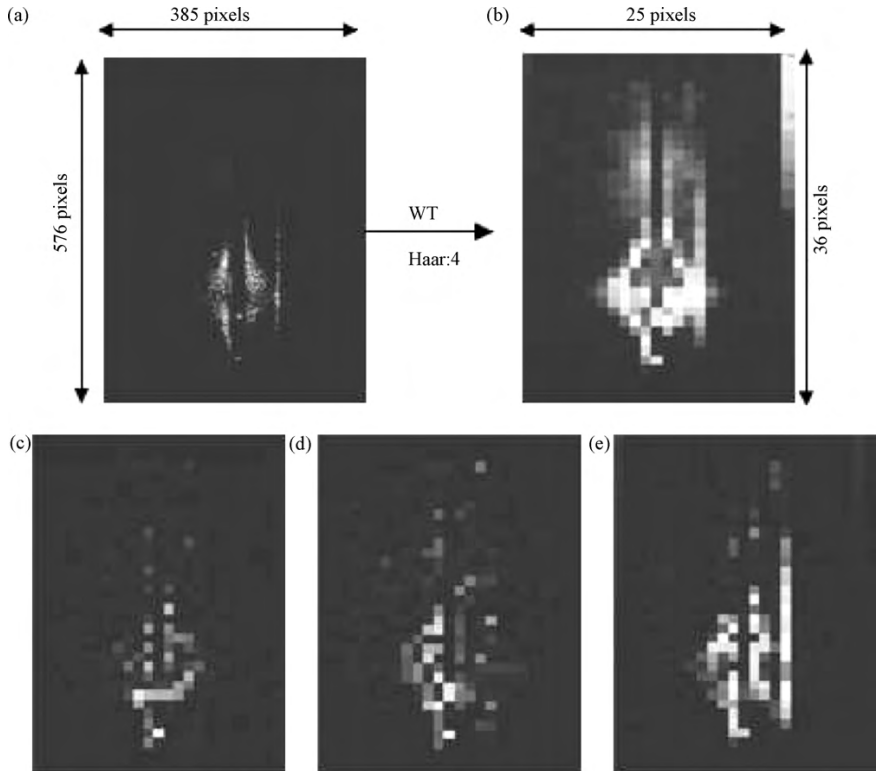


Fig. 1. (a) Original signal, (b) aproximation, (c) horizontal detail, (d) diagonal detail and (e) vertical detail.

2.1. Wavelet transform

Analysis of bi-dimensional signals can be greatly improved by using Wavelet based methods [4,5]. Due to the fact that the WT decomposition is multi-scale, images can be characterized for a set of approximation coefficients and three sets of detailed coefficients (horizontal, vertical and diagonals). The approximation coefficients represent coarse image information (they contain the most part of the image's energy), whereas the details are close to zero, but the information they represent can be relevant in a particular context.

We have found that the best coefficient to characterize the TJ-II Thomson images is the vertical detail, when selected the Haar Wavelet at level 4. With these setting, the attributes are reduced from 221,760 to 900 (0.4% of the initial attributes). Fig. 1 illustrates the WT applied to the image of a signal belonging to COFF class using the Haar Wavelet at level 4.

2.2. Support vector machines

SVM is a very effective method based on kernels for general purpose pattern recognition. In a few words, given a set of input vectors which belong to two different classes, the SVM maps the inputs into a high-dimensional feature space through some non-linear mapping (kernel functions), where an optimal separating hyper-plane is constructed in order to minimize the risk of misclassification. The hyper-plane is determined by a subset of points of the two classes, named support vectors.

The decision function that defines this hyper-plane is the following:

$$f(x) = \sum_{i=1}^{n} \alpha_i y_i K(x_i, x) + b \tag{1}$$

The parameters α_i, $i = 1,\ldots,n$ are the solution of the following quadratic optimization problem (QP-problem). Maximize the function:

$$L(\alpha) = -\frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, y_j) + \sum_{i=1}^{n} \alpha_i \tag{2}$$

subject to these constraints

$$\sum_{i=1}^{n} \alpha_i y_i = 0, \quad 0 \le \alpha_i \le C/n \quad i = 1, \ldots, n$$

where (x_i, y_i), $i = 1,\ldots,n$ are training data, $x_i \in R^n$, $y_i \in \{-1,1\}$, K is a kernel function and C is a regularization parameter. There are different kernel types: linear, polynomial or radial basis function (RBF). Table 1 shows the kernel functions used in this work.

SVM has been applied with a great success in binary classification problems. Many authors proposed methods to extend their application to classification problems with multiple classes (multi-class classifiers). In general, there are three fundamentals approaches: 1) *one versus the rest*, each classifier is trained to separate one class from the rest, 2) *one versus one*, there is a classifier for each pair of classes, and 3) *Weston and Watkins algorithm* [6] that enables to solve a multi-class problem in a single optimization.

Table 1
Kernel functions.

Linear	$K(x,x') = \langle x,x' \rangle$
Polynomial of degree d	$K(x,x') = (\langle x,x' \rangle + 1)^d$
RBF	$K(x,x') = \exp\{-\|\|x,x'\|\|^2/2\sigma^2\}$

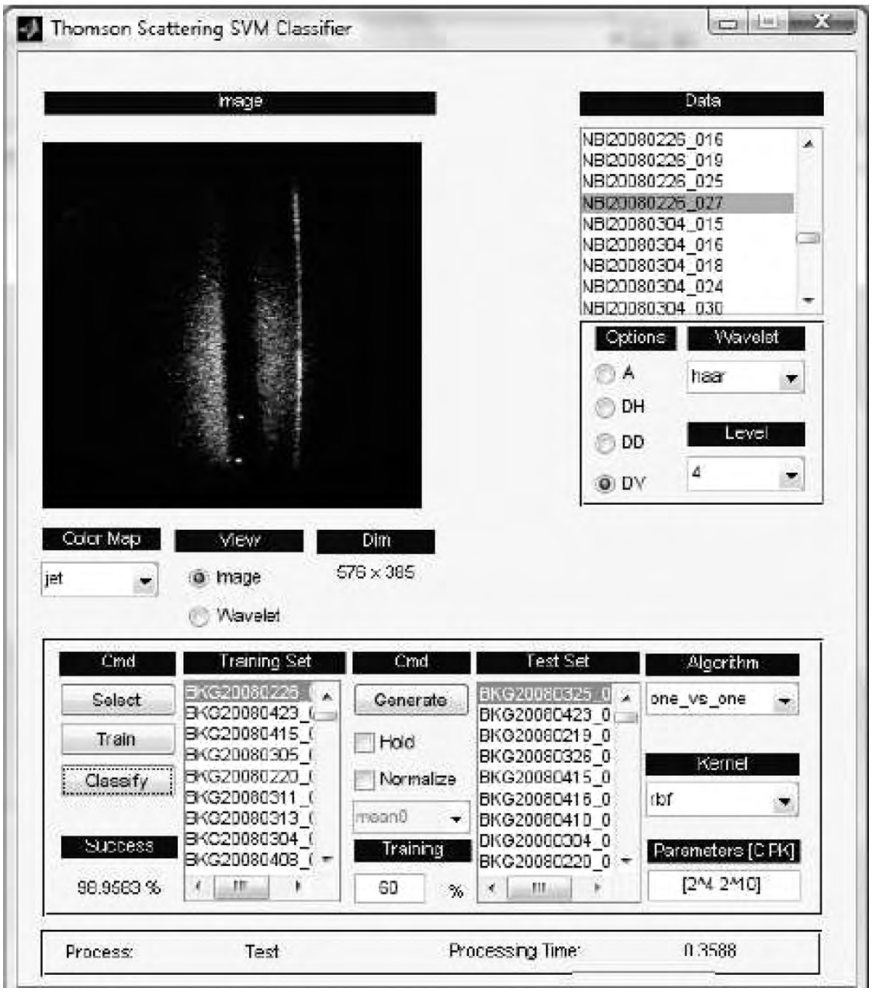


Fig. 2. Thomson scattering multi-classifier.

In this work has been made three multi-class classifiers with the TSD images to compare the efficiency of the different techniques.

3. Experimental results

To implement the previous ideas an application with a graphical user interface has been designed in MATLAB (Fig. 2). This application allows to manipulate a set of labeled images, whose main function is to evaluate the performance of the different classifiers.

These classifiers can be obtained easily by modifying some parameters as: the type of algorithm, the kernel function, the kernel parameters and the training and test set sizes.

Also, it is possible to specify the different parameters associated with the WT in the feature extraction stage: a decomposition level (reduction factor of dimensionality), the wavelet algorithm and the set of obtained wavelet coefficients, that is, the approximation or the detail (horizontal, vertical or diagonal).

In the experiments a total of 242 images were used. These images belong to one of the following classes: BKGND (50), COFF (42), ECH (50), NBI (50) and STRAY (50).

To validate the efficiency of the classifier, the images were divided randomly in training and testing. In this case, the training set was composed by 60% of the all images. The success rates depend strongly on kernel parameters. Consequently, previously to training and classification process, a search for the best values of the regularization parameter (C) and kernel parameters (σ o d, according to the case) was made.

Table 2 shows the success rates with different multi-class algorithms for various kernel functions and the best values in the parameters. KP represents the kernel parameter for the corresponding kernel function.

Table 2
Success rates SVM multi-class classifiers.

Kernel	[C, KP]	Success %
(a) One versus the rest algorithm		
Linear	$[2^{-2},-]$	98.68
Poly	$[2^{-2},2^{0}]$	98.68
RBF	$[2^{6},2^{11}]$	98.68
(b) One versus one algorithm		
Linear	$[2^{0},-]$	98.68
Poly	$[2^{0},2^{-3}]$	96.05
RBF	$[2^{5},2^{11}]$	98.68
(c) Weston and Watkins algorithm		
Linear	$[2^{1024},-]$	34.21
Poly	$[2^{1024},2^{-4}]$	60.52
RBF	$[2^{1024},2^{10}]$	98.68

It is important to note the poor results with Weston and Watkins algorithm. This method attempts to directly solve a multi-class problem modifying the binary class objective function and adding a constraint to it for every class. So, it is necessary to use a kernel more complex (such as rbf) to obtain good results.

4. Conclusions

In this paper, we present a new approach for the classification of TSD images. The method proposed here contains two processing stages, pre-processing of the original images by wavelet transform and multi-class classification by support vector machines. In the first stage, wavelet transformations are applied to signals to reduce the number of dimensions of the feature vectors. After that, a SVM-based multi-class classifier is constructed using the preprocessed signals as input space.

From observation of several experiments, our WT + SVM method is very viable and efficient time (approximately 200 times faster WT + SVM than only using SVM), and the results seem promising. However, we have further work to do. We have to finish the development of a Matlab toolbox for WT + SVM processing and to include new relevant features in the SVM inputs to improve the technique, even developing new kernel functions.

Acknowledgements

This work was partially funded by the Spanish Ministry of Science and Innovation under the Project No. ENE2008-02894/FTN.

This work, supported by the European Communities under the contract of Association between EURATOM/CIEMAT, was carried out within the framework of the European Fusion Development Agreement. The views and opinions expressed herein do not necessarily reflect those of the European Commission.

References

[1] J. Vega, I. Pastor, et al., Application of intelligent classification techniques to the TJ-II Thomson Scattering diagnostic, 32th EPS Plasma Physics Conference, 27 junio – 1 julio 2005, Tarragona (España) (http://eps2005.ciemat.es).
[2] V. Vapnik, Statistical Learning Theory, John Wiley & Sons, INC., 1998.
[3] A. Cherkasski, F. Mullier, Learning from Data, 2nd ed., Wiley-IEEE Press, 2007.
[4] I. Daubechies, Ten Lectures on Wavelets, SIAM, 1992.
[5] S. Mallat, A Wavelet Tour of Signal Processing, 2nd ed., Academia Press, 2001.
[6] J. Weston, C. Watkins, Support Vector Machines for multi-class pattern recognition, in: Proceeding of the Seventh European Symposium on Artificial Neural Networks, 1999.

Article 6

Laboratorios virtuales de procesamiento de señales

6.1 Bibliographic Description

Title

Laboratorios virtuales de procesamiento de señales.

Citation

M. Santos, G. Farias (2010) Laboratorios Virtuales de Procesamiento de Señales, *Revista Iberoamericana de Automática e Informática Industrial (RIAI)*, ISSN 1697-7912, Volume 7, Number 1, Pages 91-100.

Abstract (In Spanish)

En este trabajo se exponen diversas contribuciones que se han ido desarrollando en el contexto de la enseanza universitaria para carreras tcnicas que facilitan la comprensin de los principales conceptos del procesamiento digital de seales y reconocimiento de patrones. Las prcticas se realizan bien en laboratorios virtuales mediante herramientas de simulacin, bien de forma distribuida a travs de internet y, en algunos casos, se han implementado en sistemas reales mediante DSP. La comprobacin visual y en algunos casos auditiva del procesamiento aplicado permite la autoevaluacin y motivacin del alumno. Se presentan tambin algunas experiencias de su implantacin en la Universidad

Complutense de Madrid.

References

F.A. Candelas et al.(2004); R. Chassaing (1999); R. Chassaing (2005); S. Dormido (2004); T. S. Elali (2003); F. Esquembre (2008); European Union (1999); G. Farias et al. (2008); G. Farias, M. Santos, S. Dormido-Canto (2005); J. L. Guzmán et al.(2005); B. S. Heck (1999); J. Hilera J, V. Martínez (1995); E.C. Ifeachor, B.W. Jervis (1993); U.A. Karrenberg (2002); Mathworks Inc. (2002); A.V. Oppenheim, A.S. Willsky, S. Hamid (1997); A. Oppenheim, R. Schafer (1999); M. Rahkila, M. Karjalainen (1997); M. Rahkila, M. Karjalainen (1998); J. Sánchez et al. (2002); M. Santos, J. Klaus González (2007); M. Santos (2007); S. Smith (2002); S. D. Stearns (2002); V. Vapnik (2000).

Impact Factor

Revista Iberoamericana de Automática e Informática Industrial (RIAI) has an impact factor of 0.231 according to Thomson Reuters Journal Citation Reports (2011).

Revista Iberoamericana de Automática e Informática Industrial
CEA comité español de automática
RIAI
ISSN: 1697-7912. Vol. 7, Núm. 1, Enero 2010, pp. 91-100
http://www.revista-riai.org

Laboratorios virtuales de procesamiento de señales

M. Santos Peñas*, G. Farias Castro**

** Dpto. de Arquitectura de Computadores y Automática,*
Facultad de Informática, Universidad Complutense de Madrid,
C/ Profesor García Santesmases s/n, 28040-Madrid, España (email: msantos@dacya.ucm.es)
*** Dpto. de Informática y Automática. Escuela Superior de Ingeniería Informática. UNED.*
C/ Juan del Rosal s/n. 28040-Madrid, España (e-mail: gfarias@bec.uned.es)

Resumen: En este trabajo se exponen diversas contribuciones que se han ido desarrollando en el contexto de la enseñanza universitaria para carreras técnicas que facilitan la comprensión de los principales conceptos del procesamiento digital de señales y reconocimiento de patrones. Las prácticas se realizan bien en laboratorios virtuales mediante herramientas de simulación, bien de forma distribuida a través de internet y, en algunos casos, se han implementado en sistemas reales mediante DSP. La comprobación visual y en algunos casos auditiva del procesamiento aplicado permite la auto-evaluación y motivación del alumno. Se presentan también algunas experiencias de su implantación en la Universidad Complutense de Madrid.

Palabras Clave: Laboratorios Virtuales, Laboratorios Virtuales Distribuidos, Educación, Procesamiento de Señales, Automática.

1. INTRODUCCIÓN

El siglo XXI se presenta como el siglo de las comunicaciones, como lo han sido las últimas décadas del XX. Las comunicaciones son el sustento de la información, que sigue siendo la herramienta más poderosa a nivel social, cultural, político, económico, técnico, científico, etc. Pero la información se encuentra hoy día en unos formatos y se transmite con unas técnicas que son producto de la más reciente tecnología. De hecho, somos testigos del avance vertiginoso de las comunicaciones debido a la incorporación de nuevas estrategias y dispositivos que hacen más eficiente su transmisión.

Al mismo tiempo, es fácil detectar en los alumnos universitarios de carreras científicas y técnicas un interés creciente por saber trabajar con información que puede provenir de fuentes muy diversas. Los estudiantes de áreas experimentales y técnicas deben ser capaces de procesar las señales físicas, independientemente del ámbito en el que se hayan generado. Así, ya sea una imagen de satélite o tomada con una cámara digital, ya sea un registro que recoge una serie de operaciones bancarias, bien un encefalograma que representa un tumor cerebral, o una señal de voz, etc., todas estas señales pueden aportar una información muy valiosa, en algunos casos crucial, que hay que saber tratar.

Dentro del campo del tratamiento de las señales, el simular y visualizar los pasos que experimenta una señal a lo largo de su transmisión, desde su emisión hasta que llega a su destino final y cumple su misión de transportar una información, es fundamental para los alumnos que se van a mover en un mundo donde priman las comunicaciones. Los conceptos teóricos necesitan confirmarse con experiencias realizadas en laboratorios, acercándose a la realidad. Es más, los alumnos que formamos en titulaciones científicas y técnicas deben recibir una educación eminentemente práctica y aplicada, experimental y científica, que ayude a comprender –en el sentido más profundo de la palabra- los conceptos abstractos aprendidos en las clases teóricas, que además en esta materia tienen una alta carga matemática (Ifeachor and Jervis, 1993; Oppenheim and Schafer, 1999; Smith, 2002)

Esta experimentación no siempre se da en el aula debido a la masificación, falta de recursos, escasez de tiempo, coste de la instrumentación, etc. Para paliar esta carencia se han desarrollado los últimos años, al amparo de Proyectos de Innovación Educativa y Mejora de la Calidad Docente de la Universidad Complutense de Madrid (UCM), una serie de laboratorios virtuales en los que los alumnos, mediante herramientas de simulación, pueden realizar experiencias prácticas de procesamiento de señales con un coste muy bajo y gran eficiencia pedagógica. Algunas de estas propuestas se han implementado también de forma distribuida a través de internet, o sobre sistemas reales.

Con este tipo de experiencias el alumno puede profundizar en los conceptos que le resulten más necesarios o que le interesen particularmente, al ritmo que marque su base teórica o su preparación previa. En general es interesante introducir innovaciones en las funciones y métodos docentes previstos para esta materia ya que facilitan la participación de los alumnos y por tanto el aprovechamiento de las enseñanzas. Otro aspecto positivo es que se fomenta la aplicación de herramientas computacionales que facilitan el trabajo del alumno y están enfocadas al ejercicio profesional. En general, como ayudan a relacionar sus conocimientos con el mundo real al tratar con ejemplos tomados de diversos ámbitos, resulta muy atractivo para los estudiantes y motiva su interés. De hecho estas iniciativas se están aplicando recientemente en el ámbito del control (Heck, 1999; Sánchez et al., 2002; Dormido, 2004; Candelas et al., 2004; Guzmán et al., 2005)

Publicado electrónicamente: 08/01/2010
DOI: 10.4995/RIAI.2010.01.09

Respecto a los docentes, estos laboratorios les dotan de herramientas que puede utilizar con gran flexibilidad, tanto para apoyarse en ellas a la hora de explicar la teoría, mediante demostraciones y ejemplos, como a la hora de proponer prácticas y ejercicios, por lo que permiten una mejor asimilación por parte de los alumnos de los contenidos que se enseñan y un mejor seguimiento de los mismos por parte del profesor. Además el trabajo en laboratorios virtuales fomenta la comunicación entre los alumnos entre sí, porque muchas veces se realiza en pequeños grupos, y con el profesor.

Este trabajo tiene como objetivo mostrar el desarrollo de algunas prácticas de tratamiento de señales que incorporan elementos de la realidad. Para ello se utilizan herramientas de simulación como Matlab (Mathwork, 2002), Easy Java Simulation (EJS) (Esquembre, 2008) para la implementación a través de internet, y tarjetas de procesamiento digital de señales (DSP) como plataformas de soporte y desarrollo. Se han venido utilizando en los estudios de Ingeniería Electrónica, CC. Físicas e Ingeniería Informática de la Universidad Complutense de Madrid, así como en la Red Docente de posgrado con Iberoamérica AIASYB: Aplicaciones de la Inteligencia Artificial en los Sensores y Biosensores, y en la Red ALFA BioSenInt. Los alumnos se han mostrado muy receptivos a este tipo de iniciativas. La utilización de este tipo de estrategias docentes persigue la implantación de metodologías más activas y participativas –más acordes con el espíritu del Espacio Europeo de Educación Superior (EEES)- orientadas a facilitar y mejorar el proceso del aprendizaje de los estudiantes así como su consolidación (European Union, 1999).

En este artículo se presenta en primer lugar una serie de herramientas de procesamiento de señales desarrolladas para trabajar en laboratorios virtuales con acceso local (sección 2) y a través de internet (Sección 3), y un laboratorio tradicional (Sección 4). En la Sección 5 se expone el ámbito y metodología de aplicación de estas experiencias prácticas. En la sección 6 se analiza el impacto de su aplicación sobre estudiantes de la UCM, así como los beneficios derivados. El artículo termina con las conclusiones.

2. LABORATORIOS VIRTUALES DE PROCESAMIENTO DE SEÑALES

Actualmente existe una gran inquietud por la aplicación de las nuevas tecnologías a la enseñanza (Heck, 1999; Dormido, 2004). De hecho, el personal docente cuenta con algunos desarrollos para la realización computacional de ejercicios de procesamiento de señales. Algunas de estas herramientas son cerradas, como por ejemplo, DASILab (Karrenberg, 2002), que requiere licencia ya que es un entorno profesional desarrollado al amparo de la Compañía National Instruments. Otras propuestas (Stearns, 2002; Elali, 2003) presentan ejemplos usando el entorno de programación de Matlab, pero no son interactivas ni han desarrollado una plataforma de simulación; su objetivo es ilustrar algunos conceptos del procesamiento de señales mediante ejercicios. Chassaing (1999), presenta ejemplos resueltos en ensamblador y C para experimentos en tiempo real sin apenas capacidad de visualización e interacción. En general estas herramientas no cubren todos los aspectos remarcados en este trabajo y, fundamentalmente, no están estructuradas como una unidad entre las asignaturas que hacen relación al tratamiento de las señales en las carreras para las cuales se han desarrollado. Por ello se han ido generando una serie de escenarios de simulación que permiten, tanto a los alumnos como al profesor, analizar de forma gráfica e interactiva señales de diversos ámbitos en distintos dominios. Estas herramientas se están utilizando en asignaturas de varias carreras y a distintos niveles, según la preparación de los alumnos y los objetivos formativos de cada curso.

En Farias (2008) se presenta una taxonomía de los laboratorios, atendiendo al acceso y al tipo de recurso. El acceso puede ser local o remoto (a través de internet), y el recurso puede ser real (físico) o simulado. Independientemente del tipo de recurso, si el acceso es remoto, se dicen que son laboratorios (virtual y remoto) basados en Web. Se han seguido esas pautas para clasificar las distintas herramientas docentes de procesamiento de señales que se presentan:

- PDS y SiSCoD: simulación local (acceso local, recurso virtual)
- GUI-TAIS: simulación local (acceso local, recurso virtual)
- GUI-TAIS simulación distribuida a través de internet (acceso remoto, recurso virtual)
- Lab DPS: laboratorio tradicional (acceso local, recurso real)

Para programarlas se ha optado por el entorno Windows, usando el paquete de Software Matlab, que proporciona un potente interfaz gráfico de usuario para aplicaciones y unas elevadas prestaciones matemáticas que facilitan el tratamiento de diferentes tipos de señales (Mathworks, 2002). Es una herramienta ampliamente consolidada tanto en el ámbito investigador como en el educacional, a la que tienen acceso con facilidad también los alumnos. Como resultado de utilizar Matlab, las herramientas desarrolladas participan de sus características: rápido aprendizaje, herramienta de bajo coste (existen licencias especiales para educación), pueden utilizarse en distintas plataformas, y bajo distintos sistemas operativos, etc. Por otra parte, se ha optado por EJS (Esquembre, 2008) para realizar la implementación distribuida de una de las prácticas debido a las facilidades que esta herramienta proporciona para este fin.

Por último, último, hacer hincapié en que todas las herramientas se han diseñado de forma modular, que se presta a futuras mejoras y ampliaciones. Además la transferibilidad de estos laboratorios se ve facilitada porque este conjunto de experiencias se proponen en un entorno estándar, por lo que pueden funcionar en cualquier máquina con pocas prestaciones, y no tienen especiales requerimientos.

A continuación se presentan las herramientas que componen los laboratorios virtuales desarrollados y utilizados en las prácticas.

2.1 PDS: Procesador Digital de Señales

Esta herramienta ha sido diseñada con el propósito de visualizar y analizar el comportamiento de señales continuas y discretas tanto en el dominio temporal como en frecuencia. Para llegar a entender su comportamiento se han implementado operaciones básicas tales como composición de señales, correlación cruzada, transformada de Fourier, inversa de la misma, filtrados de señales, etc. La aportación más interesante de esta herramienta es que permite observar de forma gráfica los cambios que experimentan las señales al realizar esas operaciones según van ocurriendo, y sacar conclusiones (Santos y González, 2007).

El modulo principal del programa es un código programado íntegramente en Matlab que se ejecuta desde el fichero: **PDS.m.** Este programa proporciona un interfaz gráfico, desde donde el usuario puede elegir distintas opciones o cambiar parámetros de diseño mediante menús (Figura 1).

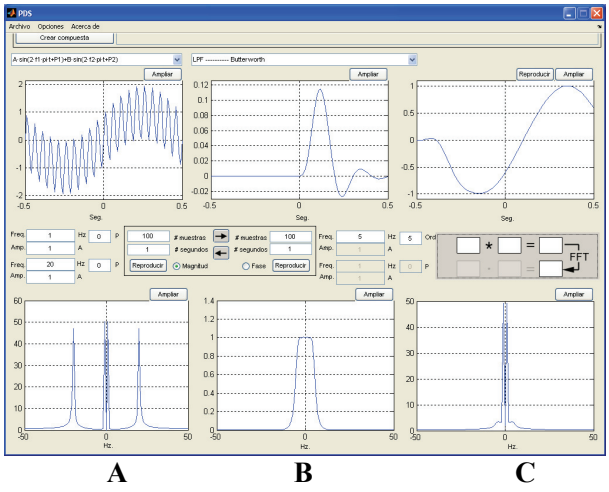


Figura 1. Ventana principal de PDS.

La ventana principal presenta un interfaz de usuario que se puede dividir inicialmente en tres áreas generales, que de izquierda a derecha podemos nombrar como A, B y C respectivamente. Cada una de las áreas consta de dos ventanas. La superior muestra la amplitud de la señal en el dominio temporal (eje X en segundos) mientras que la ventana inferior representa la correspondiente señal en el dominio de la frecuencia (eje X en hertzios). Cualquier señal representada en la ventana superior tiene su equivalente en frecuencia en la ventana directamente inferior y viceversa.

Básicamente, en el área A se muestran las señales a analizar y las posibilidades de composición para obtener señales complejas; en el área B se pueden seleccionar distintos filtros sintonizando sus parámetros para aplicarlos sobre las señales del área A, y el área C presenta el resultado de filtrar A con B.

Entre las figuras de la ventana superior de las áreas A y B se encuentra un menú que permite seleccionar el número de muestras y la duración en segundos de la señal. Además permite pasar del gráfico de magnitud al de fase (Figura 1).

Para las áreas A y B (señales y filtros, respectivamente), la respuesta en frecuencia correspondiente se calcula automáticamente mediante un algoritmo de la FFT (Fast Fourier Transform). En el área C el proceso puede ser bidireccional: este área C puede funcionar en modo temporal o frecuencial. Dicho modo se puede cambiar en cualquier momento de la ejecución del programa y los resultados deberían ser idénticos. Para cambiar entre los dos modos simplemente hay que pulsar sobre el cuadro que se encuentra entre las dos ventanas del área C (Figura 2).

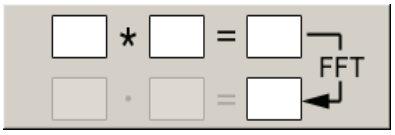


Figura 2. Cambio de modo temporal-frecuencial.

Además la herramienta tiene unas serie de ayudas adicionales (reproducir una señal de voz, ampliar, atenuar, etc.).

Esta herramienta está siendo utilizada en la asignatura de Transmisión de Datos (Complemento de Formación de Ingeniería Electrónica y optativa de CC. Físicas), y en Procesamiento de Señales (obligatoria en Ingeniería Electrónica).

2.2 SiSCoD: Sistema de Comunicación Digital

La visión global de los sistemas de comunicación digitales, así como la comprensión de las transformaciones que sufren las señales a su paso por cada uno de ellos, constituye un aspecto muy importante en la formación de cualquier persona que trabaje dentro del ámbito del procesamiento digital de las señales.

La herramienta SiSCoD, que se ejecuta mediante el comando **SiSCoD** en Matlab, muestra de forma gráfica y cercana a la realidad los distintos procesos que tienen lugar en un sistema de comunicación digital. Asimismo, permite ver de forma simultánea las representaciones temporales, espectrales y fasoriales más importantes que pueden aparecer en un sistema de estas características dependiendo del punto del esquema en el que se encuentre, a la vez que permite observar sus cambios cuando se modifica algún parámetro de diseño (Figura 3).

Figura 3. Tipos de gráficos disponibles de las señales.

A grandes rasgos, el diagrama de bloques del sistema de comunicación implementado se compone de una fuente de información, un modulador, un canal y un demodulador. Cada uno de estos elementos deber ser modelado seleccionando las características de diseño de cada uno de ellos (Figura 4).

Para el modulador se han elegido dos técnicas de modulación: una de tipo IQ (*Inphase-Quadrature*) para las modulaciones de amplitud, fase e híbridas de amplitud y fase; y otra de tipo CP (*Continuous-Phase*) para las modulaciones de frecuencia y de fase continua. El usuario dispone de una opción para generar cualquier modulación híbrida de amplitud y fase (MPAK) hasta un orden 16.

El esquema de demodulación simulado depende del tipo de modulación elegida. Para las modulaciones de amplitud, de fase e híbridas de amplitud y fase, el demodulador se corresponde con el modelo de detector coherente con criterio de decisión de Máxima Semejanza; mientras que para demodular las de frecuencia y fase continua se utiliza un discriminador de frecuencia clásico.

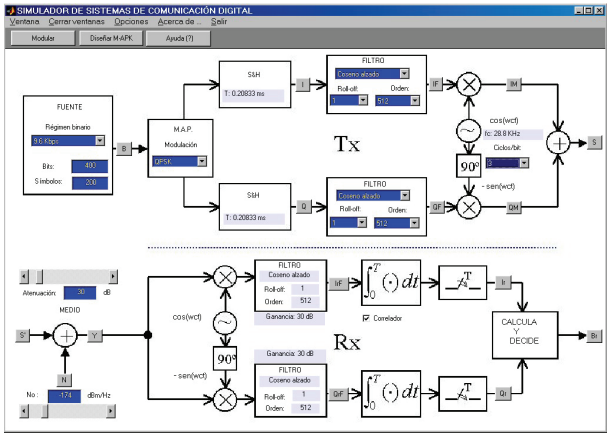


Figura 4. Pantalla principal de SiSCoD.

El simulador da al usuario la posibilidad de escoger entre dos tipos de filtro: coseno alzado y gaussiano. Si escoge un filtro en una de las ramas del transmisor, tanto la otra rama como el receptor usarán ese tipo de filtro. Lo mismo ocurre con los parámetros de configuración del mismo, si se escogen en uno automáticamente se igualarán los del resto. Otra de las opciones disponibles es la de diseñar una constelación por parte del usuario, para lo que permite interaccionar con el ratón y colocar los puntos sobre el gráfico (Figura 5).

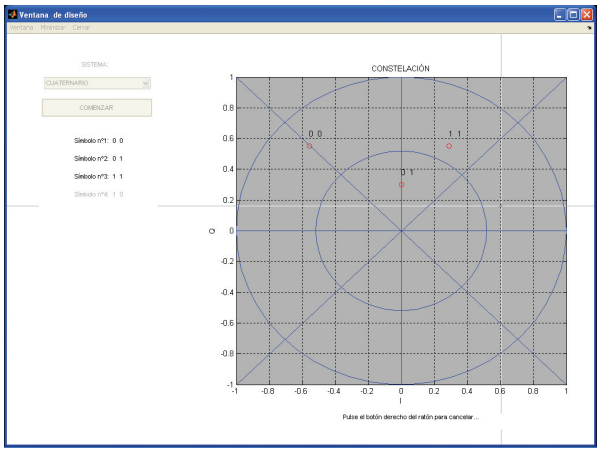


Figura 5. Diseño de una constelación en SiSCoD.

La herramienta tiene además un sistema de ayuda en cada punto del diagrama de bloques de la transmisión que facilita información sobre la señal que se está tratando, los parámetros que se pueden variar, etc. También tiene ciertas facilidades para su uso, como cerrar varias ventanas de forma simultánea, grabar datos y figuras, etc.

Este laboratorio virtual está disponible en la asignatura de Física de las Radiocomunicaciones y en Transmisión de Datos (ambas Complemento de Formación de Ingeniería Electrónica y optativas de CC. Físicas).

3. GUI-TAIS: TRATAMIENTO AVANZADO E INTELIGENTE DE SEÑALES, LABORATORIO VIRTUAL DISTRIBUIDO

Para la consecución del objetivo global de un mayor acercamiento desde el punto de vista práctico al tratamiento de señales por parte de los alumnos, se ha desarrollado la herramienta de simulación GUI_TAIS (Tratamiento Avanzado e Inteligente de Señales) en Matlab, que permite aplicar técnicas de procesamiento inteligente a señales de distintas fuentes, para su clasificación, obtención de características, identificación de patrones, etc. (Farias et al., 2005). Posteriormente se ha implementado una versión de acceso a través de internet de este laboratorio, lo que permite al estudiante realizar la práctica aún cuando no posea Matlab en su ordenador (Farias et al., 2008). En el desarrollo de este laboratorio basado en web se utilizó el software Easy Java Simulations (EJS) que permite crear simulaciones interactivas en Java para construir la interfaz gráfica de usuario.

En este laboratorio se plantea el estudio y análisis de señales reales en las que la información es crucial, y propone una serie de pasos para su tratamiento (desde el procesamiento mediante diversas transformaciones hasta su clasificación final). Los alumnos aprenden a aplicar técnicas de procesamiento a señales de ámbitos muy diversos, tanto de una dimensión como imágenes, para posteriormente analizar los resultados y obtener conclusiones respecto a diagnóstico, clasificación, detección, prevención, etc. El hecho de ver que las señales se corresponden con situaciones reales (espectroscopia de tumores cerebrales, plasma, operaciones bancarias, imágenes de fusión, señales sísmicas, etc.) motiva su interés para aprender y profundizar en esas técnicas.

Esta herramienta es de fácil manejo, altamente ilustrativa, y capaz de mostrar de forma gráfica y cercana a la realidad los resultados de aplicar técnicas de pre-procesamiento, descomposición de señales, agrupamiento, recuperación de información, identificación de patrones, etc. Asimismo, permite ver las representaciones temporales de las señales a la vez que observar sus cambios cuando se modifica algún parámetro del sistema de tratamiento.

El esquema de experimentación que se debe seguir para aplicar esta herramienta es el que se muestra en la figura 6.

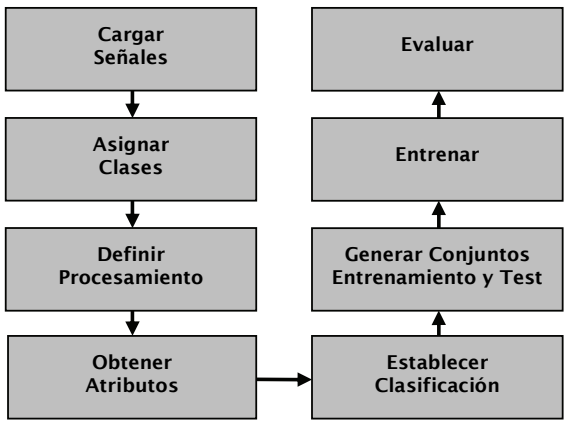


Figura 6. Fases de las prácticas con la herramienta GUI-TAIS.

El modulo principal del programa es un código programado íntegramente en Matlab que se ejecuta desde el fichero **clasificador.m.** El usuario (alumno) tiene inicialmente acceso a la ventana principal del programa, donde mediante botones, opciones desplegables, menús, etc., puede establecer todos los parámetros necesarios para aplicar una estrategia de procesamiento u otra, y visualizar los resultados correspondientes. En la Figura 7 se muestra el interfaz gráfico de la herramienta con las posibles opciones que contiene y da una idea de cómo es su entorno.

A grandes rasgos, las principales técnicas que se han implementado son de procesamiento estadístico (media, varianza, máximo y mínimo, etc.) y de compresión, en concreto, wavelets (tanto unidimensional como bidimensional). Esta última técnica puede aplicarse con distintos niveles de descomposición a seleccionar por el usuario. Seguidamente a ese grupo de señales procesadas se les pueden aplicar dos técnicas de agrupamiento (Figura 7): *Máquinas de Vectores Soporte* (SVM: Support Vector Machines) (Vapnik, 2000) y *Redes Neuronales*, (Hilera y Martinez, 1995) cuyos parámetros pueden también ser seleccionados según los objetivos de las prácticas (haciendo hincapié en el entrenamiento, o en la estructura, etc.).

Esta herramienta permite ver las representaciones temporales de las señales, tanto registros unidimensionales como imágenes, (Figura 8), y visualiza las semejanzas o características obtenidas de cada una de ellas o de grupos de señales. También presenta de forma gráfica los resultados de aplicar ciertas técnicas de procesamiento a las señales ya que se puede elegir en la ventana de visualización de la señal si se desea mostrar la señal o grupo de señales, o éstas ya procesadas (por ejemplo, la señal que resulta tras aplicar wavelets).

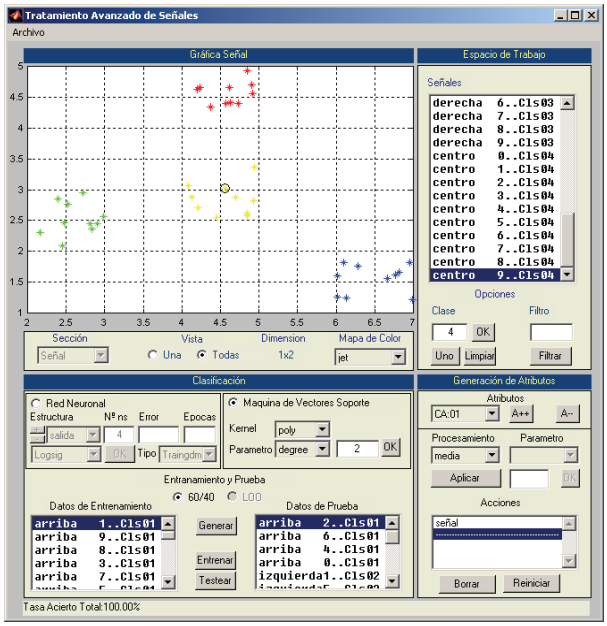


Figura 7. Interfaz gráfica de la herramienta GUI_TAIS.

Además muestra los resultados de los distintos métodos de agrupamiento, informando de los resultados de la clasificación (porcentaje de aciertos) y de cómo se ha llevado a cabo (estrategia de entrenamiento, datos utilizados para el mismo, etc.)

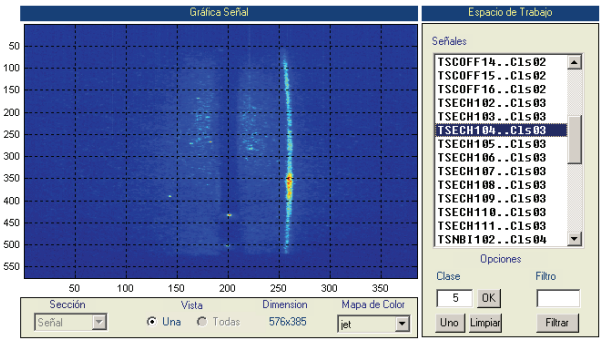


Figura 8. Carga y visualización de imágenes bidimensionales para su procesamiento.

Cuando se aplican redes neuronales se pueden configurar parámetros de su estructura como: número de neuronas de cada capa, el número de capas, el algoritmo de entrenamiento, número de épocas, etc. Con la técnica de SVM se puede elegir el tipo de kernel que se va a emplear así como los parámetros del mismo. En cuanto a los atributos, se pueden seleccionar características estadísticas de la señal o aplicar métodos de comprensión como las wavelets, que se pueden configurar eligiendo la wavelet madre y el nivel de descomposición (Figura 9).

En definitiva, es una herramienta final muy completa, gráfica e ilustrativa. Está disponible para prácticas en la asignatura de Procesamiento de Señales, de Ingeniería Electrónica, así como para apoyo a la docencia de clases del master y doctorado en lo que hace relación al tema de Aprendizaje Automático, que se viene impartiendo en la asignatura de Control Inteligente.

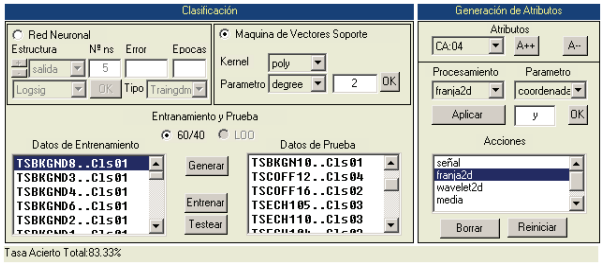


Figura 9. Parámetros de las estrategias de clasificación y de los atributos del pre-procesamiento.

3.1 Laboratorio virtual distribuido GUI-TAIS de procesamiento de señales

Como se presenta en Farias et al. (2008), se ha desarrollado una versión distribuida, de acceso a través de internet, del laboratorio virtual GUI-TAIS. Para esta aplicación computacional se ha utilizado la combinación EJS y un conjunto de funciones (archivos.m) que se ejecutarán de forma remota en Matlab. Con ello se tiene un laboratorio con la gran capacidad de visualización e interacción que resulta del uso de Java, y con la flexibilidad y potencia que le proporciona Matlab como motor de cálculo. El uso de estas herramientas remotas presenta una gran ventaja, y es que el estudiante no requiere tener Matlab

instalado en su ordenador para poder hacer uso del laboratorio virtual.

La funcionalidad de este laboratorio virtual distribuido es la misma que la del laboratorio virtual presentado anteriormente, la diferencia radica en su utilización por parte de los usuarios a través de una conexión a Internet.

En la Figura 10 se presenta el esquema de conexión entre Easy EJS y Matlab, donde se puede observar que el procesamiento y clasificación de las señales se realiza en el lado del servidor, mientras que en el lado del cliente se presentan los resultados a través de una interfaz gráfica creada con EJS.

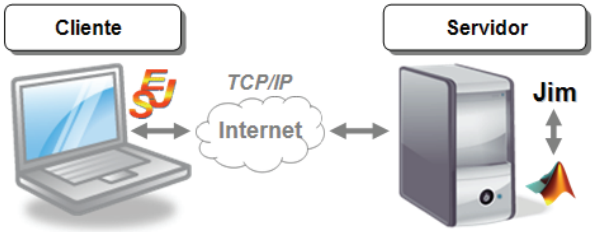


Figura 10. Conexión entre EJS y Matlab

La interfaz gráfica de usuario desarrollada es muy similar a la presentada en la sección anterior.

Este contexto de acceso a través de la web origina una serie de inconvenientes, como son los retardos en el acceso y la dependencia de la disponibilidad del servidor. Sin embargo presenta la ventaja, que ya se ha comentado, de lo que los alumnos no necesitan tener Matlab instalado en su ordenador personal para realizar las prácticas.

4. LABORATORIO DE PRÁCTICAS DE TRATAMIENTO DE SEÑALES: DE LA SIMULACIÓN AL SISTEMA REAL

Se han desarrollado distintas prácticas que constituyen el Laboratorio de Procesamiento de la Señal con la tarjeta DSP DSK6713 (Figura 11) (Spectrum Digital, 2003). Estas prácticas recogen conocimientos vistos en la teoría como: aplicación de transformadas, tanto en el dominio temporal como en frecuencia, implementación de filtros, desarrollo de algoritmos de procesamiento, etc. Con estos ejercicios se pretende hacer hincapié en la programación de los algoritmos, que en las herramientas presentadas anteriormente eran transparentes al usuario. Ahora el alumno no sólo configura los parámetros de la aplicación sino que implementa o modifica el código de las experiencias, lo que le permite un mayor dominio de las técnicas vistas en las clases de teoría.

Las prácticas se proponen tanto en simulación como en su implementación sobre la placa. Con la simulación se hace especial hincapié en la depuración del código para mejorar el rendimiento de las operaciones que se realicen sobre las señales en tiempo real. Este laboratorio de DSP tiene como base para la primera fase de simulación tanto Matlab, con sus facilidades de programación y visualización para el seguimiento de las transformaciones de la señal, como Visual-C para la programación de la placa. Es decir, los alumnos realizan las prácticas simulándolas primero con Matlab para gráficamente validar el código, para luego traducir el código a CSS (Code Composer Studio) y posteriormente ejecutarlo en tiempo real en la placa DSP, una vez depurado y validado.

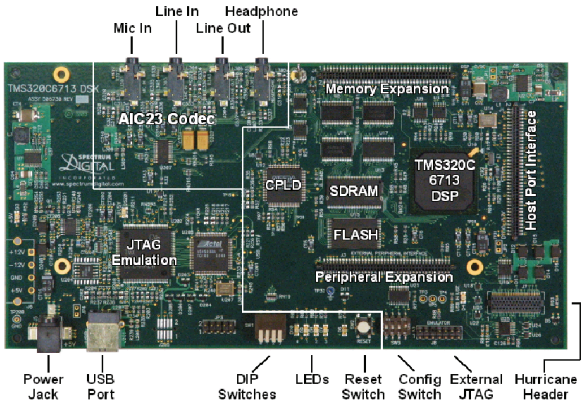


Figura 11. Vista de la DSP TMS320C6713 DSK.

El laboratorio tiene un gran alcance pedagógico: el alumno tiene que salir al paso de problemas que atañen distintos aspectos de la realización que suelen surgir al trabajar con sistemas físicos. Es muy interesante que los estudiantes tengan contacto con los sistemas reales con los que se trabaja en el mundo laboral (Rahkila and Karjalainen, 1997; Chassaing, 2005).

Para motivar el interés de los alumnos se han seleccionado señales audio para trabajar con ellas –que han resultado ser un buen marco para estas enseñanzas (Rahkila and Jarjalainen, 1998)-, de manera que puedan comprobar mediante su audición los efectos del procesamiento realizado en la DSP.

En concreto, se han diseñado las siguientes prácticas:

- *Introducción al entorno de programación Code Composer Studio* (CCS) (Figura 12), para la tarjeta de Procesamiento Digital de Señales TMS320C6713 de Texas Instrument, que es la que está disponible en los laboratorios de los alumnos.

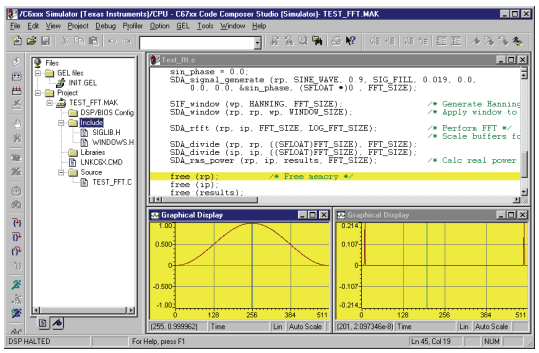


Figura 12. Entorno Code Composer Studio (CCS) de la tarjeta TMS320C6713.

- *Generación de una señal digital a partir de notas musicales*.
- *Diseño y aplicación de algoritmos para implementar efectos digitales de audio* (tremolo, eco, coro, reverberación, etc.).
- *Implementación de filtros digitales FIR ("Finite Impulse Response") e IIR ("Infinite Impulse Response")*.

Los alumnos trabajan con una señal desde su implementación inicial hasta la fase final en la que pueden comprobar los efectos de los algoritmos de tratamiento aplicados mediante la audición de la misma en tiempo real.

Por último, para estimular el interés de los alumnos se enuncian otras prácticas dedicadas al tratamiento de imágenes (detección de bordes, eliminación de ruido) donde los resultados muestran de forma gráfica las transformaciones de las señales. Se ha utilizado en las asignaturas de Transmisión de Datos y en Procesamiento de Señales.

5. ÁMBITO DE APLICACIÓN DE LOS LABORATORIOS Y METODOLOGÍA

Las herramientas desarrolladas para los laboratorios docentes han sido utilizadas, algunas desde el curso 2002-03, en las asignaturas que hacen relación al procesamiento de señales de las carreras de CC. Físicas y en Ingeniería Electrónica de la Universidad Complutense de Madrid. Es decir, en Transmisión de Datos y Física de la Radiocomunicación (Complementos de Formación de Ingeniería Electrónica y optativas de CC. Físicas), en Procesamiento de Señales (obligatoria en Ingeniería Electrónica), y en Control Inteligente (Master y Programa de Doctorado). También, de forma más tangencial, pueden dar soporte a la docencia en otras asignaturas del área de automática, como en la de Sistemas Lineales, Inteligencia Artificial aplicada al control, o Control Digital, que se imparten en la Facultad de Informática, o en concreto el laboratorio de señales con DSP en Arquitecturas Especializadas. El software está disponible para proyectos fin de carrera y fin de master, y ha sido utilizado en algunos de esos trabajos de investigación. Su alcance es, por lo tanto, interfacultativo ya que abarca varias asignaturas de distintos estudios.

En los nuevos planes de estudio, tanto en Físicas como en los diversos grados de Informática, se proponen asignaturas que hacen relación directa al tema del procesamiento de señales, por lo que pensamos que seguirán siendo de utilidad.

También se han difundido algunas de esas herramientas en el marco de la Red Docente con Iberoamérica AIASYB: Aplicaciones de la Inteligencia Artificial en los Sensores y Biosensores, dentro del módulo Procesamiento Inteligente de Señales, que se impartió como posgrado (2003-2005), así como en la Red ALFA BioSenInt.

Por otro lado, al estar la mayoría de los laboratorios virtuales basados en simulación con Matlab, los alumnos tienen gran flexibilidad para su uso ya que este software está disponible en el Aula de Informática de la Facultad de CC. Físicas (40 puestos), en varios laboratorios de la Facultad de Informática (20 puestos cada uno), y en el Laboratorio de Ingeniería de Sistemas y Automática (12 puestos) de la facultad de Físicas, donde están las placas DSP. La utilización de la herramienta distribuida a través de la web no tiene requerimientos espaciales ni temporales para los alumnos puesto que tienen conocimientos y fácil acceso a Internet.

Además, estas herramientas pueden emplearse como recurso de apoyo a la docencia en la impartición de clases de teoría, aprovechándose del realismo y dinamismo de las experiencias prácticas y de la simulación por ordenador. Igualmente pueden utilizarse como valioso apoyo en el laboratorio, donde los alumnos pueden ejercitar sus conocimientos manejando ellos mismos las distintas posibilidades que les brindan y favoreciendo así el aprovechamiento de las clases y la asimilación de la materia.

Por último, además de poderse utilizar en distintas asignaturas, destacar el carácter multidisciplinar de estos laboratorios ya que se pueden aplicar a numerosos problemas en áreas de interés creciente. Por ejemplo, para señales de fusión termonuclear, señales de voz o de audio, imágenes, señales de dispositivos de control, señales médicas, etc. Se pueden también aplicar con distintos enfoques dentro de cada área: para diagnóstico y reconocimiento de señales, para clasificación, para eliminación de ruido y perturbaciones, detección de fraude, prevención, etc.

5.1 Metodología

Al alumno se le facilitan una serie de herramientas de simulación muy sencillas e inmediatas de manejar, donde puede observar de forma instantánea y de una manera gráfica lo que está haciendo. Además, la utilización de estos laboratorios le permite trabajar en temas de procesamiento pero abstrayendo toda la problemática de la programación y de la matemática que hay detrás. El programa ya se encargará de realizar los cálculos y llamar a las rutinas de simulación necesarias y además visualizará de forma gráfica la forma de las señales.

La realización de las prácticas con estos laboratorios virtuales se puede estructurar en varias fases bastante generales, cada una de ellas con distintos objetivos:

- la primera, la realización de la práctica en su formulación básica, para adquirir los conceptos fundamentales necesarios para llevarla a cabo y entender su funcionamiento; así el alumno comprende cómo funciona, la finalidad de la práctica, los conceptos que se quieren resaltar con ella, etc.; sería una labor de síntesis.
- otra, el trabajo experimental, donde el alumno puede modificar varios parámetros de la práctica, observar su funcionamiento y sacar conclusiones que le ayuden a comprender las explicaciones teóricas; en definitiva, una labor de análisis.
- Se puede también entender como un tercer paso la autoevaluación por parte del alumno de su aprendizaje y el asesoramiento del profesor para resolver dudas, etc.

Así, la metodología propuesta, junto con la supervisión del profesor, proporciona un modo de aprendizaje rápido y eficiente, sobre todo para establecer relaciones entre conceptos vistos en teoría y los experimentos. Además permite la repetición por parte del alumno de las prácticas las veces necesarias hasta conseguir el resultado que busca o entender un determinado comportamiento, y le ayuda a avanzar en su estudio de una forma natural, motivándole a preguntar el por qué y cómo del funcionamiento de los distintos tipos de señales, el conocimiento que se puede extraer de su procesamiento, etc.

El seguimiento del proceso de aprendizaje tiene un doble objetivo: guiar al alumno en los aspectos que debe hacer hincapié al realizar la práctica, puesto que en ocasiones los alumnos no aprecian ciertos aspectos de la realización de las mismas que son importantes; y permitir al profesor conocer en

qué grado cada alumno va dominando los conceptos impartidos en las clases teóricas.

Con esta metodología de realización de prácticas en los laboratorios tradicionales y virtuales, tanto locales como de acceso a través de la web, se pretende también incentivar al alumno a que pregunte y participe en las clases y tutorías, a raíz de las dudas que surjan al realizar esas experiencias, y que supongan una motivación para su estudio y profundización, y fomenten su relación con otros alumnos y con el profesor.

Por lo tanto, la metodología propuesta incorpora estilos de aprendizaje autónomos y aplica procesos de evaluación acordes con la renovación metodológica del EEES (Santos, 2007).

6. EVALUACIÓN DEL IMPACTO

Los laboratorios virtuales planteados se han utilizado, como se ha comentado en la sección 5, en distintas asignaturas de carreras que se imparten en las Facultades de CC. Físicas y de Informática de la UCM. A modo ilustrativo se presentan algunos datos de la impartición de la asignatura Transmisión de Datos (4'5 créditos, optativa en CC. Físicas y Complemento de Formación en Ingeniería Electrónica), para evaluar el impacto de la incorporación de estas metodologías de prácticas.

En primer lugar cabe remarcar que esta asignatura se ha impartido de forma teórica junto con ejercicios y ejemplos matemáticos en los anteriores planes de estudio, en la especialidad de Electrónica de la carrera de Físicas. No se contaba con laboratorios para esta materia para los alumnos. Desde el curso 2002/03 se empezó a impartir Transmisión de Datos con el programa que actualmente se mantiene, y se incorporaron de forma paulatina los laboratorios virtuales presentados en este trabajo.

Para evaluar la aceptación por parte de los alumnos y el efecto sobre el aprendizaje que conlleva el laboratorio virtual se ha realizado un estudio estadístico de los resultados durante los cursos 2002/03, 2003/04, 2004/05 y 2007/08 así como con los datos de una encuesta sobre dedicación a la asignatura. Los datos medios que resultan de la evaluación de los resultados de esta asignatura durante los cuatro cursos académicos citados se presentan en las Figura 13 y 14. La asignatura se ha impartido a un solo grupo de unos 35 alumnos matriculados cada año.

Los guiones de las prácticas están disponibles para los alumnos en el campus virtual (anteriormente en la web docente del profesor). Además se imparte una explicación sobre las mismas en las clases de teoría, en las que también se comentan aspectos prácticos del uso de las herramientas. Los alumnos pueden realizarlas durante los horarios previstos para las prácticas tuteladas en los laboratorios correspondientes o por su cuenta, en los horarios que tiene el aula de informática para libre disposición de los alumnos, donde está instalado el software y las herramientas para que puedan hacerlas. En cualquier caso, en el momento de la entrega los alumnos deben responder a una serie de cuestiones sobre las mismas.

La realización de prácticas fomenta la asistencia de los alumnos y les exige una mayor dedicación, que redunda en la obtención de mejores calificaciones. En las figuras 13 y 14 se puede observar, sobre el 76% de alumnos que se presentan al examen (el resto no lo hacen en ninguna de las convocatorias anuales, por lo que no se les ha podido en hacer la encuesta), la media de asistencia y dedicación por curso en % de horas a las que asisten.

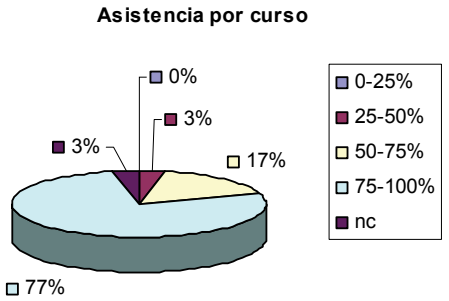


Figura 13. Datos medios de asistencia por curso de la asignatura Transmisión de Datos.

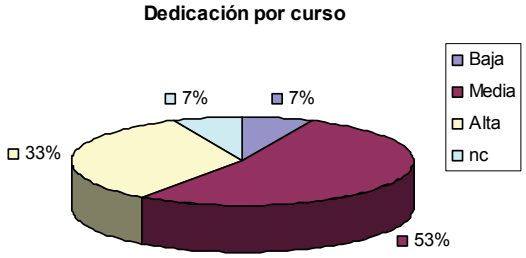


Figura 14. Datos medios de dedicación por curso a la asignatura Transmisión de Datos.

En el curso 2007/08 se ha incluido esta asignatura como piloto, por lo que a los alumnos se les ha realizado una encuesta específica sobre algunos aspectos concretos de su dedicación a esta materia, que se muestran en la Figura 15. Es significativo el hecho de que al estudio de la asignatura le estén dedicando más de tres horas de media semanales, que son las que se imparten de esta materia.

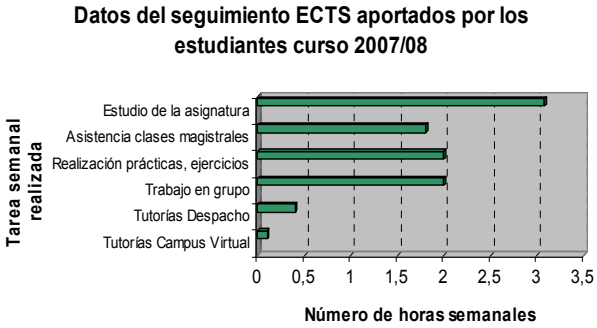


Figura 15. Resultados de dedicación de los alumnos a la asignatura Transmisión de Datos en el curso 2007/08.

Se ha constatado también que la asistencia de los alumnos a las clases de laboratorio es mayor que a las clases teóricas. La mayoría de los alumnos que han hecho las prácticas han aprobado la asignatura. Entre los que no han hecho las prácticas están todos los que han suspendido y alguno que sí ha aprobado. Esto permite concluir que el uso de laboratorios en general les ayuda a comprender mejor los conceptos del procesamiento de

las señales. Además se les facilita que pregunten ya que el contacto con el profesor es mayor en el laboratorio, y acuden con más frecuencia a tutorías puesto que las dudas que se les plantean no están resueltas en un manual o libro de texto.

A la pregunta del cuestionario sobre "Las sesiones prácticas impartidas por este profesor me han servido para entender la asignatura", la media que resulta sobre 5 es de 3. En general los alumnos en las preguntas abiertas sugieren más horas de laboratorio y reducir las de demostraciones matemáticas.

Tras una fase inicial de familiarización con las herramientas, los alumnos han encontrado éstas muy sencillas de manejar lo que les ha permitido hacer las prácticas dedicándoles menos tiempo del previsto. Las principales dificultades que han encontrado al realizar los experimentos eran debidos al acceso a la red desde la universidad (la licencia de Matlab está en un servidor), y a la presentación de los interfaces por la resolución de las pantallas. La mayoría no han tenido problemas para realizarlas.

Respecto a las diversas herramientas, los alumnos han manifestado que la que les ha resultado más sencilla de utilizar ha sido PDS: Procesamiento Digital de Señales, pero las que más les han interesado han sido las prácticas con señales de audio, en las que han tenido que involucrarse más a la hora de componer una melodía y han podido comprobar los efectos del procesamiento no sólo de forma gráfica sino también reproduciendo las señales y escuchándolas.

Además los alumnos no han realizado ningún tipo de prácticas sobre procesamiento de señales con anterioridad, por lo que estas herramientas tienen gran aceptación.

7. CONCLUSIONES

Se ha generado un material de prácticas que constituyen una serie de laboratorios virtuales de apoyo a asignaturas de gran importancia en algunas carreras científico-técnicas. Estas prácticas permiten entender y trabajar con conceptos avanzados del procesamiento de señales en general.

Las herramientas computacionales desarrolladas, de fácil manejo, incorporan unas facilidades gráficas que permiten al alumno realizar las experiencias propuestas para incorporar adecuadamente las enseñanzas teóricas de las materias relacionadas, que en algunas ocasiones requieren mucho tratamiento matemático, lo que les puede hacer perder la visión global o la utilidad de esos conocimientos

La experiencia demuestra que, efectivamente, el aprendizaje mejora cuando se utilizan recursos que motivan el interés del alumno, lo que suele estar unido en el caso de materias técnicas a encontrar aplicaciones prácticas a lo que aprenden. En el estudio hecho sobre el impacto de estos laboratorios virtuales en el aprendizaje se ha podido constatar que los alumnos valoran muy positivamente la realización de este tipo de experiencias.

Además, por la metodología seguida para las prácticas utilizando estas herramientas, y por la interactividad que permiten, el alumno puede autoevaluarse y por lo tanto aprender de su propio trabajo, analizar resultados, sacar conclusiones, etc. También se dota al profesor de mecanismos para el seguimiento del aprendizaje de los alumnos.

En otro orden, se ha conseguido también introducir innovaciones en las funciones y métodos docentes previstos para esta materias técnicas que facilitan la participación de los alumnos y por tanto el aprovechamiento de las enseñanzas. Otro aspecto positivo es que se fomenta la aplicación de herramientas computacionales que facilitan el trabajo del alumno y generan estilos autónomos y activos de aprendizaje, o de sistemas reales (placas DSP) con la que se enfrentará en su futuro profesional.

El desarrollo de las experiencias descritas tiene como principal objetivo reducir la brecha existente entre la teoría y la práctica real del procesamiento digital de señales.

Como trabajo futuro se propone incorporar nuevos laboratorios virtuales que cubran otros aspectos del contenido de esta materia, por ejemplo, plantear pequeños proyectos de tratamiento de imágenes, de señales de audio, generación de señales a partir de un modelo de la glotis, etc. Sería interesante trabajar con señales reales para acercar a los alumnos aplicaciones del ámbito de la medicina, de la fusión, etc. Otra línea abierta es la de implantar un laboratorio remoto con las placas DSP para beneficiar a todos los alumnos con su uso, ya que actualmente contamos con un número muy limitado y deben realizar las experiencias en grupos.

8. AGRADECIMIENTOS

Los autores quieren agradecer la colaboración de los alumnos J. Klaus González y Cristina de Santos en el desarrollo de algunas de las herramientas, así como la ayuda económica de los Proyectos de Innovación Educativa de la UCM 2003/7 y 2004/96.

9. REFERENCIAS

Candelas, F.A.; Torres, F.; Gil, P.; Ortiz, F., Puente, S.;, Pomares, J. (2004). *Laboratorio virtual remoto para robótica y evaluación de su impacto en la docencia*. RIAI **1**(2), 49-57.

Chassaing, R. (1999) *Digital Signal Processing. Laboratory Experiments Using C and the TMS320C31 DSK*. Wiley-Interscience.

Chassaing, R. (2005) *Digital Signal Processing and Applications with the C6713 and C6416 DSK*. Wiley-Interscience.

Dormido, S. (2004). *Control Learning. Present and Future*. IFAC Annual Reviews in Control, Willey, **28**(1), 115-136.

Elali, T.S. (2003). *Discrete Systems and Digital Signal Processing with Matlab*, 2003. CRC Press.

Esquembre, F. (2008) EJS. [http://fem.um.es/Ejs]

European Union (1999). The European Higher Education Area. Convened in Bologna, 19th of June 1999. [http://www.sc.ehu.es/siwebso/Bolonia/textos/AEES_EHEA/Bologna_declaration.pdf]

Farias, G., Dormido, S., Esquembre, F., Santos, M., Dormido-Canto, S. (2008). Laboratorio virtual de reconocimiento de patrones usando Easy Java Simulations y Matlab. *Actas XXIX Jornadas de Automática*, Tarragona.

Farias, G., Santos, M., Dormido-Canto, S. (2005). Desarrollo de una aplicación para la integración de técnicas de reconocimiento de patrones. *Actas XXVI Jornadas de Automática*, 209-215.

Guzmán, J.L.; Rodríguez, F.; Berenguel, M.; Dormido, S. (2005). *Laboratorio virtual para la enseñanza de control*

Heck B.S. editor (1999) Special report: Future directions in control education, IEEE Control Systems Magazine, **19**(5), 35-58.

Hilera J, Martínez V. (1995) *Redes Neuronales Artificiales. Fundamentos, modelos y aplicaciones*. Ed. Rama.

Ifeachor, E.C., Jervis, B.W. (1993) *Digital Signal Processing.* Ed. Addison-Wesley.

Karrenberg**,** U.A. (2002). *An Interactive Multimedia Introduction to Signal Processing***.** Springer

Mathworks Inc. (2002). The Student Edition of MATLAB. Prentice Hall

Oppenheim, A.V., Willsky A.S. and S. Hamid (1997). *Signal and Systems*, Ed. Prentice Hall.

Oppenheim, A. and R. Schafer (1999). *Discrete-time signal processing*, Ed. Prentice Hall.

Rahkila, M. and M. Karjalainen (1997). *An interactive DSP tutorial on the web.* ICASSP'97, IEEE, 2253-2256.

Rahkila, M. and M. Karjalainen (1998). *Considerations of computed based education in acoustic and signal processing.* In Frontiers in Education Conference 98, IEEE, 679-684.

climático de invernaderos, RIAI, **2**(2), 82-92.

Sánchez, J.; Morilla, F.; Dormido, S.; Aranda, J.; Ruipérez, P. (2002). *Virtual control lab using Java and Matlab: a qualitative approach*, IEEE Control Systems Magazine, **22**(2), 8-20.

Santos, M. and J. K. González (2007). A Simulation Tool for digital signal processing teaching. *Proceedings of INTED* International Technology, Education and Development Conference, IATED.

Santos, M. (2007). Integrating Different Teaching Methodologies for Technical Subjects. *Proceedings of INTED* International Technology, Education and Development Conference*,* IATED.

Smith, S (2002). *Digital Signal Processing: A Practical Guide for Engineers and Scientists*. Ed. Newnes

Spectrum Digital Incorporated. *TMS320C6713 DSK Technical Reference*. (2003).

Stearns, S.D. (2002). *Digital Signal Processing with examples in Matlab*. CRC Press.

Vapnik V. (2000) *The Nature of Statistical Learning Theory*, 2º Edition, Springer.

Article 7

Dynamic clustering and modeling approaches for fusion plasma signals

7.1 Bibliographic Description

Title

Dynamic clustering and modeling approaches for fusion plasma signals.

Citation

J.A. Martín, M. Santos, G. Farias, N. Duro, J. Sánchez, R. Dormido, S. Dormido-Canto, J. Vega, H. Vargas, (2009) Dynamic Clustering and Modeling Approaches for Fusion Plasma Signals, *IEEE Transactions on Instrumentation and Measurement*, ISSN 0018-9456, Volume 58, Number 9, Pages 2969-2978.

Abstract

This paper presents a novel clustering technique that has been applied to plasma signals to show its utility. It is a general method based on a partitioning scheme that has been proven to be efficient for purposes of analysis and processing of fusion plasma waveforms. Moreover, this paper shows how the information given by the clustering can be used to produce a concise and representative model of each class of signals by applying different

modeling approaches. Neuro-fuzzy identification and time-domain techniques have been used. These models allow the application of procedures to detect anomalous behaviors or interesting events within a continuous data flow that could automatically trigger the execution of some experimental procedures. Previously, an in-depth analysis and a preprocessing phase of the waveforms have been carried out. These procedures have been applied to plasma signals of the TJ-II Stellarator fusion device with encouraging results.

References

C. Alejaldre et al. (1999); K. Byungwhan, C. Seongjin (2007); S. Dormido-Canto et al. (2004); S. Dormido-Canto et al. (2006); R. O. Duda, P. E. Hart, D. G. Stork (2001); N. Duro et al. (2006); G. Farias et al. (2006); A. Gammerman, V. Vovk, G. Shafer (2005); M. R. Garey, D. S. Johnson (1979); J. S. Jang (1993); J. S. Jang (1994); J. S. Jang, C. T. Sun, E. Mizutani (1997); T. R. Jensen, B. Toft (1995); J. B. MacQueen (1967); S. Mallat (2001); J. A. Martín et al. (2007); H. Nakanishi, T. Hochin, M. Kojima(2003); J. Shawe-Taylor, N. Cristianini (2000); J. Vega (2007).

Impact Factor

IEEE Transactions on Instrumentation and Measurement has an impact factor of 1.214 according to Thomson Reuters Journal Citation Reports (2011).

IEEE TRANSACTIONS ON INSTRUMENTATION AND MEASUREMENT, VOL. 58, NO. 9, SEPTEMBER 2009

Dynamic Clustering and Modeling Approaches for Fusion Plasma Signals

J. A. Martín H., M. Santos Peñas, G. Farias, N. Duro, J. Sánchez, R. Dormido, S. Dormido-Canto, J. Vega, and H. Vargas

***Abstract*—This paper presents a novel clustering technique that has been applied to plasma signals to show its utility. It is a general method based on a partitioning scheme that has been proven to be efficient for purposes of analysis and processing of fusion plasma waveforms. Moreover, this paper shows how the information given by the clustering can be used to produce a concise and representative model of each class of signals by applying different modeling approaches. Neuro-fuzzy identification and time-domain techniques have been used. These models allow the application of procedures to detect anomalous behaviors or interesting events within a continuous data flow that could automatically trigger the execution of some experimental procedures. Previously, an in-depth analysis and a preprocessing phase of the waveforms have been carried out. These procedures have been applied to plasma signals of the TJ-II Stellarator fusion device with encouraging results.**

***Index Terms*—Dynamic clustering, fusion plasma signals, hybridizing intelligent techniques, neuro-fuzzy identification, signal modeling.**

I. Introduction

MEASUREMENTS in long-pulse devices require the use of intelligent techniques to detect interesting events, unexpected behaviors, or anomalies within a continuous data flow. The importance of these discoveries lies in the physical interpretation of those anomalous events and in their prevention. This detection might trigger the execution of some experimental procedures, such as increasing the sampling rates, starting data sampling in additional channels, or notifying the event to other diagnostics sensors. In a first approach, an interesting event can be any nonaverage behavior in the temporal evolution of the waveforms.

To be able to detect these events, models that represent the expected behavior of the different types of signals are required. Before obtaining the model of each class of waveform, a novel dynamic clustering procedure has been applied. This general clustering method has been proven to be efficient for the analysis and processing purposes of the type of plasma signals for which it has been applied on so far, that is, the plasma signals of the TJ-II fusion device, with encouraging results.

The proposed clustering technique is based on a partitioning method. The strategy consists of generating a triangular matrix with the values of a mathematical measurement of the similarity, i.e., the normalized scalar product (NSP), between each couple of waveforms and the application of a threshold to generate dynamic clusters based on it. From other works in which we have dealt with these signals [3], [4], [6], [7], [16], it can be derived that the most efficient procedure for the real-time measurement of similarity of the TJ-II plasma fusion signals is the NSP. For this reason, the new dynamic clustering method that we have developed is directly based on the information provided by this distance measurement. However, there are some remarkable differences between the common "prototype-based" clustering and other similar classification techniques and the dynamic partitioning clustering technique that is proposed in this paper.

Furthermore, this paper shows how the information provided by the clustering can be used to obtain a concise and representative model of each class of plasma signals by applying different modeling approaches. In that way, the expected patterns of each group are obtained, and they make it possible to detect anomalies.

Two different approaches have been used to obtain these models: neuro-fuzzy identification and time domain. They confirm some of the results obtained by the clustering.

Neuro-fuzzy techniques can be used to characterize signals. For example, in [2], we can read how these techniques are used to develop a predictive model of another type of plasma signals and how it succeeds in capturing the nonlinear plasma dynamics. In our case, we first applied a classification procedure. Then, the model that represents each class of plasma signal is identified by means of fuzzy inference systems (FIS), which are generated by applying adaptive neuro-fuzzy techniques. The purpose of this neuro-fuzzy modeling is to identify patterns for natural groups of data from large data sets to produce a concise representation.

In addition, we have obtained a rough model of the signals in the temporal and frequency domain. These patterns allow us to confirm that the shape of the models gives significant information that can help to identify the signals rather than other quantitative parameters, such as the Fourier transform, for example. They also help to prove the validity of the synergy that

Manuscript received January 30, 2008; revised July 28, 2008. Current version published August 12, 2009. The Associate Editor coordinating the review process for this paper was Dr. Jesús Ureña.

J. A. Martín H. and M. Santos Peñas are with the Facultad de Informática, Universidad Complutense de Madrid, 28040 Madrid, Spain (e-mail: jamartinh@fdi.ucm.es; msantos@dacya.ucm.es).

G. Farias, N. Duro, J. Sánchez, R. Dormido, S. Dormido-Canto, and H. Vargas are with the Department of Computer Science and Automatic Control, Universidad Nacional de Educación a Distancia, 28040 Madrid, Spain (e-mail: gfarias@bec.uned.es).

J. Vega is with the Data Acquisition Group, EURATOM/CIEMAT Association for Fusion, 28040 Madrid, Spain.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIM.2009.2016798

combines the new clustering method that is proposed here and the application of the intelligent neuro-fuzzy technique, as the models obtained are quite similar.

Previously, an in-depth analysis of the waveforms has been carried out, and an analytical preprocessing has been applied to prepare the signals for their classification.

To summarize the whole process, first, a processing of the signals is carried out to make them ready to apply some procedures to them (measurement of similarity, comparisons, a clustering strategy, etc.). Once they have been processed in this way (normalization, offset removal, noise reduction, etc.), a clustering procedure is applied to them to obtain the groups of similar signals. Then, we obtain models for those clusters that represent each class by using two different approaches. The goal is to show how those models give similar information about the signals and to allow us to detect anomalous behavior, besides giving a better understanding of the signals.

This paper is organized as follows. Section II summarizes the in-depth analysis and the preprocessing that have been carried out on the plasma signals for the clustering. In Section III, a new dynamic clustering procedure is proposed, and some classification results are presented. Section IV shows the models obtained when applying neuro-fuzzy identification techniques for these groups of signals. Section V shows the model obtained for the plasma waveforms in the temporal domain. This paper ends with conclusions in Section VI.

II. Signal Analysis and Preprocessing for Classification

The proposed clustering and modeling methods have been applied to the TJ-II stellarator database. TJ-II is a medium-sized stellarator fusion device [1] (Heliac type, magnetic field $B_0 \leq 1.2$ T, average major radius $R(0) = 1.5$ m, average minor radius ≤ 0.22 m) located at CIEMAT (Madrid, Spain) that can explore a wide rotational transform range. TJ-II plasmas are produced using electron cyclotron resonance heating (ECRH) (two gyrotrons, 300 kW each, 53.2 GHz, second harmonic, X-mode polarization) and additional neutral beam injection (NBI, 300 kW). At present, 940 digitization channels are available for experimental measurements in TJ-II. Fusion devices generate a massive database. Typically, thousands of signals with high dimensionality are collected per discharge.

Two different steps are needed before obtaining the models. First, a mathematical preprocessing is applied so that all the signals present a uniform representation that allows them to be compared. It may be needed to make the number of samples equal, to remove the offset, to normalize the amplitude and time duration, etc. Second, the clustering (extracting features) and the classification of the signals into groups are carried out. These procedures allow us to identify similar waves and to group them. This way, a model of each family of signals can be obtained by applying identification techniques. This phase produces valid models of each group of signals. This pattern is necessary to detect anomalous behaviors within the millions of waveforms of the database as the space where to compare is reduced.

TABLE I
Classes of Signals of the TJ-II Database

Classes of signals	Description
BOL5	Bolometer signal
ECE7	Electron cyclotron emission
RX306	Soft x-ray
ACTON275	Spectroscopic signal (CV)
HALFAC3	Hα
Density2	Line averaged electron density

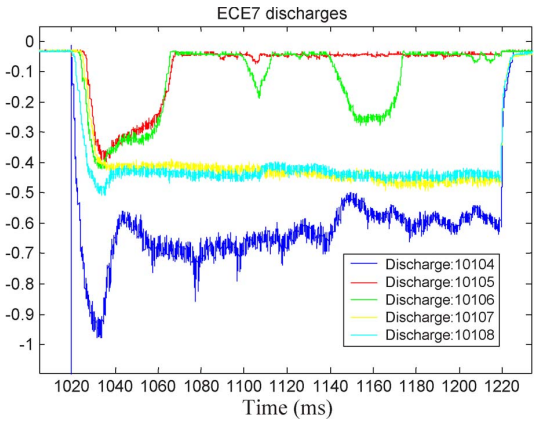


Fig. 1. Amplitude of discharges of the ECE7 signal.

The signals of the TJ-II database belong to one of the classes shown in Table I. Each of them describes a particular measurement of a physical characteristic of the plasma. For instance, a combination of the bolometer and X-ray systems can be used to characterize the temporal evolution of the plasma density. The data that these sensors provide are 2-D data, where one of the coordinates is time, and the other coordinate corresponds to the amplitude. These signals can be made up of millions of samples.

From each of these six sensors (Table I), thousands of signals are obtained. In Fig. 1, it is possible to see some of the features of the electron cyclotron emission (ECE7) waveform for five discharges (10104–10108). Each of them represents the temporal evolution of the plasma during a discharge of the TJ-II stellarator. The 2-D signals start when the logical OR of the three reference inputs is positive. These inputs correspond to the so-called signals GR1 (first gyrotron), GR2 (second gyrotron), and IACCEL1 (neutral beam injector).

It is also possible to see in Fig. 1 that these signals present an amplitude offset of −0.05. The initial warm-up time starts at around 1020–1030 samples. The negative polarization of these signals and the high frequency noise that they present can be noticed.

The norms of the signals, the number of samples, and the sample period have been calculated to gain insight into the behavior of the signals.

However, not all the signals from the same sensor present the same behavior. It is obvious in Fig. 1 that the duration of the plasma for two of the discharges (10105 and 10106) was

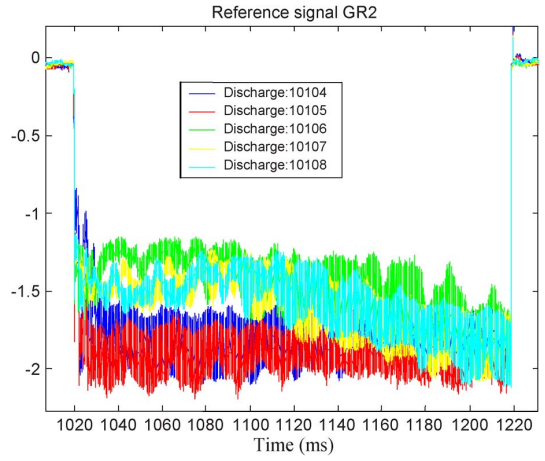


Fig. 2. Amplitude of the reference input GR2.

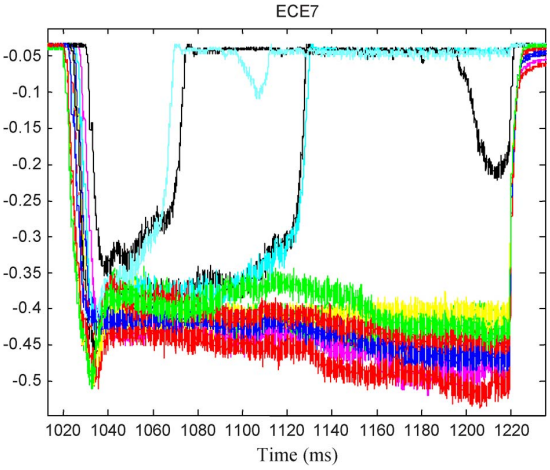


Fig. 3. Amplitude and profile of different discharges of the ECE7 signal.

too short. Otherwise, the shape of the signals is quite similar (a flat area between the beginning and the end, and an initial overshoot). That is why it is possible to assume that there should be a pattern for each class of signals as long as there is plasma, and this model would help to find anomalous behavior and outliers.

To rule out every possible factor that could have influenced the behavior of the discharges in Fig. 1, the reference signals were checked. Fig. 2 shows the GR2 input for the previous discharges. Although they do not have any physical meaning, they are the ones that produce the plasma. It is possible to see that there is no odd behavior on them that justifies the two anomalous discharges of ECE7 in Fig. 1.

We have carried out a similar analysis of all the discharges of the different classes that are available. Most of them present similar behaviors, and they are quite different regarding the anomalous behavior (see Fig. 3, where discharges 10107–10121 have been depicted). It is important to emphasize that the raw waveforms present strong differences between them, although they had been obtained by the same sensor. It depends on the configuration of each experiment. For example, the offset, the amplitude, or the polarity can change from one experiment to another for the same type of signal, although the plasma behavior they represent is the same. These features do not really affect the shape of the signal and, therefore, neither the model in that sense, but they can produce some errors when identifying the patterns.

For this reason, signal preprocessing is required to establish a right correlation between the different signals of each experiment. Therefore, this comparison will be independent of the amplification factor and of the polarity.

Thus, when obtaining a robust model for each group of signals, the following parameters must be normalized: sample time, offset, amplitude, number of data points (length), and polarity.

The procedure that has been applied to normalize those characteristics for each signal is the following.

1) Apply a spline interpolation so that the data are then homogeneous and, therefore, the number of samples is the same for all the signals.
2) Remove the offset of the signals. The mean value μ of the 100 first samples of the signal is calculated. Then, for each component y_i of the signal, the offset is eliminated by applying the simple formula

$$y_i = y_i - \mu. \tag{1}$$

3) Normalize to one. For each signal, the norm is calculated, and all the samples of the signal are divided by it.
4) Change the polarity if it is needed. Some signals present the inverse behavior to others. This modification is made by changing the sign of all the data.
5) Obtain the length of the signal. The criterion that has been applied is to compute the cumulative sum of the waveform data until it reaches the value of 0.995, and then the signal is truncated. This value has been chosen because, due to the experimental method, after 99.5% of the normalized signal, there is no phenomenon of scientific significance to observe, as the plasma drops when the heat ends (see Figs. 1 and 3). That is, the 99.5% of the signal contains all the important information.

In Fig. 4, the effects of this normalization process on a signal are shown.

A filter could have been applied to smooth the signal, but in this case, it is not necessary as the modeling produces that effect, and some information could be lost before clustering.

Some techniques can be applied to reduce the dimension of the signal, such as wavelet transform [15]. In this paper, we have chosen to work with the entire signal, i.e., the size of the signals we are dealing with is kept to 28501 samples. This way, the accuracy of the model is preserved.

However, even when applying this normalization procedure, the signals may present different characteristics. Fig. 5 shows an example of ECE7 signals that were considered valid for the experts, as there was plasma for all of them. The same result was given for our procedure when we apply a preclassification method to reassure that the signals were valid. This method consists of the following.

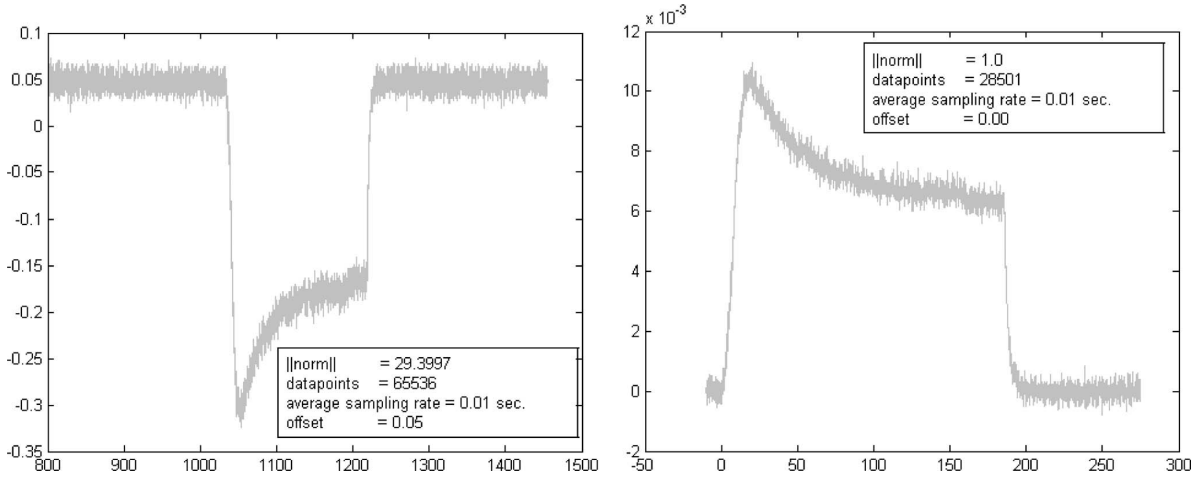


Fig. 4. Example of a signal before and after preprocessing.

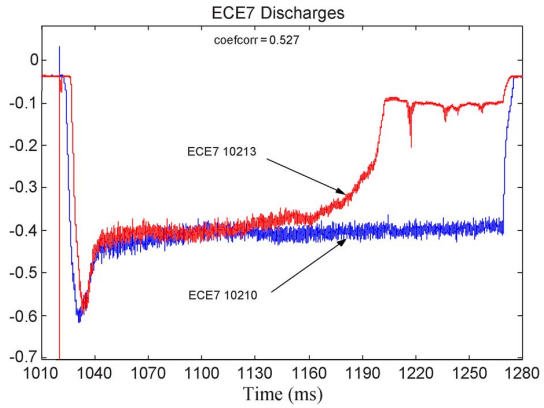


Fig. 5. Correlation threshold of 0.527.

1) Calculate the correlation matrix for all the discharges of that signal. Then, all the correlation coefficients of each column are multiplied. The largest value defines the discharge that is most similar to the rest, that is, the one that is considered the best pattern of all.
2) The correlation coefficients of the other signals and the pattern are compared. A correlation threshold has been defined, and if the correlation value exceeds this threshold, then the discharge belongs to that class of signals. Otherwise, it is an outlier or presents an anomalous behavior.

The efficacy of this detection procedure depends on the correlation threshold. After several experiments, a value of 0.7 for this factor has been proved to be a good value. Sometimes, even lower values can also work well. For example, taking a lower threshold of similarity (0.527), both signals in Fig. 5 were classified as ECE7. When consulting the experts about the correction of this result, they confirmed that both discharges were valid ECE signals. The plasma can be longer or shorter, but as far as there is plasma, the behavior of the signal is correct. These results will be confirmed in Section V when obtaining the temporal model.

Once this preprocessing has been applied to every signal, the clustering procedure is carried out to obtain the different groups of similar signals.

III. New Dynamic Clustering Procedure

Some clustering algorithms, such as k-means, support vector machines (SVMs), hierarchical clustering, etc. [8], [14], [18], work with a vector of features and apply a similarity measurement that is based on a function of a mathematical distance (for example, in [17], the Euclidean distance is used for these signals). This value allows us to find out if a signal is similar to a given signal or to a pattern or model [5].

These algorithms tend to minimize a cost function that, in fact, tries to decrease as much as possible the internal dispersion of the data of each cluster and, at the same time, to maximize the distance between clusters.

In this paper, taking into account the length of the waveforms and other previous results, a dynamic clustering strategy has been designed. The method that is proposed here differs from the other clustering approaches as it is directly based on partitioning, and it implicitly carries out the minimization and maximization of the traditional methods. It does not require the extraction of discriminatory features as it is only based on the NSP. Different publications make use of the measurement of the absolute value of the NSP as an effective way of comparing how similar two signals are [16], [19]. This similarity factor is usually chosen by several reasons, such as that the geometrical interpretation of the dot product is straightforward.

The NSP can be defined, for any two vectors $\mathbf{x}$ and $\mathbf{y}$, as

$$\text{NSP} = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x} \cdot \mathbf{y}\|}. \tag{2}$$

We have taken the absolute value of the NSP as a similarity value, so it belongs to the interval [0, 1].

Our strategy consists of generating a triangular matrix $\mathbf{U}$ with the values of the similarity measurement (NSP) between each couple of waveforms and the application of a threshold to generate dynamic clusters based on it.

That is, first of all, the similarity degree of every pair of signals is computed by means of the NSP. Then, a threshold value is defined by empirical methods and expert knowledge. This threshold, i.e., $\theta = (0, 1]$, means that any two different signals cannot belong to the same cluster if the value of the corresponding NSP between them is lower than that threshold.

Thus, each cell of the matrix that does not exceed the threshold represents a constraint in the partitioning process.

From a mathematical point of view, this problem can be represented by using graph theory. The problem is then reduced to the conventional graph-coloring problem [13]. This way, each signal corresponds to a vertex of a graph, and there will be an edge between every two vertices if and only if the value of the corresponding NSP does not exceed the defined threshold θ, which means that those signals are related and may belong to the same cluster.

As it is well known, the computational complexity of the exact solution for this problem belongs to the NP-complete problem class. That is, in general, it is very unlikely that it can accurately be solved in an efficient way [9]. However, there are some computational approaches to face it up. For instance, if some constraint is not allowed to be broken, the algorithm will then generate enough number of clusters to fulfill all the requirements. Another way of solving the problem would be to relax the constraints (turn them into soft constraints or fuzzy edges) and to predetermine the number of clusters.

This is the strategy that traditional clustering algorithms such as k-means or self-organizing feature maps (SOFM) apply [9].

However, due to the fact that this clustering is going to be applied to generate congruent models for the different classes of waveforms, and since a signal that does not belong to a specific class might have a significant effect on the model of that group if it is assigned to it, the ideal method of clustering should only be based on hard constraints.

This way, it is possible to assure that every pair of plasma waveforms that are members of the same cluster has an NSP value that exceeds the threshold θ. This means that their similarity degree is enough to assure that they belong to the same class.

The procedure of this clustering is as follows. Given a database of N signals, the similarity matrix $\mathbf{U}$ $(N \times N)$ has been calculated by taking into account the average values of the NSP between every pair of signals. Based on it, a new binary matrix is defined, where "1" means that the pair of signals (i, j) cannot be grouped because its similarity is lower than the established threshold. Otherwise, the value of the cell is 0.

To summarize, the clustering process is as follows (Fig. 6).

1) Based on a similarity measure, e.g., NSP, compute the similarity coefficient u_{ij} between each pair of signals (i) and (j). Build an upper triangular matrix $\mathbf{U}$ with these coefficients, where $i, j = 1, \ldots N$, with N being the dimension of the database, that is, the number of available signals.
2) Determine a similarity threshold $\theta = (0, 1]$ and apply it to every u_{ij}. For instance, set $\theta = 0.95$, which means that signals with $u_{ij} \geq \theta$ may belong to the same group.
3) Set each $u_{ij} \geq \theta$ in the $\mathbf{U}$ matrix to 0 and set the rest to 1.

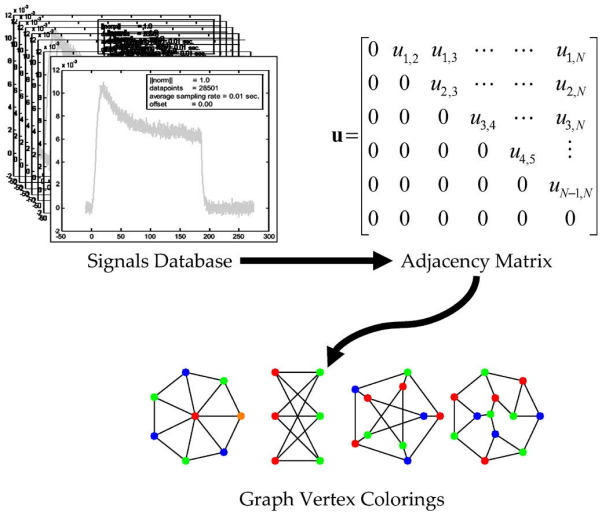


Fig. 6. Clustering process steps.

4) Finally, based on this adjacency matrix, run a graph-coloring algorithm and find a proper coloring such that the number of colors is minimized.
5) Then, take each color class as a cluster.

The result of applying this procedure as a clustering method is a set of different groups. For each of these groups, a model will be generated by neuro-fuzzy techniques, and using this model, it will be possible to detect unexpected events. In addition, the similarity matrix can dynamically be enlarged with new signals for a later classification.

As can be seen in Fig. 6, which shows the clustering process steps, the first stage is to obtain the adjacency matrix $\mathbf{U}$ (upper triangular matrix) from the signal database, where each u_{ij} in the matrix represents the similarity coefficient between the pair of signals i and j. Once this matrix is obtained, a graph-coloring algorithm is applied to the graph that represents the adjacency matrix $\mathbf{U}$, which obtains a proper coloring (i.e., a valid partitioning of the signals database) of the graph induced by $\mathbf{U}$.

This strategy has been applied to a set of $N = 289$ bolometric signals of type BOL5. They have been provided by the same sensor (in this case, a bolometer), although they belong to different discharges. Different numbers of groups and quite different classes can be found within them. Table II presents the results of this proposed clustering procedure based on the partitioning method. The signals are divided into groups according to their NSP values. Different similarity thresholds θ (0.85, 0.90, and 0.95, respectively) have been used, all of them quite demanding. The number of clusters and the number of elements in each class are shown in Table II.

As it is possible to see, in all cases, there is a main cluster that groups the vast majority of the signals. The number of clusters increases when the threshold increases, as could be expected.

All the signals are forced to be classified into any of the groups, and, therefore, more new clusters are generated to incorporate anomalous signals or spurious information.

That is, Table II shows that for all the different thresholds tested, there is always a stable group of waveforms where at least 75% of the signals are included. There are also some other

TABLE II
RESULTS OBTAINED BY DYNAMIC CLUSTERING

Clustering for a similarity threshold of 0.85
Cluster 1: 267 elements
Cluster 2: 12 elements
Cluster 3: 5 elements
Cluster 4: 3 elements
Cluster 5: 2 elements
Clustering for a similarity threshold of 0.90
Cluster 1: 268 elements
Cluster 2: 12 elements
Cluster 3: 4 elements
Cluster 4: 2 elements
Cluster 5: 2 elements
Cluster 6: 1 elements
Clustering for a similarity threshold of 0.95
Cluster 1: 216 elements
Cluster 2: 41 elements
Cluster 3: 12 elements
Cluster 4: 8 elements
Cluster 5: 5 elements
Cluster 6: 3 elements
Cluster 7: 3 elements
Cluster 8: 1 elements

clusters with fewer signals. If there are only one or two signals in a group, then that may mean that those signals are outliers. In this sense, this clustering method allows the detection of anomalies in an immediate way.

Then, a graph-coloring algorithm is applied to obtain an optimal partition, i.e., a collection of independent sets, which represent the respective clusters or signal groups.

It is important to notice that a graph-coloring problem will not necessarily produce a unique solution. In fact, it can generate a set of valid solutions for the same NSP threshold θ. This fact shows that this method is completely different from the standard clustering and classification approaches, since, in general, there is no global optimal solution for a given similarity threshold. While the majority of algorithms for data clustering and classification are prototype based, in this approach, there is no prototype at all, in the sense that a class is defined by the holistic interrelation between the signals that represent a cluster as a whole.

For pattern generation purposes and posterior detection or signal retrieval, we have decided that models will only be generated for those groups that include at least 10% of the signals.

For example, the NSP threshold of 0.95 produces two clusters of bolometric waveforms that fulfill that requirement. In the next section, neuro-fuzzy models will be obtained for those clusters.

To summarize, a clear contribution of this clustering strategy is that it helps to detect anomalous signals in an immediate way. However, the main advantage of this method is that it uses the same measure (NSP) to generate the clustering and also to retrieve similar signals from a fusion massive database in a posterior phase if the system is asked for doing so. This advantage can be compared to the methods presented in [3] and [6], where the retrieval of the signals is made by a different strategy from clustering.

IV. NEURO-FUZZY MODELING OF PLASMA WAVEFORMS

In this section, models that represent each group of plasma waveform are obtained by means of FISs, which are generated by applying adaptive neuro-fuzzy techniques to the available data [10].

The purpose of this neuro-fuzzy modeling strategy is to identify natural groups of data from a large data set to produce a concise representation of a type of signal. It can be seen as a pattern with which a new signal will be compared to classify it. On the other hand, these models will help in the searching and retrieval of similar signals, as each model represents a cluster, and, therefore, the searching space in which similar signals are more likely to be found is reduced.

A FIS is a model that maps input characteristics to input membership functions, input membership functions to rules, rules to a set of output characteristics, output characteristics to output membership functions, and output membership functions to a single-valued output or a decision associated with the output.

The computational procedure that accomplishes this membership function parameter adjustment is called ANFIS. The acronym ANFIS derives its name from "adaptive neuro-fuzzy inference system." It was developed in 1993 by J. R. Jang [10].

Using a given input/output data set, this procedure constructs a FIS whose membership function parameters are tuned using either back-propagation algorithm alone or in combination with a least-square-type method. This allows the fuzzy systems to learn from the data they are modeling.

Its adaptation properties allow the application of this method to different fields, such as adaptive control, processing and filtering, clustering, features extraction, modeling, etc. One of its main characteristics is that it can use hybrid learning methods to make it more efficient [12].

Adaptive neuro-fuzzy networks take advantage of both paradigms. On one hand, the FISs provide an intuitive mechanism to represent the knowledge at a high level by means of if/then rules. On the other hand, neural networks allow adaptation and a high level of learning and generalization. Therefore, neuro-fuzzy techniques can be very efficient for modeling real systems [11].

The ANFIS architecture is represented by a parametric FIS that is distributed in different layers. It can learn from data by adjusting the parameters of the model by using the training data. Therefore, it is necessary to have a set of data (input/output) for training the system. Once the signals have been preprocessed and classified in different groups, ANFIS is applied to each of these clusters to obtain the neuro-fuzzy model.

The number of membership functions is one of the most important parameters of the model. This number has to be determined by empirical knowledge. A small number of membership functions do not allow the optimum tuning of the model.

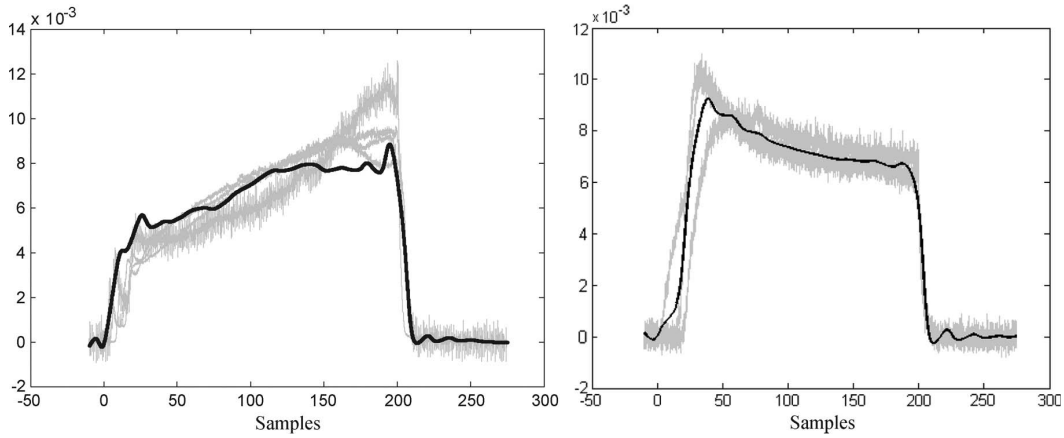


Fig. 7. (Gray) Plasma waveforms and (black) model of the cluster for (left) cluster 1 and (right) cluster 2.

However, a large number of them increases the computational time, and, therefore, it will make it difficult to implement the model in real time because of the large number of rules.

In this paper, we have applied the ANFIS procedure to identify the models of the two clusters of bolometric waveforms that contain more than 10% of the signals, which were obtained when the NSP threshold was set to 0.95 (according to Table II).

By empirical methods, 20 Gaussian membership functions have been proved to be necessary. They are initially symmetric, and their main parameters are adaptively tuned by the neuro-fuzzy learning system. First, some tests were carried out with fewer functions, but the model did not fit well the shape of the signals during the transitory state. The initial step was underestimated by the model, and the slope was not well represented.

Therefore, the number of fuzzy sets was increased to 20, and this number allowed us to model the transitory phase. This initial phase is important when analyzing the time evolution of the plasma waveforms. The final phase is less important as when there is no plasma, the signal abruptly drops, and this behavior is easier to be modeled.

Fig. 7 shows the models that have been generated by ANFIS for the first two clusters in Table II (Threshold = 0.95), i.e., those that have more than 10% of the signals. The left part of the figure shows the neuro-fuzzy model of cluster 1. The number of bolometric signals of this cluster that have been plotted (initially 216) has been reduced to make it clearer. The group of 41 bolometric signals that are plotted in Fig. 7 (right) corresponds to cluster 2.

Therefore, it is possible to see how the signals of the two clusters can be so different, although all the signals were provided by the same sensor. This way, a general perspective of the behavior of the system using different clusters is shown.

The results well-enough represent the phase where there is plasma during discharge. On the other hand, the model also quite satisfactory fits the starting and ending points of the signals.

As it was said before, a fuzzy model tends to smooth the curve, so it acts like a filter.

The neuro-fuzzy model can be seen as the pattern of that group of signals and can be used for different purposes when analyzing the signals. For example, the model of another type of bolometric signals has been used to detect outliers, as shown in Fig. 8. In this case, BOL1 waveforms and their model signals are depicted instead of BOL5.

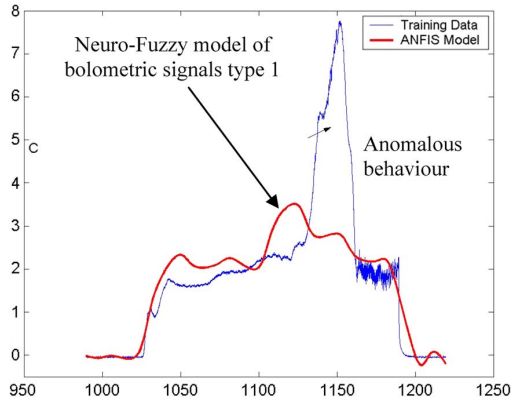


Fig. 8. Neuro-fuzzy model of bolometric signals type 1 and outlier discharge.

As far as we know, this neuro-fuzzy approach has not been applied to these specific types of plasma signals that are provided by those sensors (Table I). Other works, such as that presented in [2], used this technique for modeling, but with a different approach.

V. Classical Modeling Approach

Once we have a method to generate the clusters (Section III), to obtain a time-domain model of each class, we have analyzed the low-frequency components. As mentioned before, the entire signal approach is considered in this paper; in previous works that deal with this database, the size of the waveforms was reduced by different techniques [3], [6], [7].

The following parameters have been defined to describe the temporal behavior of the signals:

— CD: mean value of the signal before applying the reference inputs (offset);
— MS: mean value of the signal where there is plasma;

TABLE III
STATISTICAL VALUES OF THE RESPONSE TIME

TIS average	TIS deviation	TFS average	TFS deviation
10.9	5.3	10.3	10.9

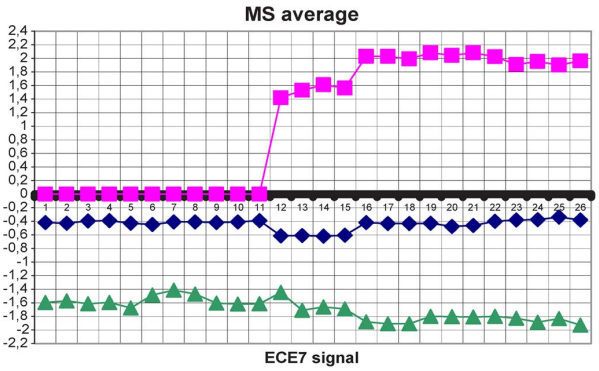


Fig. 9. MS average for 26 ECE signals (squares: GR1; triangles: GR2; rhombus: MS).

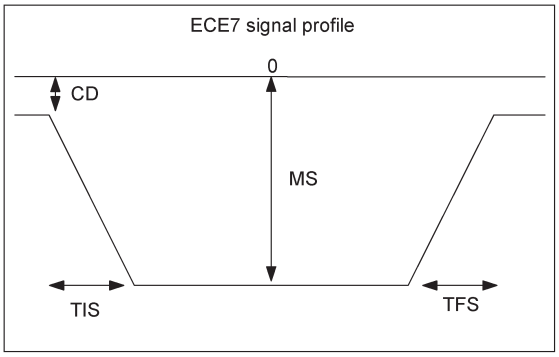


Fig. 10. Profile of the ECE7 signal.

— TIS: rising time (from CD to MS);
— TFS: dropping time (from MS to CD).

A statistical study of these values for 50 valid discharges of the electron cyclotron emission ECE7 signal gave the following average values in milliseconds (Table III).

Taking into account that the duration of the reference signals is around 200–250 ms, these ECE7 signals can be considered to be fast because the response time (rising time) is very small in comparison. That is, the mean value of the whole signal can be approximated to MS, as it does not seem to significantly vary even when considering the transitory phase of the signal.

To prove this, the MS value has been calculated for 26 ECE7 signals, and it is shown in Fig. 9 for each of then. As can be seen, this MS value remains constant, although the reference gyrotron signals that produce the plasma (GR1 and GR2) vary slightly, as can also be seen in Fig. 9. The average value obtained for the parameter MS of the 26 ECE7 signals is 0.44, and the standard deviation is 0.07.

Based on this behavior, we can sketch a profile of the ECE7 signal (Fig. 10), where the parameters defined to characterize the temporal response are also depicted.

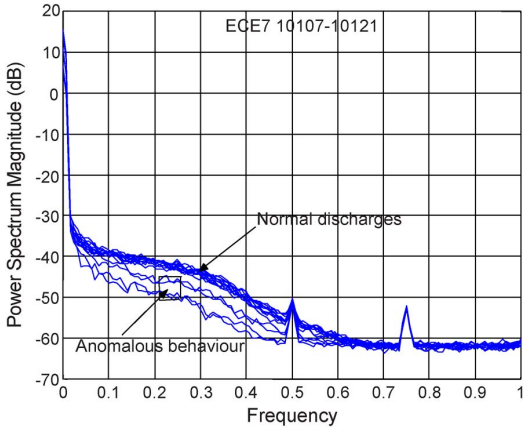


Fig. 11. Spectra of the ECE7 signals.

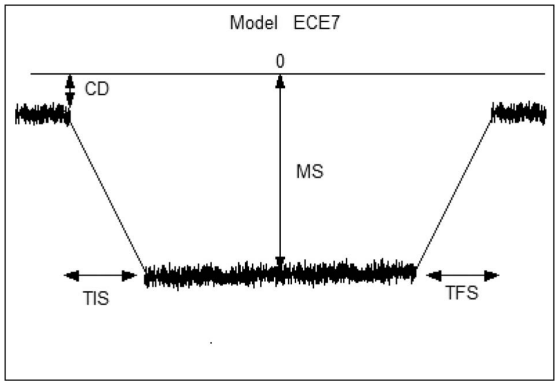


Fig. 12. Model of the ECE7 signal.

This general model in the temporal domain only shows the expected shape of that class of signals, but it helps to understand their behavior and to confirm the profile of the model obtained when applying neuro-fuzzy techniques. In this particular case, the polarity of the signals has not been changed. The same could be done with other types of plasma waveforms when the corresponding clusters have been obtained.

A. Information From the High Frequency Spectrum

We have also studied the information provided by the high-frequency components. However, in this frequency domain, one could hardly distinguish between anomalies and normal signals. Both the shape and the intensity range are quiet similar, particularly for normalized frequencies between 0.5 and 1. Fig. 11 depicts the same 15 discharges in Fig. 3, where four of them presented an anomalous behavior. If we had not known that they were irregular, then we could hardly have recognized it by the frequency spectra.

However, we can use this knowledge to refine on our model. In fact, we can add these high-frequency components to the temporal model previously obtained. By applying a high-pass frequency filter or taking the detail coefficients of the wavelet transform of the signal [15], we can obtain information to complete the model of the signal, as shown in Fig. 12, to see more clearly how the model would be.

This model, although simple, helps to confirm the statement made in Section II regarding Fig. 5. In this application, an important key to define a model is the shape of the signals. As long as the shape is maintained, the signals will be valid discharges of that specific sensor, even if the temporal parameters of this model vary. That is, the duration of the signal can vary, but if there is plasma, the signal can be classified by attending to the profile.

In previous works, we had obtained classification-oriented patterns [19]. They cannot be considered a general model in the sense that they only represent one of the signals of a specific class, i.e., the signal of the database most similar to the other waveforms of the same class. However, these patterns can help to confirm that the general model we have obtained in this section can be a good representation of the behavior of a class of plasma signals.

VI. Conclusion

This paper has presented two main contributions for the analysis of plasma signals. First, a new dynamic clustering procedure that it is based on a partitioning method is proposed. The strategy consists of generating a triangular matrix with the values of a mathematical measure of the similarity, i.e., the NSP between each couple of waveforms. Then, the dynamic clusters are generated based on this measurement, depending on a threshold.

Second, models based on different techniques have been obtained for each of these clusters. This identification phase produces valid patterns for each group of signals. The purpose of these models is to detect anomalous behaviors and interesting events within a discharge.

Another interesting contribution of this modeling is that it allows to confirm when a discharge is valid—when there is plasma—depending on the shape and values of the model.

These identification techniques (time-domain modeling and neuro-fuzzy inference) have not previously been applied to these specific types of plasma waveforms.

These procedures have been applied to electron cyclotron emission and bolometric signals of the TJ-II stellarator fusion device with encouraging results.

References

[1] C. Alejaldre, J. Alonso, L. Almoguera, E. Ascasíbar, A. Baciero, R. Balbín, M. Blaumoser, J. Botija, A. Brañas, B. de la Cal, E. Cappa, A. Carrasco, R. Castejón, F. Cepero, J. R. Cremy, C. Doncel, J. Dulya, C. Estrada, T. Fernández, A. Francés, M. Fuentes, C. García, A. García-Cortés, I. Guasp, J. Herranz, J. Hidalgo, C. Jiménez, J. A. Kirpitchev, I. Krivenski, V. Labrador, I. Lapayese, F. Likin, K. Liniers, M. López-Fraguas, A. López-Sánchez, A. de la Luna, E. Martín, R. Martínez, A. Medrano, M. Méndez, P. McCarthy, K. Medina, F. van Milligen, B. Ochando, M. Pacios, L. Pastor, I. Pedrosa, M. A. de la Peña, A. Portas, J. Qin, L. Rodríguez-Rodrigo, A. Salas, E. Sánchez, J. Sánchez, F. Tabarés, D. Tafalla, V. Tribaldos, J. Vega, B. Zurro, D. Akulina, O. I. Fedyanin, S. Grebenshchicov, N. Kharchev, A. Meshcheryakov, R. Barth, G. van Dijk, H. van der Meiden, and S. Petrov, "First plasmas in the TJ-II flexible Heliac," *Plasma Phys. Control. Fusion*, vol. 41, no. 1, pp. A539–A548, Mar. 1999.

[2] K. Byungwhan and C. Seongjin, "Adaptive Network-Based Fuzzy Inference Model of Plasma Enhanced Chemical Vapor Deposition Process," in *Lectures Notes in Computer Science*, vol. 4491. Berlin, Germany: Springer-Verlag, 2007, pp. 602–608.

[3] S. Dormido-Canto, G. Farias, R. Dormido, J. Vega, J. Sánchez, and M. Santos, "TJ-II wave forms analysis with wavelets and support vector machines," *Rev. Sci. Instrum.*, vol. 75, no. 10, pp. 4254–4257, Oct. 2004.

[4] S. Dormido-Canto, G. Farias, J. Vega, R. Dormido, J. Sánchez, N. Duro, M. Santos, J. A. Martin, and G. Pajares, "Search and retrieval of plasma wave forms: Structural pattern recognition approach," *Rev. Sci. Instrum.*, vol. 77, no. 10, pp. 10F 514-1–10F 514-4, Oct. 2006.

[5] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York: Wiley, 2001.

[6] N. Duro, J. Vega, R. Dormido, G. Farias, S. Dormido-Canto, J. Sánchez, M. Santos, and G. Pajares, "Automated clustering procedure for TJ-II experimental signals," *Fusion Eng. Des.*, vol. 81, no. 15–17, pp. 1987–1991, Jul. 2006.

[7] G. Farias, S. Dormido-Canto, J. Vega, J. Sánchez, N. Duro, R. Dormido, M. Ochando, M. Santos, and G. Pajares, "Searching for patterns in TJ-II time evolution signals," *Fusion Eng. Des.*, vol. 81, no. 15–17, pp. 1993–1997, Jul. 2006.

[8] A. Gammerman, V. Vovk, and G. Shafer, *Algorithmic Learning in a Random World*. Berlin, Germany: Springer-Verlag, 2005.

[9] M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*. New York: Freeman, 1979.

[10] J. S. Jang, "ANFIS: Adaptive-network-based fuzzy inference systems," *IEEE Trans. Syst., Man, Cybern.*, vol. 23, no. 3, pp. 665–685, May/Jun. 1993.

[11] J. S. Jang, "Structure determination in fuzzy modeling: A fuzzy CART approach," in *Proc. IEEE Int. Conf. Fuzzy Syst.*, 1994, pp. 480–485.

[12] J. S. Jang, C. T. Sun, and E. Mizutani, *Neuro-Fuzzy and Soft Computing*, ser. Math-Lab Curriculum Series. Englewood Cliffs, NJ: Prentice-Hall, 1997.

[13] T. R. Jensen and B. Toft, *Graph Colouring Problems*. New York: Wiley, 1995.

[14] J. B. MacQueen, "Some methods for classification and analysis of multivariable observations," in *Proc. 5th Berkeley Symp. Math. Stat. Probability*. Berkeley, CA: Univ. California, 1967, vol. 1, pp. 281–297.

[15] S. Mallat, *A Wavelet Tour of Signal Processing*, 2nd ed. New York: Academic, 2001.

[16] J. A. Martín, H. M. Santos, G. Farias, N. Duro, J. Sanchez, R. Dormido, S. Dormido-Canto, and J. Vega, "Dynamic clustering and neuro-fuzzy identification for the analysis of fusion plasma signals," in *Proc. Conf. IEEE Int. Symp. WISP*, 2007, pp. 1–6.

[17] H. Nakanishi, T. Hochin, and M. Kojima, "Search and retrieval methods of similar plasma waveforms," *Proc. 4th IAEA TCM Control, Data Acquisition, Remote Participation Fusion Res.*, Jul. 21–23, 2003, San Diego, CA.

[18] J. Shawe-Taylor and N. Cristianini, *Support Vector Machines and Other Kernel-Based Learning Algorithm*. Cambridge, U.K.: Cambridge Univ. Press, 2000.

[19] J. Vega, "Recent results on structural pattern recognition for fusion massive databases," in *Proc. Conf. IEEE Int. Symp. WISP*, 2007, pp. 1–6.

J. A. Martín H. received the B.S. and M.S. degrees in computer science from La Universidad del Zulia, Venezuela, in 2002, the Ph.D. degree in computer science, with a dissertation entitled "Studies on adaptive systems with applications in autonomous robots and intelligent agents," from the Universidad Politécnica de Madrid, Spain, in 2009, and the Advanced Studies Diploma on "A computational model of the equivalence class formation psychological phenomenon" from the Universidad Nacional de Educación a Distancia (UNED), Madrid, where he is currently working toward the Ph.D. degree on fundamentals of basic psychology.

Since 2005, he has been with the Department of Informatic Systems and Computing, Facultad de Informática, Universidad Complutense de Madrid. His main research areas are cybernetics, machine learning and machine perception.

M. Santos Peñas was born in Madrid, Spain. She received the B.Sc. and M.Sc. degrees in physic sciences (computer engineering) and the Ph.D. degree in physics from the University Complutense of Madrid (UCM), Madrid, in 1984, 1986, and 1994, respectively.

Since 1986, she has been with the Department of Computer Architecture and Systems Engineering, UCM, where she is currently a Senior Lecturer in system engineering and automatics. She has directed and participated in several research projects, and has numerous scientific publications. Her major research interests are intelligent control (fuzzy and neuro-fuzzy), process control and signal processing (machine learning, clustering, and pattern recognition), and modeling and simulation.

G. Farias received the degree in computer science from Chile de la Frontera de Temuco University, Temuco, Chile.

He has been a Fellow Student with the Department of Computer Science and Automatic Control, Universidad Nacional de Educacióon a Distancia, Madrid, Spain. His current research interests include simulation and control of dynamic systems, web-based laboratories, and pattern recognition.

N. Duro received the Ph.D. degree in science from the Universidad Nacional de Educacióon a Distancia (UNED), Madrid, Spain.

She is currently an Associate Professor with the Department of Computer Sciences and Automatic Control, UNED. Her current research interests are control processes, modeling and simulation of continuous processes, and the design of new systems for control education.

J. Sánchez received the Ph.D. degree in sciences from the Universidad Nacional de Educacióon a Distancia (UNED), Madrid, Spain, in 2001.

Since 1993, he has been an Associate Professor with the Department of Computer Science and Automatic Control, UNED. His current research interests are web-based systems for control education, networked control systems, event-based control, and pattern recognition in nuclear fusion.

R. Dormido received the degree in physics from Madrid Complutense University, Madrid, Spain, in 1995 and the Ph.D. degree in sciences from the Universidad Nacional de Educación a Distancia (UNED), Madrid, in 2001.

Since 1995, she has been an Associate Professor with the Department of Computer Sciences and Automatic Control, UNED. Her current research interests are robust control, the modeling and simulation of continuous processes, and the design of systems for control education.

S. Dormido-Canto received the M.S. degree in electronics engineering from the Universidad Pontificia de Comillas (ICAI), Madrid, Spain, in 1994 and the Ph.D. degree in physics from the Universidad Nacional de Educación a Distancia (UNED), Madrid, in 2001.

Since 1994, he has been with the Department of Computer Science and Automatic Control, UNED, where he is currently an Associate Professor of control engineering. His research and teaching activities are related with the analysis and design of control systems via Intranet or Internet, high-performance interconnection networks for cluster of workstations, and optimal control.

J. Vega was born in Madrid, Spain, on November 1, 1958. He received the M.S. degree from the Universidad Complutense de Madrid (UCM), Madrid, and the Ph.D. degree from the Universidad Nacional de Educación a Distancia, Madrid.

He is working on in nuclear fusion with CIEMAT, Madrid, and is very involved in the EURATOM Research Program on fusion projects, mainly, in the TJ-II stellarator and in the JET tokamak. He is the Head of the Data Acquisition Unit, Spanish Fusion National Laboratory by Magnetic Confinement. His previous research activities were plasma diagnostic techniques in soft X-ray radiation. His current research is focused on both remote participation systems and advanced data analysis methods.

H. Vargas received the degree in electronics from Chile de La Frontera de Temuco University, Temuco, Chile.

He is currently a Fellow Student with the Department of Computer Science and Automatic Control, Universidad Nacional de Educación a Distancia, Madrid, Spain. His current research is focused in the design of web-based systems for control education.

Article 8

Automated recognition system for ELM classification in JET

8.1 Bibliographic Description

Title

Automated recognition system for ELM classification in JET.

Citation

N. Duro, R. Dormido, J. Vega, S. Dormido-Canto, G. Farias, J. Sánchez, H. Vargas, A. Murari and JET-EFDA Contributors (2009) Automated recognition system for ELM classification in JET, *Fusion Engineering and Design*, ISSN 0920-3796, Volume 84, Issues 2-6, Pages 712-715.

Abstract

Edge localized modes (ELMs) are instabilities occurring in the edge of H-mode plasmas. Considerable efforts are being devoted to understanding the physics behind this non-linear phenomenon. A first characterization of ELMs is usually their identification as type I or type III. An automated pattern recognition system has been developed in JET for off-line ELM recognition and classification. The empirical method presented in this paper analyzes each individual ELM instead of starting from a temporal segment containing many ELM bursts. The ELM recognition and isolation is carried out using

three signals: Dα, line integrated electron density and stored diamagnetic energy. A reduced set of characteristics (such as diamagnetic energy drop, ELM period or Dα shape) has been extracted to build supervised and unsupervised learning systems for classification purposes. The former are based on support vector machines (SVM). The latter have been developed with hierarchical and K-means clustering methods. The success rate of the classification systems is about 98% for a database of almost 300 ELMs.

References

N. Duro et al.(2006); G. Saibene et al. (1999); V. Vapnik (1995); B. MacQueen (1967); S.C. Johnson (1967); JET EDFA (2009).

Impact Factor

Fusion Engineering and Design has an impact factor of 1.49 according to Thomson Reuters Journal Citation Reports (2011).

Fusion Engineering and Design 84 (2009) 712–715

Contents lists available at ScienceDirect

Fusion Engineering and Design

journal homepage: www.elsevier.com/locate/fusengdes

Automated recognition system for ELM classification in JET

N. Duro[b,*], R. Dormido[b], J. Vega[c], S. Dormido-Canto[b], G. Farias[b], J. Sánchez[b], H. Vargas[b], A. Murari[d], JET-EFDA Contributors[a,1]

[a] JET-EFDA, Culham Science Center, OX14 3DB, Abingdon, UK
[b] Dpto. de Informática y Automática - UNED, C/ Juan del Rosal 16, 28040 Madrid, Spain
[c] Asociación EURATOM/CIEMAT para Fusión, Avd. Complutense 22, 28040 Madrid, Spain
[d] Consorzio RFX-Associazione EURATOM ENEA per la Fusione, I-35127 Padua, Italy

ARTICLE INFO

Article history:
Available online 10 January 2009

Keywords:
ELMs classification
Clustering
SVM
JET

ABSTRACT

Edge localized modes (ELMs) are instabilities occurring in the edge of H-mode plasmas. Considerable efforts are being devoted to understanding the physics behind this non-linear phenomenon. A first characterization of ELMs is usually their identification as type I or type III. An automated pattern recognition system has been developed in JET for off-line ELM recognition and classification. The empirical method presented in this paper analyzes each individual ELM instead of starting from a temporal segment containing many ELM bursts. The ELM recognition and isolation is carried out using three signals: Dα, line integrated electron density and stored diamagnetic energy. A reduced set of characteristics (such as diamagnetic energy drop, ELM period or Dα shape) has been extracted to build supervised and unsupervised learning systems for classification purposes. The former are based on support vector machines (SVM). The latter have been developed with hierarchical and *K*-means clustering methods. The success rate of the classification systems is about 98% for a database of almost 300 ELMs.

1. Introduction

Plasma instabilities should be successfully controlled in order to produce fusion energy efficiently and without compromising the material boundary. Edge localised modes (ELMs) are one of these instabilities that are not fully known and further theoretical and experimental analysis are required.

In order to advance in the study of the physics behind ELMs, data-driven methods seem to be powerful techniques to extract knowledge from the experimental signals without assuming any kind of hypothesis. This knowledge could be combined with theoretical models for both exploratory and confirmatory analysis.

This article develops a data-driven approach for the characterization and automatic classification [1] of ELMs as type I or type III [2]. To this end, three steps are accomplished. The first one is to identify, isolate and extract individual ELMs from JET signals (in the present approach, each individual ELM is analysed instead of starting from a temporal segment containing many ELM bursts). Although the physical basis for this assumption is not established, we introduce it as working hypothesis. The second step is a feature extraction process to represent the ELMs with a minimum set of relevant characteristics. In the third step, three classifications methods (supervised and unsupervised) have been applied to classify the ELMs. All computations have been performed by developing several software tools based on MATLAB.

Section 2 describes the identification and feature extraction; Section 3 explains the three classification methods that we have used; Section 4 shows results and, finally, Section 5 presents some conclusions.

2. Identification and feature extraction of ELMs

As it was mentioned above, this article presents a method that analyzes each individual ELM. The ELM recognition and isolation is carried out using three signals: stored diamagnetic energy (corresponding to the JET signal MG3F/WPD), line integrated electron density (JET signal KG1 V/ LID4) and Dα (JET signal S3AD/ AD34). ELMs are recognized by an abrupt change in the diamagnetic energy and a simultaneous drop in the line integrated electron density. As a consequence of the ELM instability, a typical peaked shape appears in the Dα (Fig. 1). At present, the isolation procedure of ELMs is carried out in a manual way.

Once individual ELMs are identified, the feature extraction process must be accomplished. It consists of extracting features or attributes that are of distinctive nature. The set of attributes allow

* Corresponding author at: Dpto. de Informática y Automática - UNED, C/ Juan del Rosal 16, 28040 Madrid, Spain. Tel.: +34 913987169; fax: +34 91 3986697.
E-mail address: nduro@dia.uned.es (N. Duro).
[1] See the Appendix of F. Romanelli et al. Proc. 22nd IAEA Fusion Energy Conference, Geneva, Switzerland, 2008.

0920-3796/$ – see front matter.
doi:10.1016/j.fusengdes.2008.12.003

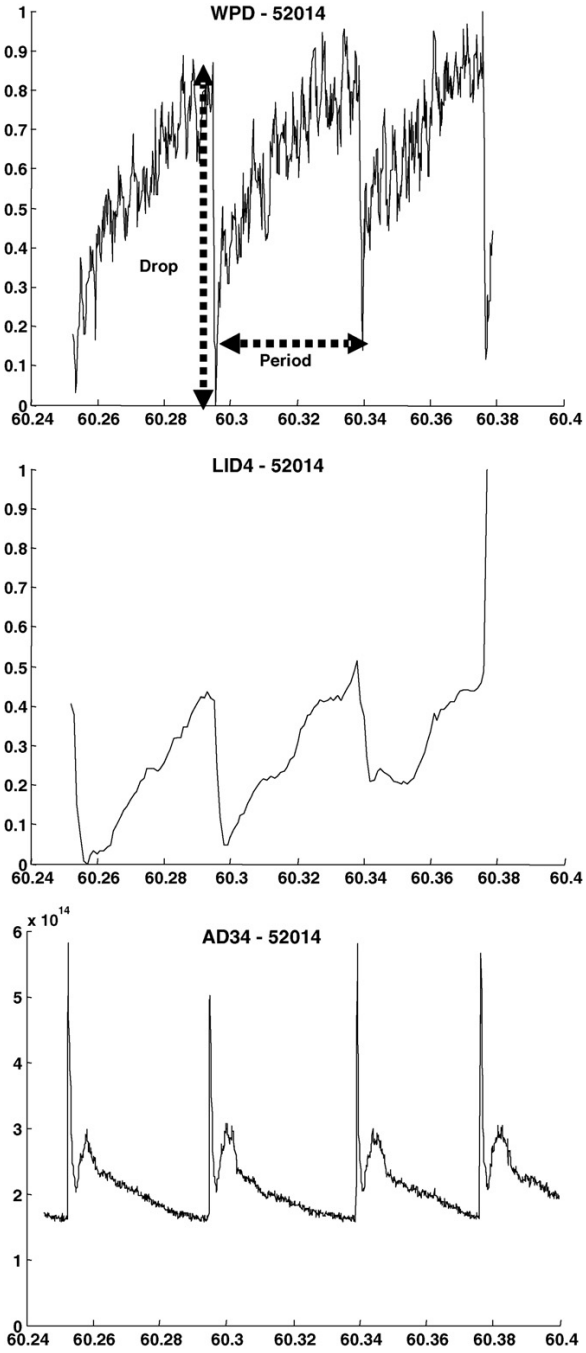


Fig. 1. WPD, LID4 and AD34 signals for shot 52014.

achieving a proper representation of the ELMs that will be used in the classification process.

Typically in classification processes, the selection of both the number of features and the specific ones that best represent the objects is problem dependent. It requires specific knowledge about the system to classify. In practice, a larger than necessary number of feature candidates is generated and then the 'best' of them are adopted.

In the case of ELMs, features have been chosen with the aim of incorporating some knowledge of the experts of ELMs. The diamagnetic energy has been the reference signal and, in particular, the drop in this magnitude has been selected. In the first approximation, type I ELMs seem to show larger drops in the diamagnetic energy than type III ELMs. Another attribute is the ELM period. The repetition frequency is often used as a characterisation parameter of ELM types. Again in the first approximation, high/low rates are assigned to type III/type I ELMs, respectively. Therefore, as individual ELMs are only considered in this article, the ELM period has been selected. Finally, an empirical criterion has been used. A simple visual inspection of the Dα waveform suggests the presence of higher frequency components in this signal for the case of type III ELMs.

Quantifying the latter aspect in a single value has not been a straightforward task and it should be emphasised that eventually it has been determined empirically. Different ways of representing the shape of the Dα emission with a single parameter have been tested (for example standard deviation of the Fourier spectrum, signal power of high frequency components, and crest factor). The best results in terms of success in the type I/type III classification process have been achieved in an empirical way with the so called crest measure (CM) that we have defined as a ratio between the crest factor (CF) and the number of relevant samples (RS) by means of:

$$\mathrm{CM} = \frac{\mathrm{CF} + \text{number of RS}}{2}$$

CF is the crest factor that is defined as the ratio of the peak (crest) value to the root mean square (RMS) value of a waveform.

$$\mathrm{CF} = \frac{\text{peak level}}{\mathrm{RMS}}$$

If the temporal evolution of the Dα signal during the ELM is represented by the samples $\{y_1, y_2, \ldots, y_n\}$, then the peak level = max $\{y_1, y_2, \ldots, y_n\}$ and the root mean square is

$$\mathrm{RMS} = \sqrt{\frac{y_1^2 + \cdots + y_n^2}{n}}.$$

The number of RS is defined as the number of signal samples which exceed a threshold level (Fig. 2). This threshold has been empirically determined as

$$\text{Threshold level} = \frac{(\text{peak level} + \text{final level})}{3}$$

where the final level is the stationary level reached by the Dα signal.

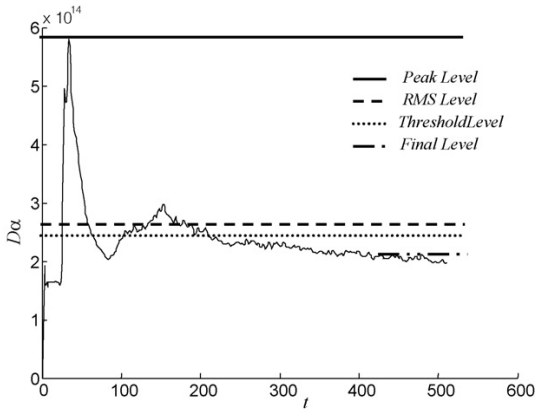


Fig. 2. Crest measure feature.

3. Classification methods

After feature extraction, the classification methods must group the ELMs into two subsets: type I and type III. The creation of a classification system implies the use of training data and test data. Training data are needed to define the classes or categories that define the classification system. Test data are used to estimate the success rate of the classification system developed with the training data. JET signals have been analysed with the different techniques explained below.

3.1. Supervised methods: support vector machines

In supervised methods, both the number of classes (in the present case two classes, type I and type III) and the category for the training samples (each training data belongs to a well-known class: type I or type III) are known. Once the system is trained, it is ready to estimate the category of input data.

Support vector machines (SVM) have been used as supervised method for ELM classification. SVM is a universal constructive learning procedure based on the statistical learning theory [3]. The SVM maps input data into a high-dimensional space using a non-linear function. Once input data are mapped into the high-dimensional space, linear functions with constraints on complexity (i.e., hyper-planes) are used to discriminate the inputs, and a quadratic optimization problem must be solved to determine the parameters of these functions.

3.2. Unsupervised methods

In unsupervised methods, only input data are given to a learning system and there is no notion of the category during learning. The outcomes of unsupervised methods are both the number of classes and the assignment of each training data. Two different techniques have been used with ELMs: *K*-means and hierarchical.

3.2.1. K-means

K-means [4] follows a simple and easy way to classify a given data set through a certain number of clusters (assume *k* clusters) estimated after the training process.

The algorithm is composed of the following steps:

1. Place *K* points into the space represented by the objects that are being clustered. These points represent initial group centroids.
2. Assign each object to the group that has the closest centroid.
3. When all objects have been assigned, recalculate the positions of the *K* centroids.
4. Repeat steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

3.2.2. Hierarchical

Given a set of *N* items to be clustered, and an $N \times N$ distance matrix, the basic steps of hierarchical clustering [5] are:

1. Start by assigning each item to a cluster, so that if you have *N* items, you now have *N* clusters, each containing just one item. Let the distances (similarities) between the clusters be the same as the distance (similarities) between the distances they contain.
2. Find the closets (most similar) pair of clusters and merge them into a single cluster, so that now you have one cluster less.
3. Compute distances (similarities) between the new cluster and each of the old clusters.
4. Repeat steps 2 and 3 until all items are clustered into a single cluster of size *N*.

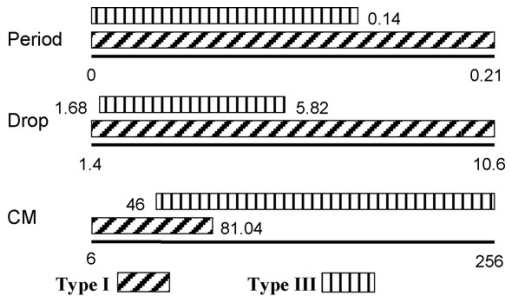


Fig. 3. Variation range of features for the training set.

Of course there is no point in having all the *N* items grouped in a single cluster but, once the complete hierarchical tree has been obtained, *k* clusters can be estimated by cutting the *k*-1 longest links in the tree.

4. Results

Each classification method presented in Section 3 (SVM, *K*-means and hierarchical) has been applied to 265 ELMs isolated from JET signals [6]. A total of 122 ELMs correspond to discharges used as training data (97 of type I and 25 of type III). Fig. 3 summarizes the numerical ranges of each feature for this training set.

The test data consists of 143 ELMs. It is worthwhile to point out that similar results are obtained when using any of the classification methods, either supervised or unsupervised. In both cases the number of classes that provides the vast classification performance is two. Moreover, the success rates with different techniques shown in Table 1 allow concluding that feature selection adopted is quite robust. In particular, using *K*-means and hierarchical, the percentage of signals included in each cluster is the same.

With SVM we obtained one set of 238 ELMs of type I and one set of 27 ELMs of type III. With *K*-means and hierarchical we obtained two set: 237 ELMs of type I and 28 of type III. Only one ELM belonging to the test data, and wrong classified by the unsupervised methods, is moved when using SVM from one class to another.

Fig. 4 shows the results for the *K*-means classification that are very similar to the ones obtained with SVM and hierarchical. It should be noted that the only difference between SVM and *K*-means is one misclassified ELM, in spite of the non-supervised character of the *K*-means method. As it can be observed by inspection of the figure, ELMs of type I have bigger drop than those of type III but this characteristic is not good enough to absolutely differentiate between them. Moreover, the period is also not a deciding feature for the classification process. However classification depends strongly on the crest measure feature. In fact, the same success rate showed in Table 1 is obtained if just the crest measure is the only feature under consideration, *i.e.* no improvement in the classification is obtained by using the other parameters. One parameter, the crest measure, is enough to discriminate ELMs of type I and type III. On the other hand, period and drop cannot be used by themselves to obtain a good classification.

Table 1
Rate of success in the classification.

	Type I	Type III
K-means	98%	93%
Hierachical	98%	93%
SVM	98%	96%

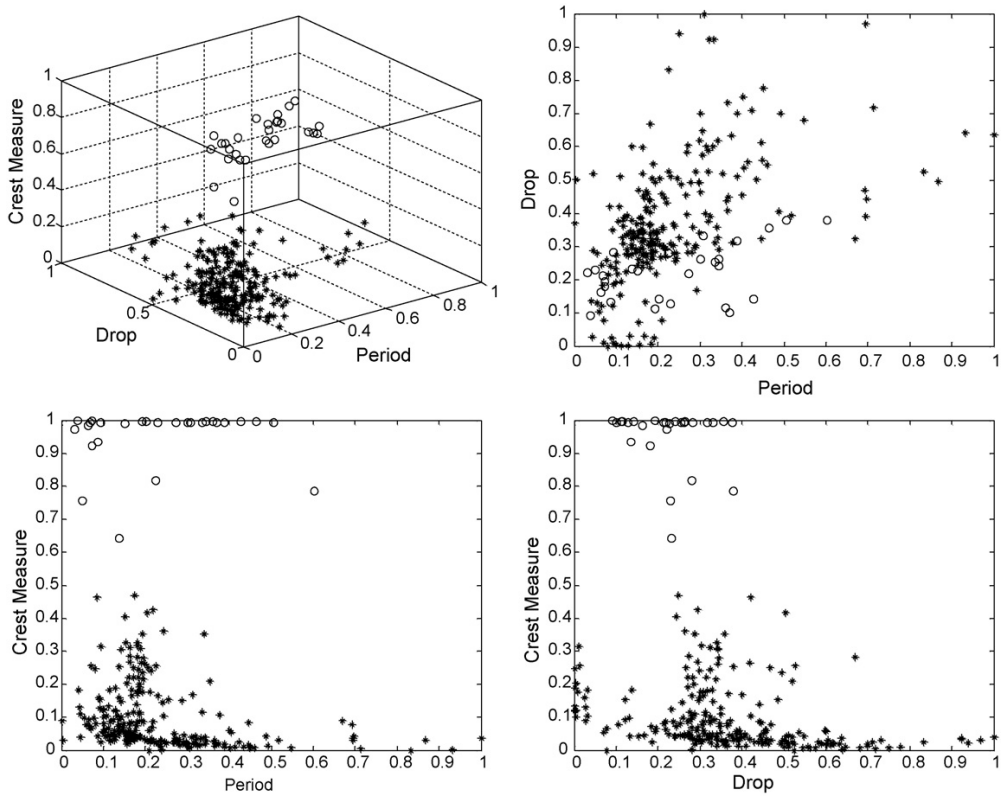


Fig. 4. *K*-means results (type I '*' and type III 'o').

It should be emphasized in the analysis of the results that it is necessary to use the CM instead of the CF in the classification process. The CF is not good enough for the discrimination of the two types of ELMs. This fact denotes that the Number of RS is an important component in the CM feature.

5. Conclusions

In this paper we presented a method for the classification of ELMs based on individual ELM analysis. Although the physical basis to this approach is not established, the method seems to work on a restricted database of 300 ELMs. In addition, the selected dataset has been classified with very high success rates and a very low dimensionality (in fact in can be reduced to a single feature, the crest measure). The several methods used (supervised and unsupervised) group the same signals into the same sets which means that the features selected are robust enough to represent the ELM instability.

On the other hand, it should be noted that the database is not completely general and a comparison with a more extended database is needed in order to draw definitive conclusions on the approach proposed here. More difficult cases could require the development of additional analysis to take into account a wider range of conditions, for example power density and collision regimes.

Acknowledgements

This work, supported by the European Communities under the contract of Association between EURATOM/CIEMAT, was carried out within the framework of the European Fusion Development Agreement. The views and opinions expressed herein do not necessarily reflect those of the European Commission.

The authors wish to thank Prof. Sebastián Dormido Bencomo (UNED) and Prof. Jesús Manuel de la Cruz (UCM) for their constructive comments and invaluable guidance.

The authors wish to thank to E. de la Luna, A. Loarte, I. Nunes and E. R. Solano for their valuable comments.

References

[1] N. Duro, J. Vega, R. Dormido, G. Farias, S. Dormido-Canto, J. Sánchez, M. Santos, G. Pajares, Automated clustering procedure for TJ-II experimental signals, Fus. Eng. Des. 81 (2006) 1987–1991.
[2] G. Saibene, L.D. Horton, R. Sartori, B. Balet, S. Clement, G.D. Conway, et al., The influence of isotope mass, edge magnetic shear and input power on high density ELM and H modes in JET, Nuclear Fusion 39 (9) (1999) 1133–1156.
[3] V. Vapnik, The Nature of Statistical Learning Theory, Springer, 1995.
[4] MacQueen, B. Proc. of 5th Berkeley Symp. Math. Statistics and Probability, Berkeley, Univ. California Press, 1967, 1, pp 281-297.
[5] S.C. Johnson, Hierarchical clustering schemes, Psychometrika 2 (1967) 241–254.
[6] http://users.jet.efda.org/pages/s1-task-force/index.html.

Article 9

Classifier based on support vector machine for JET plasma configurations

9.1 Bibliographic Description

Title

Classifier based on support vector machine for JET plasma configurations.

Citation

S. Dormido-Canto, G. Farias, J. Vega, R. Dormido, J. Sánchez, N. Duro, H. Vargas, A. Murari, and JET-EFDA Contributors (2008) Classifier based on support vector machine for JET plasma configurations, *Review of Scientific Instruments*, ISSN 0034-6748, Volume 79, Pages 10F326-1/10F326-3.

Abstract

The last flux surface can be used to identify the plasma configuration of discharges. For automated recognition of JET configurations, a learning system based on support vector machines has been developed. Each configuration is described by 12 geometrical parameters. A multiclass system has been developed by means of the one-versus-the-rest approach. Results with eight simultaneous classes (plasma configurations) show a

success rate close to 100%.

References

V. Vapnik (2000); R. Duda, P. Hart, D. Store (2001); L. Bottou, O. Chapelle, D. DeCoste, J. Weston (2007);

Impact Factor

Review Of Scientific Instruments has an impact factor of 1.367 according to Thomson Reuters Journal Citation Reports (2011).

REVIEW OF SCIENTIFIC INSTRUMENTS **79**, 10F326 (2008)

Classifier based on support vector machine for JET plasma configurations[a)]

S. Dormido-Canto,[1] G. Farias,[1] J. Vega,[2] R. Dormido,[1] J. Sánchez,[1] N. Duro,[1]
H. Vargas,[1] A. Murari,[3] and JET-EFDA Contributors[4,b),c)]
[1]*Departamento de Informática y Automática, UNED, C/Juan del Rosal 16, 5a. 28040 Madrid, Spain*
[2]*Asociación EURATOM/CIEMAT para FUSIÓN, Avda. Complutense 22. 28040 Madrid, Spain*
[3]*Consorzio RFX-Associazione EURATOM ENEA per la Fusione. I-35127 Padua, Italy*
[4]*JET-EFDA, Culham Science Centre, OX14 3DB Abingdon, United Kingdom*

(Presented 14 May 2008; received 7 May 2008; accepted 14 July 2008;
published online 31 October 2008)

The last flux surface can be used to identify the plasma configuration of discharges. For automated recognition of JET configurations, a learning system based on support vector machines has been developed. Each configuration is described by 12 geometrical parameters. A multiclass system has been developed by means of the one-versus-the-rest approach. Results with eight simultaneous classes (plasma configurations) show a success rate close to 100%. © *2008 American Institute of Physics.* [DOI: 10.1063/1.2972023]

I. INTRODUCTION

The shape of the last flux surface is an essential ingredient in the definition of the JET operation scenarios and several ones can be present during a discharge. Some kinds of data analysis are sensitive to the plasma configuration (for example, to the location of the *X*-point and strike points) and, therefore, proper identification (classification) of the plasma configuration is important.

At present, JET configurations are primarily identified by referring to an identifying keyword describing the request made, prior to the pulse, to the plasma control system. This has the disadvantage of being nonspecific, as several different identifiers can refer to the same configuration; cumbersome, as this data cannot be accessed automatically; incomplete, as some discharges are not assigned an identifier; and potentially wrong, as the identifier describes the intended rather than the resulting configuration. These problems motivated the development of an automated classifier.

Developing classifiers is a learning problem. It means that identification of different classes is needed to show the grouping of the data. The clustering can be carried out by exploiting *a priori* known information. This is known as supervised learning. Otherwise, if the data clustering does not use any prior information it is called an unsupervised classification method.

This article describes the development of a classification system for JET plasma configurations. It is a supervised system based on support vector machines (SVMs).[1] Section II reviews SVM as a technique for classification problems. It should be noted that any particular application of classification with SVM must address three key phases: feature extractions, training, and testing. These three phases for the present classification system are described in Sec. III. Section IV shows results for two different classifiers: one with three classes and one with eight classes systems. Section V is devoted to the final discussion.

II. SUPPORT VECTOR MACHINE FOR CLASSIFICATION

SVM is a universal constructive learning procedure based on statistical learning theory. SVM maps input data into a high-dimensional space using a nonlinear function. Once input data are mapped into the high-dimensional space, linear functions with constraints on complexity (i.e., hyperplanes) are used to discriminate the inputs, and a quadratic optimization problem must be solved to determine the parameters of these functions. Nevertheless for high-dimensional feature spaces, the large number of parameters makes this problem intractable. For this reason, duality theory of optimization is used in SVM to make the estimation of parameters in the high-dimensional feature space computationally affordable. The process time to classify in SVM is fast due to the matrix calculus in the algorithm.[2]

The linear approximation function corresponding to the solution of the dual problem is given in the kernel representation, $K(x,x')$, and it is called the optimal separating hyperplane. $K(x,x')$ represents a dot product of feature vectors in some high-dimensional space.[3] The solution in the kernel representation is written as a weighted sum of the support vectors, that is, a subset of the training data.

III. STAGES IN A CLASSIFICATION SYSTEM

A. Feature extraction

The description of signals in fusion databases is difficult to implement because there is no general solution for extracting generic features. So, feature extractors must be developed to extract the domain specific features most suited to

[a)]Contributed paper, published as part of the Proceedings of the 17th Topical Conference on High-Temperature Plasma Diagnostics, Albuquerque, New Mexico, May 2008.
[b)]For a full listing of names and affiliations of the JET-EFDA Contributors, see A. T. Macrander, Rev. Sci. Instrum. **79**, 10F701 (2008).
[c)]See the Appendix of M.L.Watkins *et al.*, Fusion Energy 2006 (Proceedings of the 21st International Conference, Chengdu, 2006) IAEA, (2006).

TABLE I. Geometrical parameters of the boundary of the last flux plasma surface.

Parameters	Description	Parameters	Description
elon	Elongation boundary	r_{og}	Radial outer gap
r_{geo}	Major radius	r_{ig}	Radial inner gap
r_{xpl}	Radial coordinate lower *X*-point	r_{mag}	Magnetic axis *r* coordinate
z_{xpl}	*z* coordinate lower *X*-point	z_{mag}	Magnetic axis *z* Coordinate
tri_l	Lower triangularity	el_{ax}	Elongation at magnetic axis
tri_u	Upper triangularity	vol_m	Plasma volume

the subsequent classification task. In the present case, the boundary of the last flux surface can be used to identify the plasma configuration of discharges. Therefore, geometrical parameters of the boundary have been chosen as feature vectors (Table I). These parameters were proposed by the experts.

B. Training and testing stages

The process of using data to determine the classifier is referred to as "training" of the classifier. Test data allow performing the "evaluation" of the classification system. Evaluation is important both to measure the performance of the system and to identify the need for improvements in its components (for instance, to add new attributes to the feature vectors or to eliminate redundant ones).

IV. CLASSIFICATION SYSTEMS

Results with two classification systems based on geometrical parameters of the last flux surface and SVM are presented in this section. In order to evaluate the approach, two different classifiers have been applied to some of the configurations stored in the JET database. These configurations belong to one of the following classes: HIXR_GB, StandardFat, HC_SFE_LT, SEPTUM, VLPC_SWEEP, D1F_C_SFE_LT, V_LFE_LT, D1Z_XFORM, and VH_3M5_HT3.

The first classifier has been trained to discriminate discharges belonging to the following three classes: VH_3M5_HT, HIXR_GB, and SEPTUM. Each configuration is defined by 12 geometrical parameters (Table I). Figure 1 shows that a simple visual inspection of parameters can be enough to solve the easiest cases. For the three above configurations, it is possible to discard parameters that do not contribute to the discrimination and to identify the ones that provide the most discriminative power: r_{ig} and z_{mag}. Therefore, for the three classes problem (VH_3M5_HT, HIXR_GB, and SEPTUM), two-dimensional feature vectors (parameters r_{ig} and z_{mag}) perfectly identify three clusters that correspond to the three different configurations (Fig. 2).

Figure 2 also shows three straight lines separating each class from all the others (linear discriminant functions). In this case, using a SVM classifier, in its simplest linear version, we obtain the hyperplane that maximizes the margin of separation, therefore minimizing the misclassification risk.

The data set for the three class problem is made of 199, 39, and 17 configurations, respectively, for each class. The training set is composed by 60% of the configurations and the testing set by 40% obtained from the JET database. The percentage of success is 100% for all the classes.

For the second classifier the number of magnetic configurations considered is 8. The class with a greater number of cases (VH_3M5_HT3) has been discarded to build a more

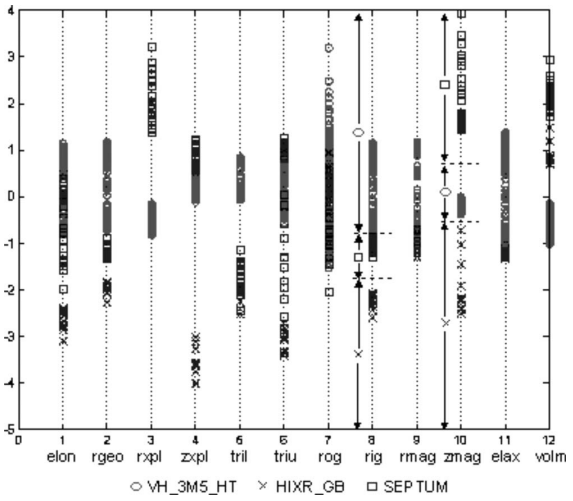


FIG. 1. Geometrical parameters of the analyzed data set for three plasma configurations.

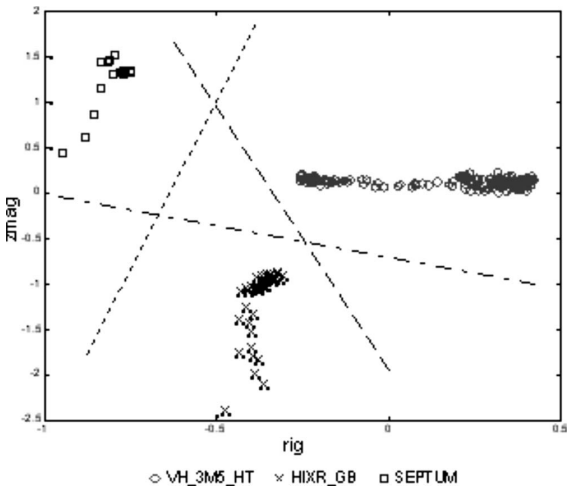


FIG. 2. Three classes described by two-dimensional feature vector (z_{mag} and r_{ig}).

TABLE II. Results for the second classifier.

	Kernel		
	Linear	Radial basis $\sigma=100$	Exponential radial basis $\sigma=100$
HIXR_GB	96.6	95.3	96
StandardFat	100	100	100
HC_SFE_LT	94	95.5	92.2
SEPTUM	95	90	95
VLPC_SWEEP	100	100	100
D1F_C_SFE_LT	100	100	100
V_LFE_LT	98	96.2	98.7
D1Z_XFORM	87.5	90	87.5

uniform data set. The data set is composed of 39, 9, 24, 17, 45, 24, 40, and 12 configurations for each class. In this case, we use as feature vectors all the parameters described in Table I. The percentage of correct classifications is illustrated in Table II for three different kernels.

V. DISCUSSION

SVM is a very competitive method to classify JET configurations. Although apparently a simple visual inspection could be enough to discriminate a limited number of different discharges, when a bigger number of categories are considered, it is necessary to resort to a general purpose system as SVM. High success rate in spite of the reduced number of training data should be emphasized. Results with eight classes are promising even for a future real-time application of the method.

ACKNOWLEDGMENTS

This work, supported by the European Communities under the contract of Association between EURATOM/CIEMAT, was carried out within the framework of the European Fusion Development Agreement. The views and opinions expressed herein do not necessarily reflect those of the European Commission.

[1] V. Vapnik, *The Nature of Statistical Learning Theory*, 2nd ed. (Springer, New York, 2000).
[2] R. Duda, P. Hart, and D. Store, *Pattern Classification*, 2nd ed. (Wiley, New York, 2001).
[3] L. Bottou, O. Chapelle, D. DeCoste, and J. Weston, *Large_Scale Kernel Machines* (MIT, Cambridge, MA, 2007).

Article 10

Structural pattern recognition methods for fusion databases

10.1 Bibliographic Description

Title

Structural pattern recognition methods based on string comparison for fusion databases.

Citation

S. Dormido-Canto, G. Farias, R. Dormido, J. Sánchez, N. Duro, H. Vargas, J. Vega, G. Ratta, A. Pereira, A. Portas (2008) Structural pattern recognition methods based on string comparison for fusion database, *Fusion Engineering and Design*, ISSN 0920-3796, Volume 83, Issue 2-3, Pages 421-424. Ed. Elsevier.

Abstract

Databases for fusion experiments are designed to store several million waveforms. Temporal evolution signals show the same patterns under the same plasma conditions and, therefore, pattern recognition techniques allow the identification of similar plasma behaviours. This article is focused on the comparison of structural pattern recognition methods. A pattern can be composed of simpler sub-patterns, where the most elementary sub-patterns are known as primitives. Selection of primitives is an essential issue in

structural pattern recognition methods, because they determine what types of structural components can be constructed. However, it should be noted that there is not a general solution to extract structural features (primitives) from data. So, four different ways to compute the primitives of plasma waveforms are compared: (1) constant length primitives, (2) adaptive length primitives, (3) concavity method and (4) concavity method for noisy signals. Each method defines a code alphabet and, in this way, the pattern recognition problem is carried out via string comparisons. Results of the four methods with the TJ-II stellarator databases will be discussed.

References

C.S. Daw, C.E.A. Finney, E.R. Tracy (2003); Y.-W. Huang, P.S. Yu (1999); H. Nakanishi, T. Hotchin, M. Kojima (2004); S. Dormido-Canto et al. (2004); G. Farias et al. (2006); S. Dormido-Canto (2006); N. Wirth (1985); The MathWorks Inc. (2006).

Impact Factor

Fusion Engineering and Design has an impact factor of 1.49 according to Thomson Reuters Journal Citation Reports (2011).

Available online at www.sciencedirect.com
ScienceDirect
Fusion Engineering and Design 83 (2008) 421–424
Fusion Engineering and Design
www.elsevier.com/locate/fusengdes

Structural pattern recognition methods based on string comparison for fusion databases

S. Dormido-Canto [a,*], G. Farias [a], R. Dormido [a], J. Vega [b], J. Sánchez [a], N. Duro [a], H. Vargas [a], G. Rattá [b], A. Pereira [b], A. Portas [b]

[a] *Dpto. Informática y Automática - UNED 28040, Madrid, Spain*
[b] *Asociación EURATOM/CIEMAT para Fusión, 28040, Madrid, Spain*

Available online 4 March 2008

Abstract

Databases for fusion experiments are designed to store several million waveforms. Temporal evolution signals show the same patterns under the same plasma conditions and, therefore, pattern recognition techniques allow the identification of similar plasma behaviours. This article is focused on the comparison of structural pattern recognition methods. A pattern can be composed of simpler sub-patterns, where the most elementary sub-patterns are known as primitives. Selection of primitives is an essential issue in structural pattern recognition methods, because they determine what types of structural components can be constructed. However, it should be noted that there is not a general solution to extract structural features (primitives) from data.

So, four different ways to compute the primitives of plasma waveforms are compared: (1) constant length primitives, (2) adaptive length primitives, (3) concavity method and (4) concavity method for noisy signals. Each method defines a code alphabet and, in this way, the pattern recognition problem is carried out via string comparisons. Results of the four methods with the TJ-II stellarator databases will be discussed.

Keywords: Structural pattern recognition; Data mining; Nuclear fusion; Information retrieval

1. Introduction

Identification problems involving time-series data (or waveforms) constitute a subset of pattern recognition applications that is of particular interest because of the large number of domains that involve such data (for instance, fusion databases). There are some previous works on pattern recognition in fusion databases. Refs. [1,2] are focused on general data mining methods devoted to analysing time series data. However, the goal of our approach is not knowledge extraction but to provide users with an easy tool to perform a first data screening. In this sense, earlier approaches concentrated the efforts in looking for similar full waveforms, i.e. signals covering the full plasma life [3–5]. In another approach, the interest is focused on searching for specific patterns within waveforms [6].

The algorithms used in pattern recognition systems are commonly divided into two tasks, as shown in Fig. 1. The description task transforms data collected from the environment into features (primitives). The classification task arrives at an identification of patterns based on the features provided by the description task.

There is no general solution for extracting structural features from data. The selection of primitives by which the patterns of interest are going to be described depends upon the type of data and the associated application. The features are generally designed making use of the experience and intuition of the designer.

This article summarizes different structural pattern recognition methods and shows specific examples in waveforms of the database of TJ-II stellarator. The difference among these methods is the way in which the primitives are computed. Section 2 describes the main concepts to consider in each technique. Section 3 emphasizes in the application scheme. Finally in Section 4 illustrative examples are shown.

2. Computation of primitives

The selection of primitives is an essential issue because they determine what types of structural components can be constructed.

* Corresponding author. Tel.: +34 91 3987194; fax: +34 91 3987690.
E-mail address: sebas@dia.uned.es (S. Dormido-Canto).

0920-3796/$ – see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.fusengdes.2007.11.009

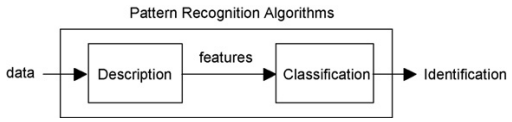


Fig. 1. Tasks in the pattern recognition systems.

We define four ways to compute the primitives of the waveforms: constant length primitives (CLP) [6], adaptive length primitives (ALP), concavity method (CM) and concavity method for noisy signals (CMNS). Each method defines a code alphabet and, in this way, the pattern recognition problem is carried out via string comparisons.

2.1. Constant length primitives

In this method we divide the original signal into segments (all the segments have the same number of samples) where each segment is represented by a straight line. A least square minimization procedure is used to obtain each straight line. Then we encode these segments into a string of primitives. We give a label to each segment and we add the amplitude between the first and the last sample into the primitive. The labelling of the segment $\{(x_i, y_i), (x_j, y_j)\}$ is based on the classification of the slope of the fitted straight line. Our discriminate values, the primitive labels and an illustrative example are depicted in Fig. 2. The classification of the angle gives us all the elementary structural information needed to construct more complex sub-patterns in waveform recognition.

We use five different values (a, c, e, d and z) to represent the classes of the angle. With a larger set of primitive classes we could have expressed more accurately the structure of signal, but the final string would have been more complex. On the other hand these five codes are just enough for the typical plasma evolution analysis. The code e represents a flat part of a signal, codes c and d represent the ascending and descending angle and codes a and z represent the extremely step slopes.

2.2. Adaptive length primitives

As in the CLP method, the original signal is decomposed into L line segments, however in this case, the length of each segment, K, is variable (i.e. the segments do not have a fixed number of samples). A fixed maximun error, E_{max}, is defined for each line segment. This value is a function of the standard deviation of the entire signal. Hence, the ALP methodology is implemented by following these steps:

1. Start of segment (initially $L = 1$).
2. Set K, the number of signal samples to regress (initially $K = 3$).
3. Generate a line regressing.
4. If (fitted error $< E_{max}$), increase K and go to 3, otherwise start a new segment, increase L, and go to 2.

After this processing, the primitives are labelled using the slope thresholds defined for the previous method (CLP).

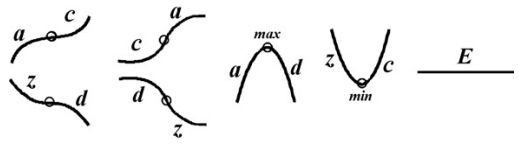


Fig. 3. Stationary points and labels for the classification using concavity methods.

2.3. Concavity method

In this method we detect stationary points of the signal using a simple derivative test. These points belong to one of the six types shown in Fig. 3: maximum, minimum or inflexion point.

Notice that stationary points of inflexion can be easily classified in four types depending on the change in concavity of the curve at that point. Then we give a label to the piece of signal between two consecutive stationary points. The label of each piece of signal $\{(x_i, y_i), (x_j, y_j)\}$ is based on the classification of its concavity. As there are only four possibilities to decide between convex or concave and increase or decrease (shown in Fig. 3); four primitives are sufficient to label any piece of signal. We use four different values (d, z, c, and a) to represent these possible changes of concavities (Fig. 3).

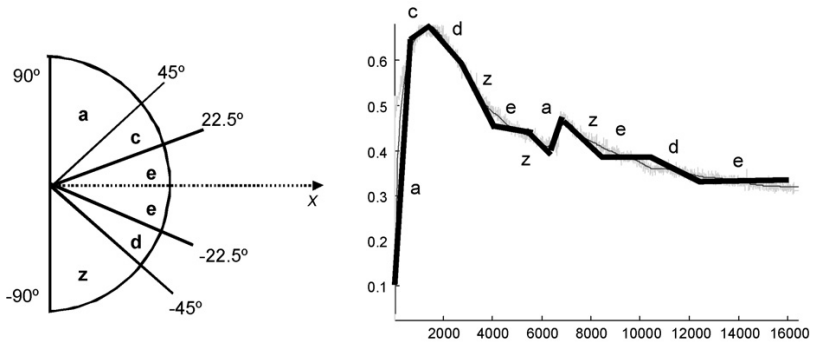


Fig. 2. Discriminates and labels for the classification of the angle of the fitted straight line with an illustrative example.

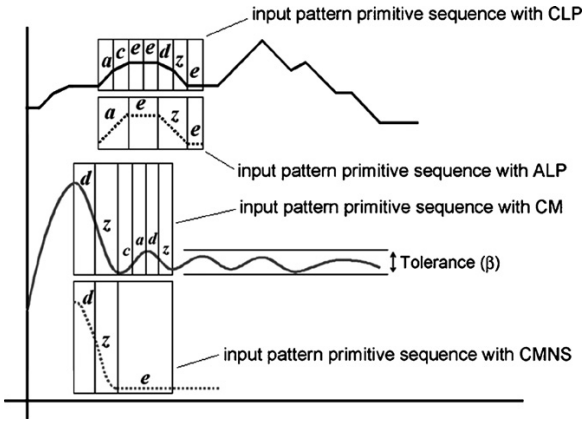


Fig. 4. Primitive sequences in an input pattern with the different methods.

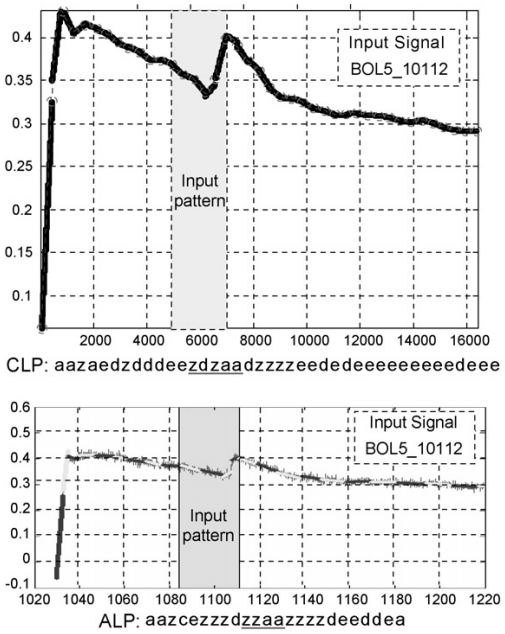


Fig. 5. Primitive sequences with CLP and ALP.

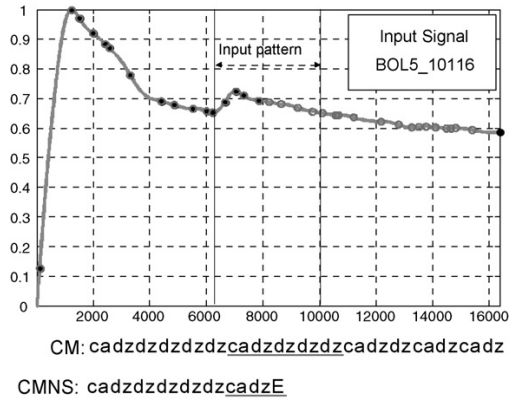


Fig. 7. Primitive sequences with CM and CMNS.

2.4. Concavity method for noisy signals

This method is an extension of the previous method to simplify the codification of noisy signals. The main idea is to add a fifth primitive coded with the label *E* (Fig. 3) to represent small horizontal oscillations of the signal (noise). Our discriminating value to determine if an oscillation is noise or not is a tolerance parameter ($0 \leq \beta \leq 1$). The bigger β the bigger number of horizontal pieces are detected.

Fig. 4 shows a generic example of the primitive sequences using the different methods. It should be emphasised that the code sequence in the input pattern varies according to the used method.

3. Application scheme

The searching patterns procedure can be implemented with a relational database management system using Structure Query Language (SQL). We will obtain from the database the patterns whose sequences of primitives match exactly with the input pattern sequence. In this work, we have used Microsoft Access™ because it is easy to build a database and test our approaches. Knuth–Morris-Pratt (KMP) algorithm is used to solve the string matching problem [7].

Preprocessing and primitive computing of waveforms were done by means of Matlab™. The Matlab Database Toolbox [8] was used for the link between Matlab and Access.

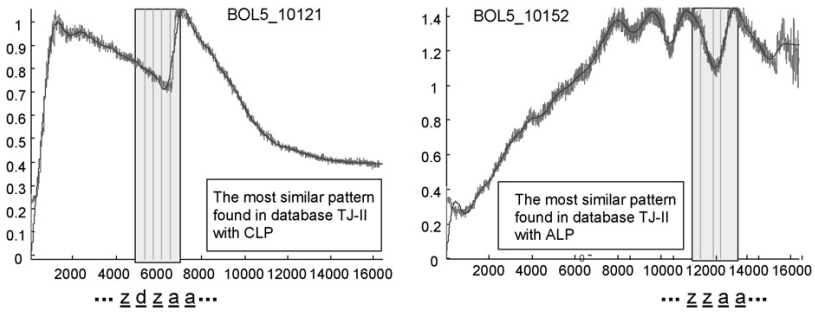


Fig. 6. Output patterns using CLP and ALP.

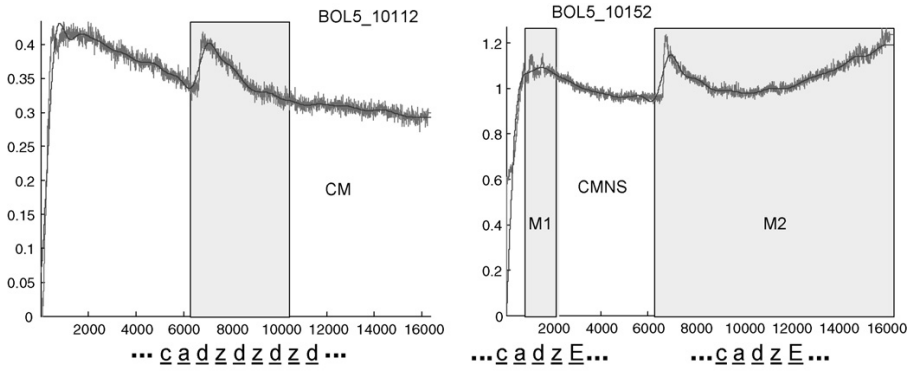


Fig. 8. Output patterns using CM and CMNS.

The application algorithm is the following: first a shot is selected and then a section of the signal (pattern) plus a method to code the pattern are chosen. The Matlab application carries out the pre-processing and primitive computation. After that, an SQL query is made and Access sends back to Matlab the SQL results and the application shows all matches in the returned signals.

4. Illustrative examples

Method comparisons were performed with the TJ-II stellarator database. Once all the matches are found, it is necessary to identify how similar two patterns are (input pattern and database pattern) by using a similarity measure. This requires the introduction of a distance (in the mathematical sense) to be used as a proximity measure.

CLP and ALP methods are tested by using as input signal a bolometer waveform (BOL5) in shot 10112. Fig. 5 shows the codification of the signal in both methods, the input pattern sequenced is underlined. Fig. 6 displays the output patterns. In case of CM and CMNS methods BOL5_10116 has been considered as input signal. Figs. 7 and 8 show the results with these methods. Notice that in Fig. 8 using CMNS two different patterns have been found in the same signal being the most similar the first one (M1).

5. Conclusion

Structural pattern recognition techniques are an efficient way to implement a pattern oriented data retrieval paradigm. Each method has some advantages over the others. For instance the main advantage of CM and CMNS is that the structure of the signal is described in an accurate way. The codification of the signal does not introduce any approximation as in CLP and ALP methods. But CM and CMNS need a greater computation time than CLP and ALP. We have tested all the methods with several waveforms. The search of patterns is accomplished in an efficient way.

References

[1] C.S. Daw, C.E.A. Finney, E.R. Tracy, Rev. Sci. Instrum. 74 (2003) 915–930.
[2] Y.-W. Huang, P.S. Yu, Proceedings of the 5th ACM SIGKDD international conference on Knowledge discovery and data mining, 1999, pp. 282–286.
[3] H. Nakanishi, T. Hotchin, M. Kojima, LABCOM Group, Fusion Eng. Des. 71 (2004) 189–193.
[4] S. Dormido-Canto, G. Farias, R. Dormido, J. Vega, J. Sánchez, M. Santos, TJ-II Team, Rev. Sci. Instrum. 75 (10) (2004) 4254–4257.
[5] G. Farias, S. Dormido-Canto, J. Vega, J. Sánchez, N. Duro, R. Dormido, et al., Fusion Eng. Des. 81 (2006) 1993–1997.
[6] S. Dormido-Canto, G. Farias, J. Vega, R. Dormido, J. Sánchez, N. Duro, et al., Rev. Sci. Instrum. 77 (10) (2006) F514.
[7] N. Wirth, Algorithms and Data Structures, Prentice Hall, 1985.
[8] The MathWorks Inc., Database Toolbox for use with MATLAB, User's Guide, Version 3, 1998–2006.

Article 11

First applications of structural pattern recognition methods at JET

11.1 Bibliographic Description

Title

First applications of structural pattern recognition methods to the investigation of specific physical phenomena at JET.

Citation

G. Rattá, J. Vega, A. Pereira, A. Portas, E. De la Luna, S. Dormido-Canto, G. Farias, R. Dormido, J. Sánchez, N. Duro, H. Vargas, M. Santos, G. Pajares, A. Murari, and JET EFDA Contributors (2008) First applications of structural pattern recognition methods to the investigation of specific physical phenomena at JET, *Fusion Engineering and Design*, ISSN 0920-3796, Volume 83, Issue 2-3, Pages 467-470. Ed. Elsevier.

Abstract

Structural pattern recognition techniques allow the identification of plasma behaviours. Physical properties are encoded in the morphological structure of signals. Intelligent access methods have been applied to JET databases to retrieve data according to physical

criteria. On the one hand, the structural form of signals has been used to develop general purpose data retrieval systems to search for both similar entire waveforms and similar structural shapes inside waveforms. On the other hand, domain dependent knowledge was added to the structural information of signals to create particular data retrieval methods for specific physical phenomena. The inclusion of explicit knowledge assists in data analysis. The latter has been applied in JET to look for first, cut-offs in ECE heterodyne radiometer signals and, second, L-H transitions.

References

J. Vega, JET EFDA (2008); J. Vega et al. (2008); S. Dormido-Canto et al. (2006); E. de la Luna et al. (2004); A. Murari (2006); S. Dormido-Canto et al. (2008).

Impact Factor

Fusion Engineering and Design has an impact factor of 1.49 according to Thomson Reuters Journal Citation Reports (2011).

Available online at www.sciencedirect.com
ScienceDirect
Fusion Engineering and Design 83 (2008) 467–470
Fusion Engineering and Design
www.elsevier.com/locate/fusengdes

First applications of structural pattern recognition methods to the investigation of specific physical phenomena at JET

G.A. Rattá [a,*], J. Vega [a], A. Pereira [a], A. Portas [a], E. de la Luna [a], S. Dormido-Canto [b], G. Farias [b], R. Dormido [b], J. Sánchez [b], N. Duro [b], H. Vargas [b], M. Santos [c], G. Pajares [c], A. Murari [d],

JET-EFDA Contributors

[a] *Asociación EURATOM/CIEMAT para Fusión, Spain*
[b] *Dpto. Informática y Automática—UNED, 28040 Madrid, Spain*
[c] *Dpto. Arquitectura de Computadores y Automática—UCM, 28040 Madrid, Spain*
[d] *Consorzio RFX—Associazione EURATOM ENEA per la Fusione, Padua, Italy*

Available online 24 October 2007

Abstract

Structural pattern recognition techniques allow the identification of plasma behaviours. Physical properties are encoded in the morphological structure of signals. Intelligent access methods have been applied to JET databases to retrieve data according to physical criteria. On the one hand, the structural form of signals has been used to develop general purpose data retrieval systems to search for both similar entire waveforms and similar structural shapes inside waveforms. On the other hand, domain dependent knowledge was added to the structural information of signals to create particular data retrieval methods for specific physical phenomena. The inclusion of explicit knowledge assists in data analysis. The latter has been applied in JET to look for first, cut-offs in ECE heterodyne radiometer signals and, second, L-H transitions.

Keywords: Structural pattern recognition methods; Specific physical phenomena identification; JET

1. Introduction

At present, thousands of signals are acquired in a JET discharge with a data storage that can be over 10GB. Up to now, data has been exclusively indexed according to shot number. However, a new data retrieval model has been proposed to search for data [1]. It is based on the fact that similar signals represent similar plasma behaviour. Therefore, structural pattern recognition techniques can be used to query databases. This avoids the manual data inspection to look for particular structural shapes and allows high-level queries based on physical criteria instead of shot number.

General purpose methods to search for similar structural shapes have been applied to JET databases. They are founded on structural pattern recognition techniques. Feature extraction is accomplished taking into account the morphological structure of waveforms. Two different techniques allow data retrieval in an intelligent way. First, a search method to retrieve entire waveforms (Section 2). Second, a search for structural forms within signals (Section 3). These techniques have also been applied to study particular physical phenomena as described in Section 4. Specific system knowledge is included in the feature extraction process to optimize searching criteria for the physical properties of the phenomenon. The application of the techniques to ECE signals and L-H transitions must be understood as a proof of principle of the method and not as detailed analyses on both processes.

2. Entire waveform search

This approach allows the search of entire waveforms similar to a given one. The application to JET databases has

* Corresponding author.
E-mail address: giuseppe.ratta@ciemat.es (G.A. Rattá).

0920-3796/$ – see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.fusengdes.2007.09.008

been carried out on temporal evolution signals covering the whole discharge, *i.e.* from plasma start to plasma extinction.

Essential elements to develop the searching system are [1]: feature vectors, a classification system and a similarity measure. Feature extraction is performed by means of the "Haar" wavelet transform. A supervised clustering system based on shot length groups the data [2]. This classification avoids traversing the whole database looking for similar signals. Instead, the similarity computation is only carried out within the waveforms of one cluster, the more likely one to contain similar signals. The similarity measure is computed according to the normalized inner product of the feature vectors [2].

A software application implementing this data retrieval technique is available in a concurrent way on the JET analysis cluster (JAC) computer environment.

3. Pattern search within signals

This approach allows the search of structural shapes (patterns) inside time-series data. Patterns are composed of simpler sub-patterns. The most elementary ones are known as primitives. Primitives are represented by characters, converting the pattern recognition problem into a string matching problem.

Feature extraction is carried out by dividing the initial waveform into segments and each segment is fitted with a straight line through a least squares minimization process [3]. The segments are encoded according to the slopes of the straight lines. Only a reduced set of five slopes (primitives) is enough to encode the waveforms. The search of patterns is accomplished by means of a relational database which is an optimal system to handle strings of characters.

Two different techniques have been developed to define segments [6]. The first one assumes segments with the same number of samples: equal length segments (ELS). The second technique uses segments with a variable number of samples: variable length segments (VLS).

This searching method has been put into operation on the JET JAC cluster and can be executed in a concurrent way by simultaneous users.

4. Searching for specific physical phenomena

The recognition of structural shapes plays a central role in distinguishing particular physical properties. Sometimes just one structural form (a bump, an abrupt peak or a sinusoidal component), is enough to identify a specific phenomenon. In other occasions it is necessary to check the coincidence of multiple events in more than one waveform, to detect plasma behaviours.

There is not a general rule to describe the structure – or structure combinations – of various phenomena, so specific knowledge about their characteristics has to be taken into account. In other words, signal structural shape may be not enough for a complete description of physical properties. Therefore, domain knowledge has to be added to the structural information provided by ELS or VLS techniques.

In this first approach, we have applied the mentioned techniques to identify ECE cut-offs in temperature signals. It is as an example of physical phenomena recognition by just a single pattern in temperature waveforms. As an application of multiple pattern recognition we applied the searching method to identify regime transitions, combining the analysis of density and Dα signals.

4.1. ECE cut-offs

A temperature diagnostic in JET is the ECE heterodyne radiometer [4]. One of the limitations of any ECE system in high-density plasmas is the appearance of cut-off (internal reflection of the ECE radiation). When the density between the ECE antenna (located in the low field side of the plasma) and the emitting region rises above the cut-off density, the radiation is unable to propagate out to the detecting system. When that happens, a sharp reduction of the signal in the central channels is observed, keeping its temporal response low and steady. At a later time in the discharge, when the density diminishes below the cut-off value, the signal reaches the proper level with a sudden upward

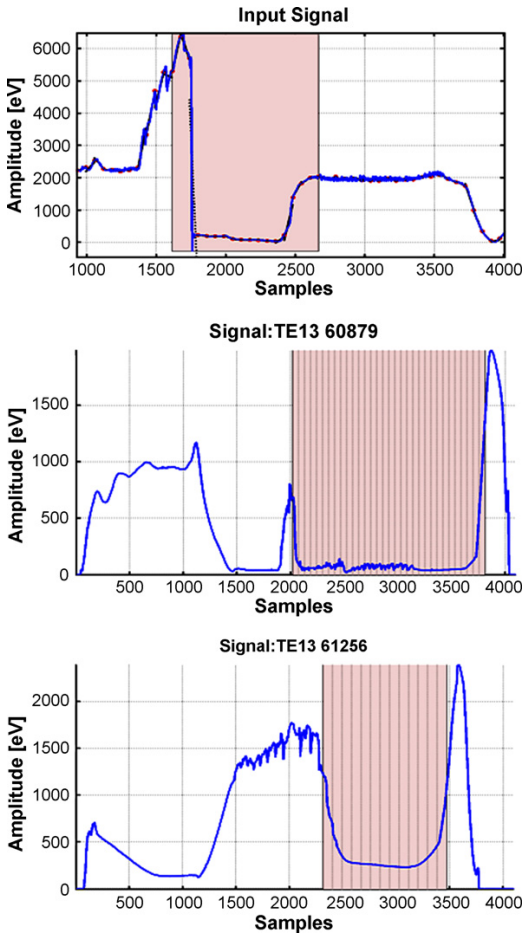


Fig. 1. Input waveform (above), match (middle) and mismatch (below).

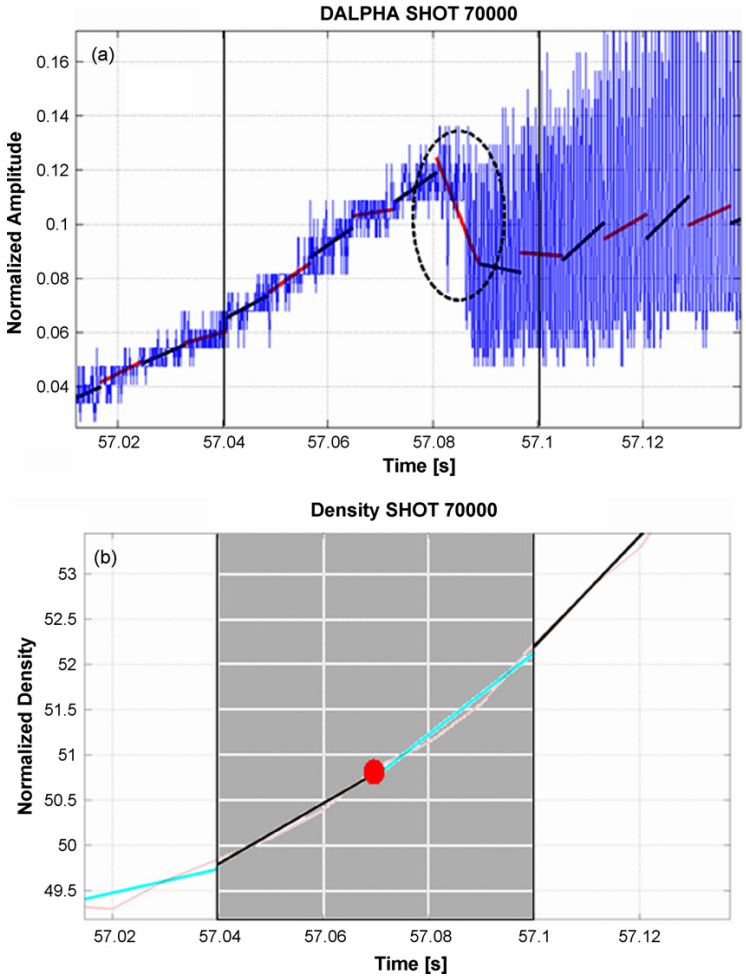


Fig. 2. L-H transition recognition. (a) In a Dα signal and (b) density signal.

transition. The time interval of the cut-offs varies differently in each shot, so the pattern to search must be "flexible" in this respect.

A large set of JET signals (from shot 60,628 to 68,749) was used to create the test database. After the ELS computation of the primitives, the searching procedure was slightly changed: the database must return all the strings that show a quick drop, subsequently a "flexible" time of flat slope (channel in cut-off) and finally a high positive slope. As outputs the starting and ending times of the cut-off and their corresponding densities are estimated for each shot.

About 92% of the signals presenting this behaviour were detected. However such pattern does not correspond univocally to ECE cut-offs because it could be produced by abrupt changes in the plasma temperature. So, some of those similar shapes were also recognized as due to the onset of cut-offs and that is the reason of the mismatch (bottom image of Fig. 1). Also, it is possible to observe in Fig. 1 that the cut-off time length in the input signal is different to the retrieved ones.

4.2. Regime transitions

Time instants for L-H transitions can be identified in an automated way [5]. However, structural pattern recognition techniques can also be used (as a first approach) easily.

The L-H transition is determined by a rapid drop in the Dα signal. However, this behaviour is difficult to identify because its amplitude changes considerably signal by signal. Also, the waveform is corrupted with noise and it presents sudden amplitude variations; so, if it is independently analyzed, there remains the risk of missing the real transition time. Consequently, the analysis of these signals needs extra accuracy, achieved using the VLS method. An example of the codification is depicted in Fig. 2a. Notice that when the transition takes place approximately at 57.08 s, there is a high negative slope and the length of the primitive is bigger than the average. The searching criterion, defined with this additional knowledge, was the following: the first of the primitives with a high negative slope and with a length at least two times bigger to the average would mark the estimated transition instant.

As it was mentioned before, it is necessary to look simultaneously for other patterns in order to minimise the errors.

A sudden increase in the plasma density could evidence a confinement improvement, so we studied this waveform to confirm the regime change. However, a density augment can be also a consequence of other factors (neutral beam injections, gas injection or particle recycling). Thus, the precise selection of the encoding angles was essential to identify only the required phenomena. Once again, to improve the accuracy and "tune" the search, the VLS computation of primitives was selected. The transition time was determined by the first primitive with a high positive gradient, and the error bars were set to the temporal length of its adjacent primitives (Fig. 2b).

In summary, the recognition of the L-H transition is accomplished by the combination of two patterns: a drop in the Dα and a simultaneous increment of the density.

The technique was applied to 50 shots picked from JET database. When the identified instant in the Dα signal (57.08 s in the example) was inside the error bars (grey highlighted in Fig. 2b) an L-H transition was recognized, setting this Dα instant as the estimated transition time. That happens in the 76% of the cases.

Comparing them to the real L-H mode times, the average error was 34 ms, with a maximum error of 89 ms and a standard deviation of 19.2 ms.

5. Discussion

In the present article the results of structural pattern recognition methods applied to JET signals are shown. General techniques (entire waveforms identification and structural recognition within signals) are already available on the JAC Linux Clusters of JET.

To recognize particular physical phenomena, the specific knowledge about the behaviour of the waveforms involved in each case was exploited, guiding the searching methodologies.

These *ad hoc* modified methods were applied, as a proof of principle, to analyze two specific phenomena in JET plasmas. Firstly, we tuned the technique to identify ECE cut-offs. Secondly, we determined the L-H transition time by means of multiple patterns searching, combining the structural shapes of density and Dα waveforms.

References

[1] J. Vega and JET EFDA Contributors, Intelligent methods for data retrieval in fusion databases, these proceedings, in press.
[2] J. Vega, A. Pereira, A. Portas, S. Dormido, G. Farias, R. Dormido, et al., Data mining technique for fast retrieval of similar waveforms in Fusion massive databases, Fusion Eng. Des, submitted for publication.
[3] S. Dormido-Canto, G. Farias, J. Vega, R. Dormido, J. Sánchez, M. Santos, et al., Search and retrieval of plasma waveforms: structural pattern recognition approach, Rev. Sci. Instrum. 77 (2006) 10F514.
[4] E. de la Luna, G. Conway, J. Fessey, R. Prentice, D.V. Bartlett, J.M. Chareau, et al., Electron cyclotron emission radiometer upgrade on the JET tokamak, Rev. Sci. Instrum. 75 (10) (2004) 3831–3833.
[5] A. Murari, G. Vagliasindi, M.K. Zedda, R. Felton, C. Sammon, L. Fortuna, et al., Fuzzy logic and SVM approaches to regime identification in JET, IEEE Trans. Plasma Sci. 34 (3) (2006).
[6] S. Dormido-Canto, G. Farias, R. Dormido, J. Sánchez, N. Duro, H. Vargas, et al., Comparison of structural pattern recognition methods for fusion databases, these proceedings, submitted for publication.

Article 12

Data mining technique for fast retrieval in fusion massive databases

12.1 Bibliographic Description

Title

Data mining technique for fast retrieval of similar waveforms in fusion massive databases.

Citation

J. Vega, A. Pereira, A. Portas, S. Dormido-Canto, G. Farias, R. Dormido, J. Sánchez, N. Duro, M. Santos, E. Sánchez, G. Pajares (2008) Data mining technique for fast retrieval of similar waveform in Fusion massive databases, *Fusion Engineering and Design*, ISSN 0920-3796, Volume 83, Issue 1, Pages 132-139. Ed. Elsevier.

Abstract

Fusion measurement systems generate similar waveforms for reproducible behavior.A major difficulty related to data analysis is the identification, in a rapid and automated way, of a set of discharges with comparable behaviour, i.e. discharges with "similar" waveforms. Here we introduce a new technique for rapid searching and retrieval of

"similar" signals. The approach consists of building a classification system that avoids traversing the whole database looking for similarities. The classification system diminishes the problem dimensionality (by means of waveform feature extraction) and reduces the searching space to just the most probable "similar" waveforms (clustering techniques). In the searching procedure, the input waveform is classified in any of the existing clusters. Then, a similarity measure is computed between the input signal and all cluster elements in order to identify the most similar waveforms. The inner product of normalized vectors is used as the similarity measure as it allows the searching process to be independent of signal gain and polarity. This development has been applied recently to TJ-II stellarator databases and has been integrated into its remote participation system.

References

C. Alejaldre et al. (1999); J. Vega et al. (2005); J. Vega et al. (2006); R.O. Duda, P.E. Hart, D.G. Stork (2001); V. Cherkassky, F. Mulier (1998); N. Duro et al. (2006); H. Nakanishi, T. Hochin, M.Kojima (2004); H. Nakanishi, T. Hochin, M.Kojima (2006); S. Dormido-Canto et al. (2004); G. Farias et al. (2006); S. Mallat (2001); V. Vapnik (2000); The Mathworks Inc (2007); R. Castro, D.R. López, J. Vega (2006); J.Vega et al. (1996).

Impact Factor

Fusion Engineering and Design has an impact factor of 1.49 according to Thomson Reuters Journal Citation Reports (2011).

Available online at www.sciencedirect.com
ScienceDirect
Fusion Engineering and Design 83 (2008) 132–139
Fusion Engineering and Design
www.elsevier.com/locate/fusengdes

Data mining technique for fast retrieval of similar waveforms in Fusion massive databases

J. Vega [a,*], A. Pereira [a], A. Portas [a], S. Dormido-Canto [b], G. Farias [b], R. Dormido [b], J. Sánchez [b], N. Duro [b], M. Santos [c], E. Sánchez [a], G. Pajares [c]

[a] *Asociación EURATOM/CIEMAT Para Fusión, Madrid, Spain*
[b] *Departamento de Informática y Automática, UNED, Madrid, Spain*
[c] *Departamento de Arquitectura de Computadores y Automática, UCM, Madrid, Spain*

Received 7 September 2006; received in revised form 9 July 2007; accepted 22 September 2007
Available online 31 October 2007

Abstract

Fusion measurement systems generate similar waveforms for reproducible behavior. A major difficulty related to data analysis is the identification, in a rapid and automated way, of a set of discharges with comparable behaviour, i.e. discharges with "similar" waveforms. Here we introduce a new technique for rapid searching and retrieval of "similar" signals. The approach consists of building a classification system that avoids traversing the whole database looking for similarities. The classification system diminishes the problem dimensionality (by means of waveform feature extraction) and reduces the searching space to just the most probable "similar" waveforms (clustering techniques). In the searching procedure, the input waveform is classified in any of the existing clusters. Then, a similarity measure is computed between the input signal and all cluster elements in order to identify the most similar waveforms. The inner product of normalized vectors is used as the similarity measure as it allows the searching process to be independent of signal gain and polarity. This development has been applied recently to TJ-II stellarator databases and has been integrated into its remote participation system.

Keywords: Fusion databases; Similar waveforms; Pattern recognition; Data mining; TJ-II

1. Introduction

Fusion devices generate very large databases with a million or more signals and tens or hundreds of thousands of samples per waveform. Moreover, the database contents are not just related to scientific data but also to technical systems. Diagnostics permit the temporal evolution of plasma properties to be followed and "similar" waveforms are generated for reproducible plasma behavior. Control systems record time-dependent signals and similar waveforms characterize analogous discharges.

In general, data analysis in fusion requires searching for "similar" waveforms: statistical analysis, seeking specific behaviours or reviewing previous results. This means selecting a large enough number of signals from different discharges. Such a selection process is usually a manual and tedious procedure in which the signals need to be examined individually.

To automate the searching process for identifying the waveforms that are most "similar" to a reference one, two aspects must be taken into account. The first aspect is the concept of "similar waveforms" itself. Intuitively, one thinks that two signals are similar when one resembles the other. Typically, the identification of similarity in manual searches is carried out by means of visual data analysis. However, several experimental factors connected with signal conditioning (i.e. amplification gain and/or signal polarity) may hide the analogous appearance. Thus, the automation of a searching process implies the definition of a similarity criterion, which requires the introduction of a distance (in the mathematical sense) that can be used to compare how similar two waveforms are. Nevertheless, gains and polarities must be borne in mind as issues.

Once a similarity criterion has been established, the second aspect to be considered is the means to reach for the most similar waveforms. A linear approach might be to compute the similarity

* Corresponding author. Tel.: +34 913466474; fax: +34 913466124.
E-mail address: jesus.vega@ciemat.es (J. Vega).

0920-3796/$ – see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.fusengdes.2007.09.011

factor of a given waveform with all database signals and to sort the waveforms according to the similarity value. However, this procedure is unrealistic in very large databases with a lot of samples per waveform. Therefore, it is necessary to develop methods to reduce the searching space to just the most probable waveforms of being similar.

This article describes a new technique by means of which, given a waveform, it is possible to retrieve similar ones from large databases in a fast and automated way. The method needs an input signal from a researcher and then, it looks for the most similar waveforms within the database. The similarity factor is a measure of how similar the waveforms are in relation to the given one. This factor allows establishing an order in the signals to determine the degree of similarity with the initial one. The technique is based on the development of a classification system that groups waveforms into clusters in accordance with certain rules. Waveform clustering is the essential element to speed up the search and to save computational resources. The searching process is carried out by means of a one by one comparison method but only within those waveforms inside a cluster, rather than by computing the similarity between all database signals.

The technique has been applied to a fusion database. The searching pattern is a full waveform, i.e. a waveform that stores the whole plasma evolution during a discharge, from plasma beginning to extinction. A similar waveform recognition system (SWRS) has been developed for the TJ-II device databases. TJ-II is a medium size stellarator (heliac type) [1] located at CIEMAT in Madrid (Spain). It is a four period device whose main parameters are: $B(0) \leq 1.2$ T, $R(0) = 1.5$ m, $\langle a \rangle \leq 0.22$ m. Two gyrotrons (300 kW each, 53.2 GHz, second harmonic, X-mode polarization) and one NBI (300 kW) provide plasma heating. Presently, the SWRS has been integrated into the TJ-II remote participation system (RPS) [2,3].

Section 2 provides an overview of the similar waveform recognition system. Section 3 describes a general model to develop a very flexible classification system. Finally, Section 4 explains a specific implementation of a SWRS for the TJ-II environment.

2. Similar waveform recognition system overview

Fusion devices can collect thousands of waveforms per discharge and hence, the database is made up of thousands of signal collections. A signal collection signifies the complete set of recorded signals for an individual waveform for all discharges. For instance, in a Tokamak, the plasma current collection is a collection made up of all plasma current waveforms. The SWRS has been designed to look for similar waveforms within collections. Of course, the present technique would be applicable to all the waveforms of a database but, in a practical environment, the searching process is restricted to waveforms of the same collection.

The waveforms of every signal collection are classified into a series of categories (or clusters). This classification process tries to achieve a convenient set of groups, with a suitable number of waveforms in each cluster, in order to reduce the searching space when looking for similar waveforms for an input signal. The clustering process begins with a feature generation stage to identify measurable quantities that represent the waveforms with a lower dimensionality. As a result, waveforms are replaced by their feature vectors.

However, creating classifiers involves the use of patterns (from the feature vectors) for learning. Learning refers to some form of algorithm to assign each object to a cluster. There are two common types of learning problems, known as supervised learning and unsupervised learning. Supervised learning is used to estimate an unknown (input, output) mapping from known (input, output) samples. The term ‘supervised’ denotes the fact that output values for training samples are known. In the unsupervised learning scheme, only input samples are given to a learning system, and there is no notion of the output during the learning [4,5].

Signal collections can be very different to each other. Thus, several clustering criteria (supervised and unsupervised) could be necessary for optimum classification of the waveforms in each collection. In general, supervised clustering is related to a classification based on physical properties whereas unsupervised clustering is performed when physical criteria are not apparent. Several clustering procedures for fusion experimental signals are discussed in [6].

After building the classification system from feature vectors, the searching process of most similar signals is carried out in four steps. Given a waveform, the first step performs feature extraction. The second one is the classification of the feature vector into one of the existing clusters. The third step is the computation of the similarity factor between the input feature vector and the rest of the cluster feature vectors. Finally, waveforms are sorted according to the similarity measure in descending order.

Recently, other similar waveform recognition systems have been published [7–10]. The first two are based on Fourier analysis and recognize, on the one hand, slow varying full waveforms and, on the other hand, patterns within waveforms with, at most, one major frequency component. The searching process of similar waveforms is accomplished by means of an “R-tree” multi-dimensional indexing system. The other two published references are based on the discrete wavelet transform (DWT) [11] and the support vector machine (SVM) learning system. SVM is a universal constructive learning procedure based on statistical learning theory [12]. In references [9,10], each full waveform is replaced by its wavelet coefficients computed at some decomposition level. To look for similar full waveforms, the wavelet coefficients of a signal are the input to a SVM based learning system. This system identifies the collection that the waveform belongs to and then, a similarity measure is computed between the input signal and all the waveforms (wavelet coefficients) in the collection (it should be noted that no classification system is considered). Two kinds of similarity factors are proposed: Euclidean distance and bounding envelop methods [10].

3. General purpose classification system model

As was mentioned before, any classification process requires a previous feature extraction from the objects to classify, i.e.

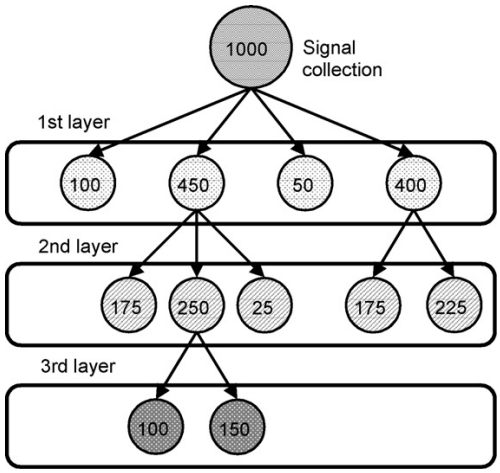


Fig. 1. Classification system model.

to extract a set of characteristics that represent the object main features. This process is essential for data clustering and it allows the dimensionality of the problem to be reduced considerably.

Because the SWRS can manage several signal collections, a classification system is required for each one. Bearing in mind that each collection may consist of thousands of waveforms and new signals can be incorporated as new discharges are produced, the classification system must follow a very flexible scheme to evolve according to dynamic requirements. To this end, a multilayer classification system model is proposed. The first layer is made up of the set of clusters that result after the classification of all waveforms of a collection. Some clusters may contain a high number of waveforms with different patterns and, therefore, they can be sub-classified again. The new clusters form the second layer. The clustering refinement can continue up to reach an optimal classification (Fig. 1).

The main properties of the model are:

- A tree structure is generated for each collection. Each cluster is a *node* of the tree. The cluster at the top is the *root* and the clusters at the bottom are the *leaves*.
- The nodes contain waveforms (characterized by their features).
- Each node represents one category (or class) of the classification process performed with the parent node.
- The union of all child nodes is the parent node.
- Different clustering methods can be applied to the several nodes that form a layer.
- The clustering criterion of any node is different from the clustering criterion of any ancestor node.
- Different branches can have different decomposition layers.
- Horizontal expansion: new kind of signals inside a collection can be added as new clusters at any moment in time.
- Vertical expansion: leave clusters can be split in new nodes without affecting the tree structure.

This model enables fine-tuning of the classification system at any moment. Also, it should be noted that the model allows the use of different clustering criteria (supervised and unsupervised) with the several nodes of the tree structure.

4. SWRS development for the TJ-II database

This section describes the SWRS for the TJ-II environment as well as the TJ-II classification system (feature extraction, clustering method and similarity measure), the client/server architecture for data logging and retrieval, and the integration of the SWRS into the TJ-II remote participation environment.

4.1. TJ-II SWRS: classification system and similarity measure

The main aim of the classification system is to group the waveforms of a collection into several classes in order to reduce the searching space when looking for similar waveforms. The measurements used for classification are known as features. In the more general case, l features, termed x_i, where $i = 1, 2, \ldots, l$, are used and form the *feature vector*

$$x = [x_1, x_2, \ldots, x_l]^{\mathrm{T}}$$

where T denotes transposition. Each feature vector identifies *uniquely* a single object.

Some waveform pre-processing must be performed to build a suitable classification system with the feature vectors. First, it is necessary to bear in mind that searching for similar full waveforms signifies looking for plasmas whose temporal evolution demonstrate similar behaviour. This means that the classification system has to be constructed by referring all signals to the same temporal interval from a single reference (particular event). In medium size devices for example, it is possible to speak about an interval of 400 ms from the plasma start or a 100 ms segment from the beginning of neutral beam injection or a 50 ms interval after an L–H transition. In large devices like JET, time intervals can be several seconds long. In addition, it should be highlighted that different collections can be considered for the same signal. The differences among them are the length of the temporal segments and/or the event that defines the segment start.

The reference time for the TJ-II databases is related to the beginning of the TJ-II discharge. Each discharge has a initial time (τ_0) defined by when the heating starts (ECH or NBI).

In addition, waveform pre-processing is responsible for making the classification system independent of several other factors: signal offset, sampling rates, number of samples or sampling instants.

Waveform pre-processing is made up of three stages:

1. *Offset removal*. This stage sets the waveform line base to the 0 V level. This is accomplished by computing the mean value of all samples up to the time instant τ_0, and then subtracting this mean value from all sample in the waveform. This step is essential for the TJ-II first layer clustering criterion as explained below.
2. *Linear interpolation*. In this phase all waveforms are restricted to an interval of 300 ms and signals are aligned

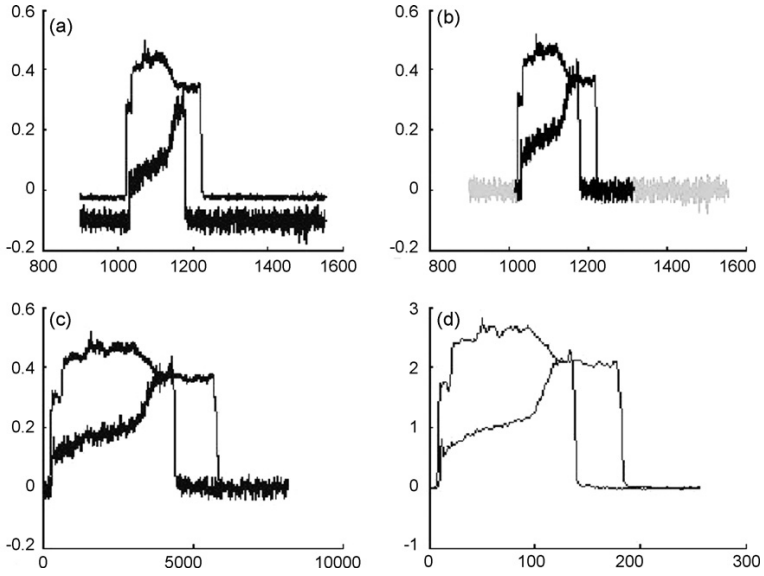


Fig. 2. Waveform pre-processing. (a) Initial waveforms. (b) Offset removal and linear interpolation. (c) Signal alignment for feature extraction. (d) Feature extraction: Haar wavelet coefficients.

with the beginning of the discharges. Linear interpolation of N_I points is carried out between $\tau_0 - 5$ ms and $\tau_0 + 295$ ms, where N_I is the nearest power of 2 (exceeding) to the number of samples in the original waveform in the fore-mentioned interval. The reason for choosing a power of two is related to the application of the wavelet transform for feature extraction.

3. *Feature extraction*. Feature extraction in the TJ-II classification system is achieved by means of the Haar wavelet transform, which has two main advantages. First, it can be computed quickly and easily, and second, the Haar wavelet retains the time and frequency information simultaneously. Feature extraction also allows reducing the problem dimensionality from several tens of thousands of samples to just a few points. Therefore, the waveform obtained in the linear interpolation phase is transformed in accordance with a Haar wavelet transformation. Different decomposition levels can be chosen (1, 2, 3, ...) and a feature vector with a reduced number of characteristics ($N_I/2$, $N_I/4$, $N_I/8$, ..., respectively) is obtained. Analysis with several decomposition levels were carried out. In conclusion, feature vectors with 256 points allow developing classification systems equivalent to ones built with greater number of features.

Waveform pre-processing is summarized in Fig. 2.

To classify the feature vectors, a supervised clustering criterion is used. The criterion is based on computing the number of features required to reach 99.5% of the feature vector Euclidean norm. In other words, a waveform w belongs to cluster K when its feature vector $\boldsymbol{v}_w$ satisfies

$$\sqrt{\sum_{j=1}^{K} \boldsymbol{v}_{w,j}^2} \geq 0.995||\boldsymbol{v}_w||$$

Therefore, there are 256 possible clusters (as much clusters as features) in the first layer of the classification model. In the present system, no additional sub-classifications have been carried out.

From a physical point of view, the above criterion means that each cluster contains discharges with equivalent pulse lengths. This is a direct consequence of removing the signal offset in the pre-processing stage. After finishing a discharge, the signal level comes back to 0 V and, therefore, the main contribution to the feature vector Euclidean norm takes place during plasma life time.

Taking into account that (1) the waveform processing stage handles a temporal segment of 300 ms and (2) the length of the feature vector is 256, then, the pulse length of two signals in adjacent clusters differs, at most, by 1.172 ms. Therefore, cluster K contains discharges whose length, T_K, satisfies

$$(K-1)\Delta T < T_K \leq K\,\Delta T$$

where ΔT is the maximum difference between signals of adjacent clusters.

As mentioned previously, the searching process of similar waveforms is accomplished among waveforms of a single cluster. However, the T_K parameter can be considered small to completely discriminate similar waveforms in adjacent clusters. Hence, for practical purposes, the searching process is not limited to a single cluster, rather to an odd number (N_C) of them. The search is symmetrically distributed around the initial cluster, covering both sides (left and right) with $(N_C - 1)/2$ clusters each.

A similarity measure must be introduced in order to identify signals that are most similar to a given waveform. When the angle between two vectors is a meaningful measure of their similarity, then the normalized inner product may be an appro-

collection	K	feature vectors	time (s)
BOL5	136	75	2.283

Similarity	Shot
1.000000	13774
0.999249	13775
0.997523	13770
0.997363	13764
0.996873	12881
0.996541	13773
0.996502	13760
0.995959	13777
0.995498	13761
0.995408	12957

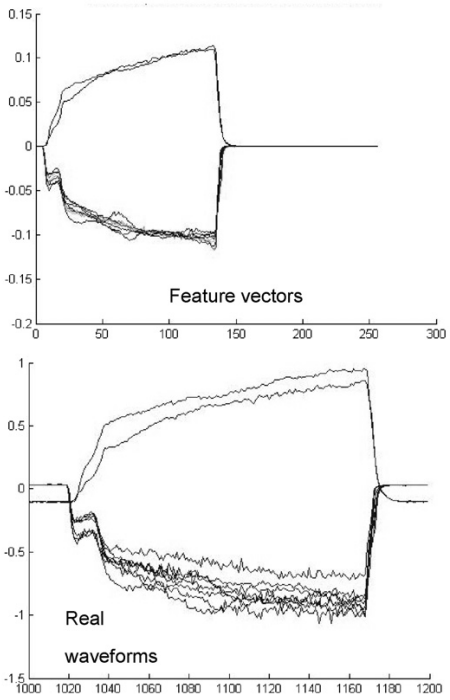


Fig. 3. SWRS results for a bolometry signal.

priate similarity function. The absolute value of this quantity has been chosen as the similarity measure and the vectors are feature vectors ($\boldsymbol{u}_w$ and $\boldsymbol{v}_w$).

$$S_{\boldsymbol{uv}} = |\cos\,\alpha| = \frac{|\boldsymbol{u}_w \cdot \boldsymbol{v}_w|}{||\boldsymbol{u}_w|| \cdot ||\boldsymbol{v}_w||}, \quad 0 \le S_{\boldsymbol{uv}} \le 1$$

The absolute value of the normalized inner product provides two main advantages. The method does not depend on either amplification gains (i.e. waveforms whose difference is a gain factor are recognized as equal signals) or signal polarity (i.e. inverted waveforms are perceived as equal signals).

It should be noted that the present technique retrieves the most similar waveforms to a given one, but it does not imply that the signals are almost equal (similarity near 1). The method finds the most similar signals but the similarity factor can be low (close to 0). In other words, the technique can get a list of signals from the database although the waveforms do not resemble between them. When this happens, the initial signal can be considered as an outlier.

The classification system model and the inner product similarity measure constitute a very powerful recognition system when searching for any kind of waveform. The waveforms, first, are not restricted by signal characteristics (for instance, frequency components) and, second, they do not need to be defined in advance (any kind of signal can be considered a pattern). These facts ensure capabilities to seek for any kind of waveform at any moment. Computational resources are maintained at a minimum. The major requirement is disk storage, although in reality this is not very large.

A first SWRS was developed in a Windows XP Pentium IV computer with the Matlab software package [13]. Single layer classification systems for several collections were created with 846 feature vectors of dimension 256. To test the searching process, one waveform was chosen in a random way and the ten most similar waveforms were retrieved by looking for similarity in nine clusters ($N_C = 9$). The reason of choosing this value is to compute the similarity between signals whose discharge lengths differ at most in 10 ms (about 3% of the temporal segment, which is 300 ms). It should be taken into account that the temporal difference between adjacent clusters is 1.172 ms.

Of course, the method always finds the initial waveform with a similarity factor of 1. Fig. 3 shows results for a collection of bolometry signals. In this case, the 99.5% point of the feature vector norm was reached with 136 coefficients ($K = 136$). The number of waveforms in the nine clusters was 75 and, therefore, the process computed 75 normalised inner products and sorted them. The total time for computations

collection	K	feature vectors	time (s)
RX306	178	280	9.394

Similarity	Shot
1.000000	13438
0.997681	13439
0.991461	13435
0.990737	13437
0.988594	13487
0.986054	13355
0.986046	13239
0.985052	12749
0.984527	13202
0.984195	13485

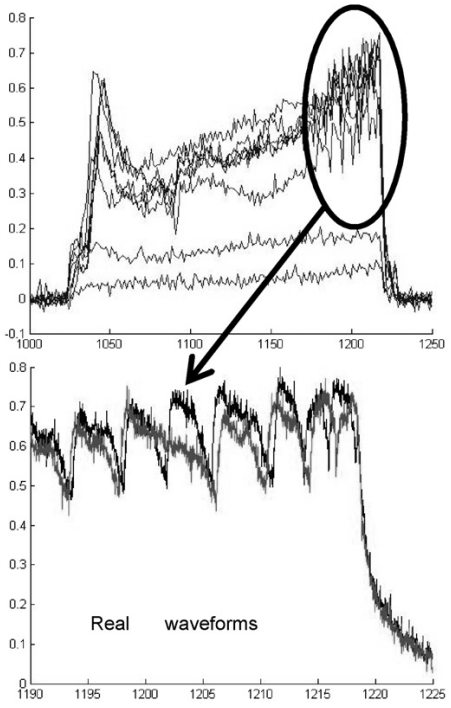


Fig. 4. SWRS results for a soft X-ray signal.

was 2.283 s and the CPU time was 1.482 s. The figure also gives information on measures of similarity, feature vectors and real waveforms. It should be noted that the similarity criterion is independent of signal polarity. Fig. 4 shows results for a soft X-ray signal collection. The CPU time for this calculation was 5.718 s. This example illustrates the independency of amplification gain and also the capability of finding oscillating patterns.

Note that computation times for the searching process are short enough and there is no need of defining additional layers to the classification system.

4.2. Client/server architecture for the TJ-II SWRS

An optimum use of the SWRS implies a general development for use in a shared environment (both local and wide area networks). Client/server architecture ensures a suitable means of interaction. The TJ-II SWRS server part resides in the TJ-II central data server (an AlphaServer computer with Tru64 UNIX operating system). Communication protocol between clients and server is TCP/IP, using Berkeley Sockets API (Application Program Interface). The communication mechanism is connection oriented and the server was developed as a concurrent server. This scheme is sufficiently general to service multiple concurrent connections and to allow the development of clients for several platforms and applications: visual data analysis applications and software library development.

The server part is in charge of managing the different classification systems for the several waveform collections (Fig. 5).

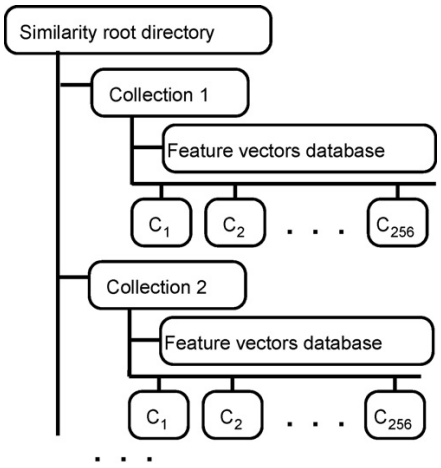


Fig. 5. Directory structure in the central server.

Each classification system is characterized by a database (that stores the feature vectors of the waveforms) and a set of files (one per cluster) containing the shot numbers that belong to the cluster. The database is based on a traditional method to handle queries on primary (i.e. unique) keys: hashing. In particular, the ndbm package of the UNIX operating system is used. Feature vectors are indexed according to shot number. The database and the 256 files (C_K, $K = 1, \ldots, 256$) share a common directory in the AlphaServer computer.

The server part provides computational resources

- to create new classification systems;
- to include new data (waveform classification);
- to retrieve information (similar waveform searches).

New data integration is carried out in two steps (Fig. 6): (1) data pre-processing and (2) feature vector classification. The former writes the wavelet coefficients into the database and the latter appends the shot number to the corresponding cluster file.

The searching process of similar waveforms takes four stages (Fig. 6): (1) input signal pre-processing, (2) feature vector classification into a cluster, (3) similarity factor computation with the feature vectors of N_C adjacent clusters and (4) similarity factor sorting in descending order. Note that the searching process does not require that the input signal should be previously classified. The existence of both the cluster files with the cluster composition and the database greatly speeds up the similarity factor computation.

In its first stage, the TJ-II SWRS system was built with four signal collections. The first one corresponds to a central chord of a soft X-ray detector array. The second one represents an integrated radiation signal from a line-of-sight near plasma centre measured by a bolometer detector. The third collection is an Hα emission measurement. Finally, the last collection groups a line averaged electron density measurement. The system has been created with data corresponding to the 2004/2005 TJ-II experimental campaigns. Each collection has 1350 waveforms approximately. Like in the Matlab case, the number of adjacent clusters to look for similar waveforms is 9.

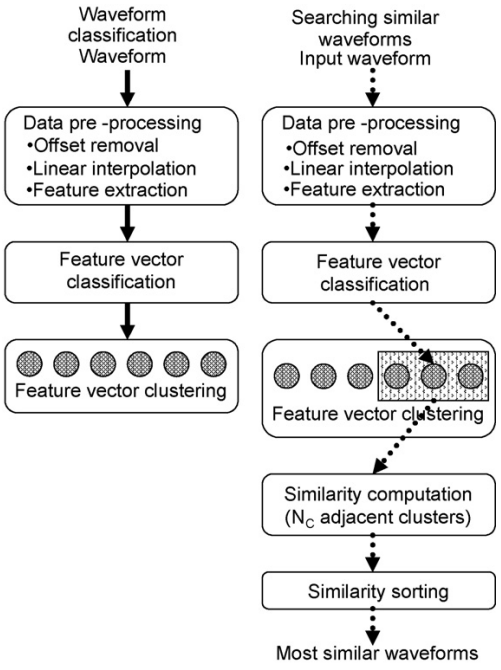


Fig. 6. Steps for waveform classification and searching processes.

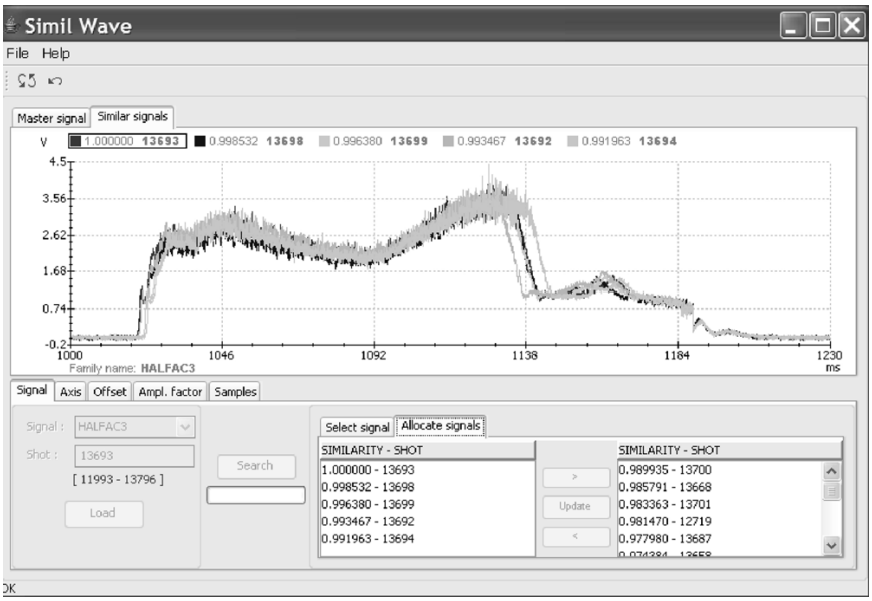


Fig. 7. TJ-II remote participation system GUI.

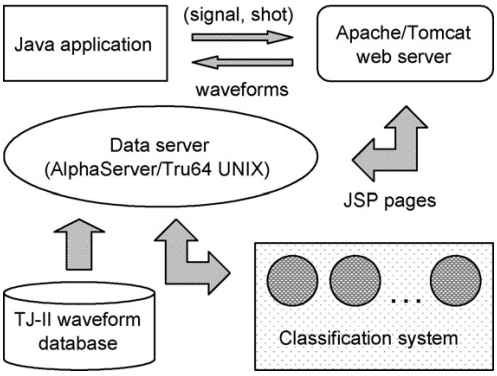


Fig. 8. TJ-II remote participation system data flow.

4.3. Integration into the TJ-II remote participation system

A first application of the SWRS was its integration into the TJ-II remote participation system. A Java application can be downloaded from the corresponding web page, according to the usual procedure in the TJ-II RPS [3]. Security for download and execution is based on the PAPI authentication and authorization system [14]. This Java application provides researchers with a point and click graphical user interface (GUI) to select waveforms and to search for the most similar ones. The application retrieves at most N_S similar waveforms and it shows the respective similarity factors. At present, N_S is normally 20, but this number can be easily modified at any moment. The GUI incorporates controls for horizontal and vertical expansion of traces, signal variable offset, absolute and relative measurements on waveforms, zoom capabilities and signal display with a variable sampling rates.

Fig. 7 shows the GUI after searching for similar waveforms. Signal selection is carried out under the 'Master signal' tab, whereas similar signal display is performed in the 'Similar signals' tab. Two list-boxes in the bottom of the window show similarity factors and shot numbers. The box in the bottom centre provides information about the signals in display. The waveforms inside the box at the bottom right are not displayed. However, signals can be moved between boxes either to appear or to disappear in the graphical area.

Fig. 8 shows the communication diagram with both remote and local users. Data exchange protocol is very simple and is carried out by means of Java Server Pages (JSP). The client application can send two types of queries. First, the user asks for a signal name and a shot number to perform visual data analysis for signal selection ('Master signal' tab). The server side transmits the data. Once the waveform to look for similar signals has been established, the user asks for them and the application resends signal name and discharge. Now, the server transfers, on the one hand, the similarity and shot number for N_S waveforms and, on the other hand, the waveforms. Signal transmission is accomplished in compressed format in order to save bandwidth and to speed up the transfer process. Data compression is realized according to standard TJ-II methods based on lossless techniques [15].

Acknowledgements

The authors wish to thank Prof. Sebastián Dormido Bencomo (UNED) and Prof. Jesús Manuel de la Cruz (UCM) for their constructive comments and invaluable guidance.

References

[1] C. Alejaldre, J. Alonso, L. Almoguera, E. Ascasíbar, A. Baciero, R. Balbín, et al., Plasma Phys. Controll. Fusion 41 (1) (1999) A539.
[2] J. Vega, E. Sánchez, A. López, A. Portas, M. Ochando, E. Ascasíbar, et al., Fusion Eng. Design 74 (2005) 775–780.
[3] J. Vega, E. Sánchez, A. Portas, A. Pereira, A. Mollinedo, J.A. Muñoz, et al., Fusion Eng. Design 81 (2006) 2045–2050.
[4] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, second ed., John Wiley & Sons, Inc., 2001.
[5] V. Cherkassky, F. Mulier, Learning from Data, John Wiley & Sons, Inc., 1998.
[6] N. Duro, J. Vega, R. Dormido, G. Farias, S. Dormido-Canto, J. Sánchez, et al., Fusion Eng. Design 81 (2006) 1987–1991.
[7] H. Nakanishi, T. Hochin, M. Kojima, LABCOM group, Fusion Eng. Design 71 (2004) 189.
[8] H. Nakanishi, T. Hochin, M. Kojima, LABCOM group, Fusion Eng. Design 81 (2006) 2003–2007.
[9] S. Dormido-Canto, G. Farias, R. Dormido, J. Vega, J. Sánchez, M. Santos, The TJ-II Team, Rev. Scient. Instrum. 75 (10) (2004) 4254–4257.
[10] G. Farias, S. Dormido-Canto, J. Vega, J. Sánchez, N. Duro, R. Dormido, et al., Fusion Eng. Design 81 (2006) 1993–1997.
[11] S. Mallat, A Wavelet Tour of Signal Processing, second ed., Academic Press, 2001.
[12] V. Vapnik, The Nature of Statistical Learning Theory, second ed., Springer, 2000.
[13] http://www.mathworks.com.
[14] R. Castro, D.R. López, J. Vega, Fusion Eng. Design 81 (2006) 2057–2061.
[15] J. Vega, C. Crémy, E. Sánchez, A. Portas, S. Dormido, Rev. Scient. Instrum. 67 (12) (1996) 4154–4160.

Article 13

A computational fusion of wavelets and neural networks

13.1 Bibliographic Description

Title

A computational fusion of wavelets and neural networks in a classifier for biomedical applications.

Citation

G. Farias, M. Santos (2007) A computational Fusion of Wavelets and Neuronal Networks in a Classifier for Biomedical Applications, *Lecture Series on Computer and Computational Sciences*, ISSN 1573- 4196, Volume 8, Pages 66-70, Brill Academic Publishers.

Abstract

The purpose of this paper is to develop a computational tool in order to identify different types of human brain tumours. The Wavelet-Neural classifier merges wavelet transform to reduce the size of the medical spectrum and to extract the main features, with a feedforward neural network. It also allows to analyze the influence of the design parameters of each of those techniques on the clustering. The classification results are promising specially taking into account that medical knowledge has not been considered. The developed tool could help to confirm the histological diagnosis.

References

J. Peeling, G. Sutherland (1992); J.M. Roda et al. (2000); G. Farias, M. Santos (2005); The MathWorks Inc.(1989); G. Hagberg (1998); A.R. Tate (1997); S.L. Howells, R. Maxwell, J.R. Griffiths (1992); I. Martínez-Pérez et al. (1995); I. Daubechies (1992).

Impact Factor

There is not impact factor for this journal.

Brill Academic Publishers
P.O. Box 9000, 2300 PA Leiden,
The Netherlands

Lecture Series on Computer and Computational Sciences

A computational fusion of wavelets and neural networks in a classifier for biomedical applications

G. Farias, M. Santos[1]

[2]Department of Computer Science and Technology,
Faculty of Computer Science,
Universidad Complutense de Madrid
28040-Madrid, Spain

Abstract: The purpose of this paper is to develop a computational tool in order to identify different types of human brain tumours. The Wavelet-Neural classifier merges wavelet transform to reduce the size of the medical spectrum and to extract the main features, with a feedforward neural network. It also allows to analyze the influence of the design parameters of each of those techniques on the clustering. The classification results are promising specially taking into account that medical knowledge has not been considered. The developed tool could help to confirm the histological diagnosis.

Keywords: Soft Computing, Neural Networks, Wavelets, Biomedical Data, Clustering

ACM Subject Classification Index: I.2 Artificial Intelligence. I.5 Pattern Recognition. I.5.3 Clustering

1. Introduction

A main concern in the medical environment is the development of nonhistological methods of diagnosis based on *in vitro* ^{1}H Magnetic Resonance Spectroscopy (MRS) biopsies of human brain tumours. Histological procedures remain mandatory for tumour diagnosis. However, pathologist may find these alternatives protocols useful in cases where a confirmation of the histological diagnosis by an independent method is advisable or in situations in which adequate anatomopathological examinations cannot be performed.

The progress in statistical techniques and in pattern recognition suggests an automatic evaluation. This automatic procedure could be implemented in surgical spectrometers. Soft computing could provide useful tools to deal with this information. They are not intrusive methods and at the same time they can incorporate the knowledge of the experts if it is available.

In this work, Neural Networks (NN) is applied. The original point of view is that in the literature, the characterization of tumours is based on the medical knowledge of the molecular or metabolic profiles. But this knowledge is difficult to obtain. In this paper the clustering is, as it should be, independent of the medical knowledge. The method uses just computational information in an automatic procedure.

Previously to the neural network, another computational tool, Wavelet Transforms (WT), is applied to extract the relevant information. The brain signals need to be pre-processed as the data are influenced by the conditions in which the samples were taken. Moreover, the spectra are large and complex. Therefore the application of some compression technique is required in order to reduce the size of the spectra and to obtain the main features while filtering the noise.

So, a hybrid computational tool that merges these two techniques is developed: an intelligent classifier that applies wavelets and neural networks. The results are encouraging taking into account that medical knowledge is not considered.

[1] Corresponding author. Matilde Santos. Dpto. de Arquitectura de Computadores y Automática E-Mail: msantos@dacya.ucm.es

2. The Medical Data and the Processing

Nuclear Magnetic Resonance (NMR) has provided a great help in the knowledge of the different pathologies. Nevertheless, its diagnostic application has been limited due to the fact that the *in vivo* ^{1}H NMR spectra only give small accuracy and because of the difficulties founded in the quantification of the metabolites. Most of the limitations could be overcome if extracts of tumour biopsies are available, by applying the *in vitro* technique [1]. This method has been used to obtain the spectra that are going to be analysed in this paper. The preparation and characterization of biopsies are described in [2].

The spectra that have obtained in this way have 16384 samples taking into account only the real part. The spectrum of each tumour represents the intensity –proportional to the concentration of protons in the tissue- (y axis) vs. the distance in ppm (part per million) (x axis), i.e., at a particular resonance frequency which depends on the magnetic field.

The one-dimensional ^{1}H MRS can be classified in eight different groups that correspond to normal brain and seven different classes of human brain tumors, as it has be stated by the WHO (World Health Organization). Table 1 shows the different classes and the number of available samples of each of them.

Table 1. Tumour classes and number of samples (Classes 01 to 08)

High Grade Astrocytoma	Low Grade Astrocytoma	Normal Brain	Medulloblastoma
12	16	16	4
Meningioma	Metastasis	Neurinoma	Oligodendroglioma
31	14	9	10

Before the classification, the signals have been normalized in both, the resonance intensity for the amount of tissue extracted and the number of samples of the spectra. In addition, the representative information of ^{1}H spectra is concentrated in the range of 0,8 to 4,22 pmm, so that this is the interval of frequencies that has been considered. The number of points of each spectrum is then 4655.

The use of the Discrete Wavelets Transform (DWT) makes possible to reach a desired decomposition level preserving the signal information. The redundant information is minimized and so the computational load is substantially cut down. After applying the wavelet transform, the number of data is reduced in an exponential way while the decomposition level increases. In our case, each spectrum has been reduced form 4655 samples to 291 attributes when decomposition level 4 is applied. However, to select the most suitable family of mother wavelets and the best scale for particular signals is a difficult task [3].

3. The Wavelet-Neural classifier

Our purpose is to develop a computational tool that fulfills the following requirements: (i) Accuracy in the diagnosis (high percentage of correct classifications), (ii) Easy to integrate in a surgery environment; (iii) Friendly use and easy to apply, i.e., it does not require specialized skills; (iv) Open to modifications and improvements.

The classifier has been implemented in MATLAB [4] by applying feeedforward Neural Networks (NN) in the spatial domain. Although other methods have been used, as SVM, this has been proved to give good results. The developed computational application (Figure 1) allows not only to classify biomedical spectra and any other kind of signals but also to evaluate the performance of different classifier structures. These classifier configurations can be easily obtained by modifying some parameters as:

- In the pre-processing step: the wavelet mother, the decomposition level, the coefficient, etc.
- In the neural classifier: number of layers, neurons, epochs, activation functions, error goal, etc.

The spectra are displayed in the image window at the left side. Once the type and parameters of WT have been chosen, the View option allows to show either the original spectrum or its wavelet transform at any stage. To start the classification process, when pressing the Generate button two sets of signals are randomly obtained for training and testing purposes. The proportion of signals of each set can be

defined by the user. After pressing the Train button, the NN works until it reaches the error goal. To evaluate the results it is necessary to press the Classify button. Automatically the classifier will also compare the obtained results with the labeled classes giving the percentage of correct classifications (success) and the processing time.

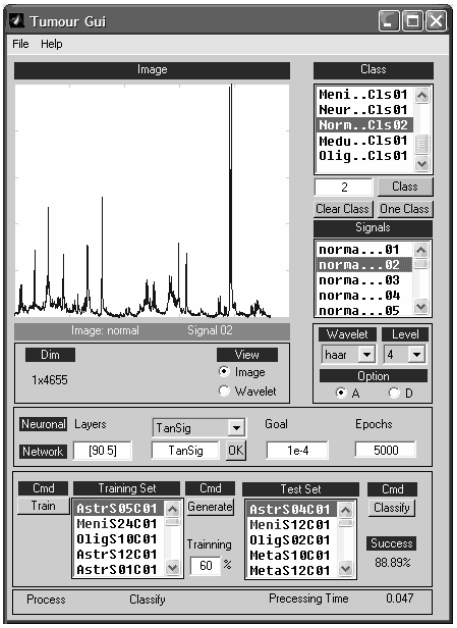


Figure 1: Graphical user interface of the Wavelet-Neural Classifier.

4. Results

After trying with different families of wavelets filters [9], wavelet Haar at level 4 with approximation coefficient has been selected based on [3]. The main reason is because it performs well both with one-dimensional signals and images. In this paper a Feed Forward NN with supervised learning is used. The NN has an input layer of 291 attributes generated by the WT. Two hidden layers with 140 and 70 cells are implemented, with activation function *Tansig*. For binary classification, the output layer has 2 neurons with *Logsig* function. After applying Back-Propagation for training the NN by the LOO (Leave-one-out) strategy, every signal is associated to its corresponding class.

To test the classifier, many experiments were carried out. Correct diagnosis is obtained when the class selected by the computational tool matches the histological diagnosis. First of all, a binary classification was carried out. Using this strategy, it is possible to classify the complete data set into only two groups, normal brain and tumor pathologies. The scores provided for correct classification were 95,7 %.

After that, binary comparisons were performed between every tissue class and each one of the remaining ones (Table 2). Classification between two classes may yield different scores for each class, depending on the number of elements and the number of correct classification in each class. The first row depicts the comparison of high grade astrocytoma. For instance, when compared with low grade astrocytoma, the correct score was 66,7 %. That is, 8 extracts of the total of 12 biopsies of high grade astrocytoma class were correctly classified. Similar interpretations are applicable to the rest of the rows and binary comparisons.

The best results were obtained for normal brain against meningioma. As it is possible to see in Table 2, the scores were 90,32% (element 5,3) and 93,75% (element 3,5), and the number of extracts of these classes is fairly large. The average weighted of these results, taking into account the number of biopsies of those classes, was 91,48%. The worse case was the comparison between medulloblastoma and high grade astrocitoma because of the scant number of elements of the medulloblastoma class, only 4

extracts, and maybe because the profiles of both tissue classes are quite similar and so difficult to distinguish between them without extra knowledge.

Table 2. Average percentage of correct classifications

	High Grade Astrocytoma	Low Grade Astrocytoma	Normal Brain	Medullo blastoma	Menin gioma	Metastasis	Neuri noma	Oligodendr oglioma
High Grade Astrocytoma		67	75	92	84	58	100	59
Low Grade Astrocytoma	75		94	81	82	88	81	82
Normal Brain	88	94		94	94	88	94	94
Medulloblastoma	50	50	75		75	50	75	75
Meningioma	87	90	90	97		87	84	90
Metastasis	57	93	93	86	79		86	93
Neurinoma	67	56	89	56	78	67		89
Oligodendroglioma	50	60	80	80	50	70	80	

Multi-class classification was also applied to establish the different groups. The output layer of the NN was set to 8 neurons, involving the eight possible classes considered. The final values were obtained by carrying out more than 30 experiments to calculate average values. The scores obtained from the multiclassification represent higher accuracy in the classification process than when considering the effects of chance (12,5 %).

5. Conclusions

A computation tool that merges Wavelets and NN is developed. The wavelet-neural classifier allows to observe the influence of the design parameters of each technique on the clustering, so to reduce the classification time and to improve the results.

A relevant aspect is the comparison of the scores obtained with this tool with those provided by alternative procedures [2, 5, 6, 7, 8]. The percentage of correct classification with this method may be a bit lower, although it reaches the 100% in some cases. However, a relevant advantage of the proposed tool is that it allows non specialists to classify any sample of the database without applying medical knowledge. It is also necessary to emphasize that the training data set of the classifier was very limited.

This tool could help the histologists to make a decision and to confirm his diagnostic, and it constitutes an alternative for automated classification of biomedical spectra.

References

[1] Peeling J, Sutherland G (1992) High-resolution ^{1}H NMR spectroscopy studies of extracts of human cerebral neoplasm. Magn. Reson. Med. 24: 123-136.

[2] Roda JM, Pascual JM, Carceller F, Gonzalez-Llanos F (2000) Nonhistological Diagnosis of Human Cerebral Tumors by ^{1}H Magnetic Resonance Spectroscopy and Amino Acid Analysis. Clinical Cancer Research 6: 3983-3993.

[3] Farias G, Santos M (2005) Analysis of the Wavelet Transform Parameters in Images Processing. *Lectures Notes on Computer and Computational Sciences*, vol. 2, pp. 51-54.

[4] MATLAB® (1989). The MathWorks, Inc., MA, USA.

[5] Hagberg G (1998) From magnetic resonance spectroscopy to classification tumors. A review of pattern recognition methods. *NMR Biomed*. 11: 148-156.

[6] Tate AR (1997) Statistical pattern recognition for the analysis of biomedical magnetic resonance spectra. J. Magn. Resonance Anal. 3: 63-78.

[7] Howells SL, Maxwell R, Griffiths JR (1992) An investigation of tumor ^{1}H NMR spectra by pattern recognition. NMR Biomed. 5: 59-64.

[8] Martínez-Pérez I, Maxwell RJ, Howells SL, van der Bogaart A, Mazucco R, Griffiths JR, Arús C (1995) Pattern recognition analysis of ^{1}H NMR spectra from human brain tumours biopsies. Proc. Soc. Magn. Reson., 3rd Annual Meeting, Abstract P1709.

[9] Daubechies, I (1992) *Ten Lectures on Wavelets*. CBMS Lectures Series, SIAM, Philadelfia.

Article 14

Search and retrieval of plasma wave forms

14.1 Bibliographic Description

Title

Search and retrieval of plasma wave forms: Structural pattern recognition approach.

Citation

S. Dormido-Canto, G. Farias, J. Vega, R. Dormido, J. Sánchez, N. Duro, M. Santos, J.A. Martín, G. Pajares (2006) Search and retrieval of plasma waveforms: structural pattern recognition approach, *Review of Scientific Instruments*, ISSN 0034-6748, Volume 77, Pages 10F514-1/10F514-4.

Abstract

Databases for fusion experiments are designed to store several million wave forms. Temporal evolution signals show the same patterns under the same plasma conditions and, therefore, pattern recognition techniques can allow identification of similar plasma behaviors. Further developments in this area must be focused on four aspects: large databases, feature extraction, similarity function, and search/retrieval efficiency. This article describes an approach for pattern searching within wave forms. The technique is performed in three stages. Firstly, the signals are filtered. Secondly, signals are encoded

according to a discrete set of values (code alphabet). Finally, pattern recognition is carried out via string comparisons. The definition of code alphabets enables the description of wave forms as strings, instead of representing the signals in terms of multidimensional data vectors. An alphabet of just five letters can be enough to describe any signal. In this way, signals can be stored as a sequence of characters in a relational database, thereby allowing the use of powerful structured query languages to search for patterns and also ensuring quick data access.

References

H. Nakanishi, T. Hotchin, M. Kojima (2004); S. Dormido-Canto et al. (2004); G. Farias et al. (2006); H. Nakanishi, T. Hotchin, M. Kojima (2006); K. S. Fu (1982); The MathWorks, Inc. (2006).

Impact Factor

Review Of Scientific Instruments has an impact factor of 1.367 according to Thomson Reuters Journal Citation Reports (2011).

REVIEW OF SCIENTIFIC INSTRUMENTS **77**, 10F514 (2006)

Search and retrieval of plasma wave forms: Structural pattern recognition approach

S. Dormido-Canto[a] and G. Farias
Departameno de Informática y Automática, UNED, C/Juan del Rosal 16 5a, 28040 Madrid, Spain

J. Vega
Asociación EURATOM/CIEMAT para FUSIÓN, Avenida Complutense 22, 28040 Madrid, Spain

R. Dormido, J. Sánchez, and N. Duro
Departameno de Informática y Automática, UNED, C/Juan del Rosal 16 5a, 28040 Madrid, Spain

M. Santos, J. A. Martin, and G. Pajares
Departamento de Arquitectura de Computadores y Automática, UCM, Ciudad Universitaria, 28040 Madrid, Spain

(Received 5 May 2006; presented on 11 May 2006; accepted 28 May 2006; published online 29 September 2006)

Databases for fusion experiments are designed to store several million wave forms. Temporal evolution signals show the same patterns under the same plasma conditions and, therefore, pattern recognition techniques can allow identification of similar plasma behaviors. Further developments in this area must be focused on four aspects: large databases, feature extraction, similarity function, and search/retrieval efficiency. This article describes an approach for pattern searching within wave forms. The technique is performed in three stages. Firstly, the signals are filtered. Secondly, signals are encoded according to a discrete set of values (code alphabet). Finally, pattern recognition is carried out via string comparisons. The definition of code alphabets enables the description of wave forms as strings, instead of representing the signals in terms of multidimensional data vectors. An alphabet of just five letters can be enough to describe any signal. In this way, signals can be stored as a sequence of characters in a relational database, thereby allowing the use of powerful structured query languages to search for patterns and also ensuring quick data access. © *2006 American Institute of Physics.* [DOI: 10.1063/1.2219409]

I. INTRODUCTION

Visual data analysis is an essential tool in plasma physics. A simple visual inspection of signals is enough to recognize a typical plasma evolution or to distinguish the presence of interesting events. A researcher identifies the plasma behavior through the recognition of patterns inside wave forms: bumps, unexpected amplitude changes, abrupt peaks, or sinusoidal components. Therefore, a big challenge in data access is the creation of fast means to look for patterns within wave forms. These techniques will allow the development of intelligent data retrieval methods instead of using manual searches by pulse number (in general) or identifiable time interval (in long pulse operation).

There are some previous works on pattern recognition in fusion databases. In an earlier approach, efforts were concentrated in looking for similar full wave forms, i.e., signals covering the full plasma life.[1–3] In another approach, the interest is centerd in searching for patterns within wave forms. A pioneer work[4] describes the search of patterns based on one major frequency component. However, more general methods are required to look for general patterns.

II. SYNTACTIC AND STRUCTURAL PATTERN RECOGNITION APPROACH

The *syntactic approach* takes the view that a pattern is composed of simpler subpatterns.[5] The most elementary subpatterns are known as primitives. A complex pattern is then expressed in terms of relationships among its primitives. An analogy between the structures of patterns and the theory of formal languages is used to establish the foundation for syntactic pattern recognition. The patterns represent the sentences in a language, while the primitives constitute the alphabet of the language. A grammar for a language generates and identifies sentences belonging to that language by employing its rules. The idea that a potentially large set of related complex patterns can be described by a finite number of primitives, and grammatical rules makes this approach appealing.

There are many applications where patterns can be described in terms of primitives and their relations. However,

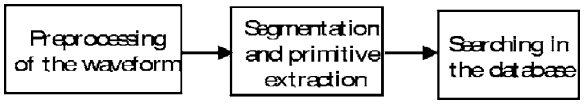


FIG. 1. The flow chart of our recognition system.

[a]Electronic mail: sebas@dia.uned.es

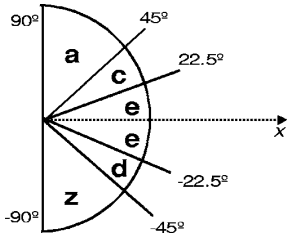


FIG. 2. Discriminates and labels for the classification of the angle of the fitted straight line.

sometimes a grammar is not suitable for a pattern class description because the patterns under consideration lack regularities and cannot be defined by rules. In such a case, the *structural approach* to pattern recognition can be adopted. In structural pattern recognition, we use symbolic data structures, such as strings, for the representation of individual patterns, similar to the syntactic approach. However, rather than use a grammar, we represent pattern classes through a number of primitives. Consequently, the recognition problem turns into a pattern-matching problem. For example, given a pattern decomposed into primitives (set of characters: string), the final goal is to find the most similar pattern from a database of strings.

The description task of a structural pattern recognition system is difficult to implement because there is no general solution for extracting structural features (primitives) from data. The result is that primitive extractors for structural pattern recognition systems are developed to extract either the simplest and most generic primitives possible or the domain specific primitives that best support the subsequent searching task. Simplistic primitives are domain independent; therefore, a deeper interpretation is postponed until the searching. At the other extreme, domain specific primitives can be developed with the assistance of a domain expert, but obtaining and formalizing the necessary domain knowledge can be problematic.

III. APPLICATIONS TO TIME-SERIES DATA

Identification problems involving time-series data (or wave forms) constitute a subset of pattern recognition applications that is of particular interest because of the large number of domains that involve such data (for instance, fusion databases). Although structural approaches are particularly appropriate in domains where domain experts classify time-series data sets based on the arrangement of morphological events evident in the wave form (e.g., speech recognition, electrocardiogram diagnosis, seismic activity identification, radar signal detection, and process control), we are interested in a domain-independent structural pattern recognition system, which is one that is capable of acting as a "black box" to extract primitives and perform searching without the need for domain knowledge.

Our method is applied to fusion databases with the aim of looking for similar patterns within wave forms. The technique consists of three stages. First we preprocess the signal by applying a low-pass filter for smoothing purposes. Then we extract primitives, which encode the most elementary pieces of structural information of the pattern. Finally, we can find similar patterns from a database of strings, where the patterns represent the temporal evolution of physical properties. It should be remarked that the technique allows searching for patterns of any time length. The general flow graph of our searching pattern method is described in Fig. 1.

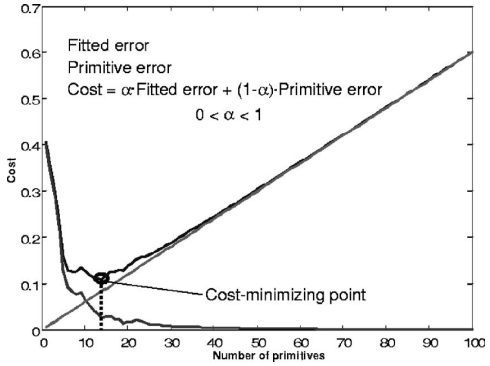


FIG. 3. Cost function for a signal.

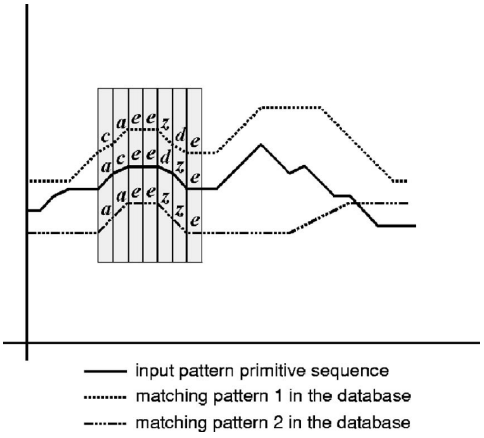


FIG. 4. Primitive sequences in the coarse searching.

IV. COMPUTATION OF PRIMITIVES

Selection of primitives is an essential issue in the structural pattern recognition of wave forms because they deter-

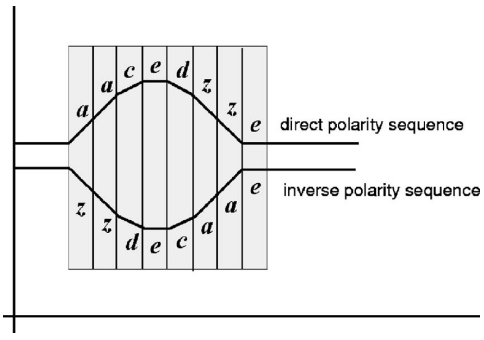


FIG. 5. Direct and inverse primitive sequences.

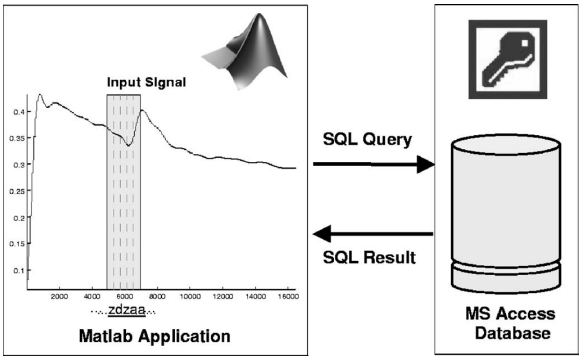


FIG. 6. Application scheme.

mine what types of structural components we can construct. There are plenty of different ways to compute the primitives of the wave forms, such as constant, straight, exponential, sinusoidal, triangular, and trapezoidal structures. We have used the straight structure (line segment) which is easy and fast to calculate.

In our method we divide the original signal into segments (all the segments have the same number of samples) which are fitted with a straight line. A least squares minimization procedure is used to obtain each straight line. Then we encode these segments into a string of primitives. We give a label to each segment, and we calculate the amplitude between the first and the last sample into the primitive. The labeling of the segment $\{(x_i,y_i),(x_j,y_j)\}$ is based on the classification of the slope of the fitted straight line. We find the primitives P where the angle of the line with the x axis belongs [Eq. (1)].

$$\mathrm{Lab}(\{(x_i,y_i),(x_j,y_j)\}) = P, \tag{1}$$

if the lower limit of primitive $P < \arctan[(y_j-y_i)/(x_j-x_i)] \leqslant$ the upper limit of primitive P *end*.

Our discriminate values and the primitive labels are depicted in Fig. 2. The classification of the angle gives us all

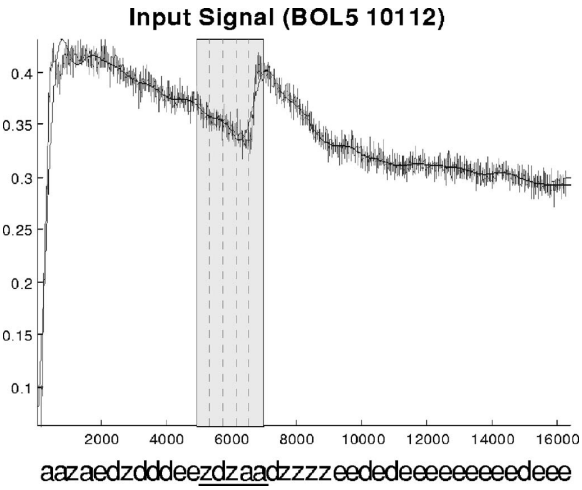


FIG. 7. Input signal and section to search.

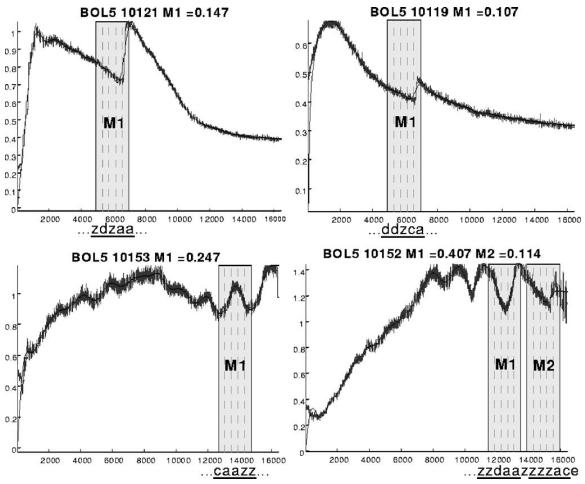


FIG. 8. Some shots returned and their matches.

the elementary structural information needed to construct more complex subpatterns in wave form recognition. The amplitude of a segment is computed in a straightforward manner [Eq. (2)]:

$$\mathrm{Amplitude}_{i,j} = |y_j - y_i|. \tag{2}$$

Thus, our input to the system is composed by (1) a string of n primitives, where n is *samples/samples per primitive*, and (2) an array with the amplitudes of each primitive.

We use five different values (a, c, e, d, and z) to represent the classes of the angle. With a bigger amount of primitive classes we could have expressed more accurately the structure of the signal, but the final string would have been more complex. On the other hand, these five codes are just enough for a typical plasma evolution analysis. The code e represents a flat part of a signal, codes c and d represent the ascending and descending angles, and codes a and z represent the extremely steep slopes.

In order to obtain a suitable number of primitives (n) we minimize a cost function [Eq. (3)] where $0 \leqslant \alpha \leqslant 1$, A is the fitted error, and B the primitive error (number of primitives). The fitted error is estimated between the filtered signal and the straight lines from the least squares minimization procedure. We evaluate the function until we obtain 100 primitives, and we select the value with minimum cost. Figure 3 shows an example for the evaluation of a cost function.

$$J = \alpha A + (1-\alpha)B. \tag{3}$$

V. TYPES OF SEARCHING

In our case, we define two types of searching: (1) fine searching and (2) coarse searching.

In the *fine searching* we will obtain from the database the patterns whose sequences of primitives match exactly with the input pattern sequence. In the *coarse searching*, it is possible to associate different primitives into the same label. For example, if $a \Leftrightarrow c$ and $z \Leftrightarrow d$ we would have the following set of primitive labels: [a or c], e, and [z or d]. Where [a or c] represent the ascending angle, e represents a flat part

TABLE I. SQL results of searching example-codes.

Shot	Code
10108	*aaazzzddeeeeeezaazzdzzeeeeeeeeeeeeeeeee*
10109	*aaazzzzzzzzzdeedeeeeeeeeeeeeeeeeeeeeee*
10110	*aaazzzzzzzzzeeeeceeeeeeeeeeeeeeeeeeeee*
10112	*aazaedzdddeezdzaadzzzzeededeeeeeeeedeee*
10115	*aaazzzzzzzzzdddecdeeeeeeeeeeeeeeeeeeee*
10116	*aaazzzzzzzeeeeecaddeeedeeeeeeedeeeeeee*
10119	*aeaazzzzzzzzdzcazzddddedeeeeeeeeeeeee*
10120	*aaazdzzzzzzzzdzaaazzdzzdeeeeeeeeeeeeeeeee*
10121	*aaazeeddddddzdzaaezzzzzzzzzdedeeeeeeee*
10152	*azaaaaaaaaeaacaaaaaezcaezeaezzdaazzzzace*
10153	*aaaaaaccaaedaaeecaeeeezzedzeazzcaazzaaae*
10155	*aaaaezzeeddecezzeeeedzaazeaaadzacedcaeze*
10176	*aaazzzzzzdeeeeeeeeeeeeeeeeeeeeeeeeeee*
10194	*aaaaaacccccccceeeeeccececaccdeaaaaaczze*
10222	*aaaczzzzzzdeeeeceeeeeeeaaaeeeecaecccecee*
10227	*aazzzeeeeeeccceeeccecaaaaezzdcceacdeeeee*
10228	*aazzzdeeeeeeeeceecacaaaaadzeezdceeaeecce*
10104	*aadzddeeeeeeeeeeeeeeceeecaazzzzzccczaazzce*

TABLE II. SQL results of searching example-distance errors.

Shot	Error M1	Error M2
10108	0.155 175 645 91	
10109	0.338 439 721 8	
10110	0.266 122 453 8	
10112	1.774 495 9321 $E-10$	
10115	0.283 615 803 3	
10116	0.205 821 807 5	
10119	0.107 452 797 05	
10120	0.190 652 696 1	0.233 227 516 3
10121	0.147 236 874 64	
10152	0.407 779 733 38	0.114 520 788 97
10153	0.247 111 540 27	
10155	0.245 258 614 59	
10176	0.642 049 205 7	
10194	0.135 704 051 57	
10222	0.132 745 187 24	
10227	0.101 018 156 59	
10228	0.079 871 870 697	
10104	0.153 679 361 30	0.533 216 355 46

and [z or d] the descending angle. In this case, if the input pattern sequence is *aceedze*, we will obtain from the database those sequences that fit in the following combined sequences: [a or c], [c or a], e, e, [d or z], [z or d], and e. Observe that, in this example, there are 16 possible sequences (Fig. 4 shows two of these).

In the application it is also possible to search, in the database, wave forms with inverse polarity with regard to the input signal. For example, if the input pattern sequence is *aacedzze*, we would substitute the labels of the primitives by theirs symmetries *(zzdecaae)*, as shown in Fig. 5.

VI. APPLICATION SCHEME

The types of searching mentioned before can be done with a relational database management system using Structure Query Language (SQL). In this work, we have used MICROSOFT ACCESS™ because it is easy to build a database and test our approach.

Preprocessing and primitive computing of wave forms were done by means of MATLAB™; the MATLAB Database Toolbox[6] was used for the link between MATLAB and Access.

The application algorithm is the following: First a user selects a shot, then chooses a section of the signal (pattern), and asks for the application for a type of searching (fine or coarse). The MATLAB application carries out the preprocessing and primitive computation. After that, an SQL query is realized with regard to the searching type selected. Finally, ACCESS sends back to MATLAB the SQL results, and the application shows all matches in the returned signals. Figure 6 shows the application scheme.

VII. SEARCHING EXAMPLE

In this example we select patterns from the TJ-II stellarator database. The input is a signal from a bolometer (BOL5) in shot 10112, as it is shown in Fig. 7. After that, a coarse direct and inverted searching is chosen. In this case, the SQL sentence is

```
select * from codes where Type like 'BOL5' and (Codce like '% [z, d] [z, d] [z, d] [a, c] [a, c] %' or code like '% [a, c] [a, c] [a, c] [z, d] [z, d] %' )
```

Once ACCESS has executed the SQL query, it sends back to MATLAB the results, and finally, the application shows the matches. Figure 8 depicts some of the matches, and Tables I and II show all SQL results. Note that errors in matches can be used to sort the SQL results.

VIII. DISCUSSION

We present a structural approach based on high level codification and SQL for pattern searching within wave forms. We have tested the system with several wave forms, and we have obtaines excellent results in the detection of subpatterns. The computation time is suitable for an interactive searching because we have used the power of a relational database management system for retrieving similar wave forms.

[1] H. Nakanishi, T. Hotchin, M. Kojima, and LABCOM group, Fusion Eng. Des. **71**, 189 (2004).
[2] S. Dormido-Canto, G. Farias, R. Dormido, J. Vega, J. Sénchez, M. Santos, and TJ-II Team, Rev. Sci. Instrum. **75**, 4254 (2004).
[3] G. Farias, S. Dormido-Canto, J. Vega, J. Sénchez, N. Duro, R. Dormido, M. Santoa, and G. Pajares, Fusion Eng. Des. (in press).
[4] H. Nakanishi, T. Hotchin, M. Kojima, LABCOM group, Fusion Eng. Des. (in press).
[5] K. S. Fu, *Syntactic Pattern Recognition and Applications* (Prentice-Hall, Englewood Cliffs, NJ, 1982).
[6] The MathWorks, Inc., Database Toolbox for use with MATLAB, User's Guide, Version 3, 1998–2006.

Article 15

Searching for patterns in TJ-II time evolution signals

15.1 Bibliographic Description

Title

Searching for patterns in TJ-II time evolution signals.

Citation

G. Farias, S. Dormido-Canto, J. Vega, J. Snchez, N. Duro, R. Dormido, M. Ochando, M. Santos, G. Pajares (2006) Searching for patterns in TJ-II time evolution signals, *Fusion Engineering and Design*, ISSN 0920-3796, Volume 81, Pages 1993-1997, Ed. Elsevier.

Abstract

Since fusion plasma experiments generate hundreds of signals, it is important for their analysis to have automatic mechanisms for searching for similarities and retrieving specific data from the signal database. This paper describes a technique for searching in the TJ-II database that combines support vector machines and similarity query methods. Firstly, plasma signals are preprocessed by wavelet transform or discrete Fourier transform to reduce the dimensionality of the problem and to extract their main features. Secondly, support vector machines are used to classify a set of signals by reference to an input signal. Finally, similarity query methods (Euclidean distance and bounding

envelope) are used to search the set of signals that best matches the input signal.

References

R. Duda, P. Hort, D. Stork, (2001); H. Nakanishi, T. Hochin, M. Kojima (2004); S. Dormido-Canto et al. (2004); I. Daubechies ((1992); S. Mallat (2001); V. Vapnik (2000).

Impact Factor

Fusion Engineering and Design has an impact factor of 1.49 according to Thomson Reuters Journal Citation Reports (2011).

Available online at www.sciencedirect.com
SCIENCE DIRECT®
Fusion Engineering and Design 81 (2006) 1993–1997
Fusion Engineering and Design
www.elsevier.com/locate/fusengdes

Searching for patterns in TJ-II time evolution signals

G. Farias [a,*], S. Dormido-Canto [a], J. Vega [b], J. Sánchez [a], N. Duro [a], R. Dormido [a], M. Ochando [b], M. Santos [c], G. Pajares [c]

[a] *Dpto. Informática y Automática – UNED, 28040 Madrid, Spain*
[b] *Asociación EURATOM/CIEMAT para Fusión, 28040 Madrid, Spain*
[c] *Dpto. Arquitectura de Computadores y Automática – UCM, 28040 Madrid, Spain*

Available online 2 May 2006

Abstract

Since fusion plasma experiments generate hundreds of signals, it is important for their analysis to have automatic mechanisms for searching for similarities and retrieving specific data from the signal database. This paper describes a technique for searching in the TJ-II database that combines support vector machines and similarity query methods. Firstly, plasma signals are pre-processed by wavelet transform or discrete Fourier transform to reduce the dimensionality of the problem and to extract their main features. Secondly, support vector machines are used to classify a set of signals by reference to an input signal. Finally, similarity query methods (Euclidean distance and bounding envelope) are used to search the set of signals that best matches the input signal.

Keywords: Search pattern; Support vector machines; Similarity query methods; Wavelet transform; Discrete Fourier transform

1. Introduction

Most experimental signals in fusion are devoted to studying the time evolution of plasma properties. Diagnostics translate their observations into electrical signals that are digitized and stored for off-line analysis. Different plasma physical behaviours are shown by the different signals generated by the diagnostics. In general, a linear mapping can be established to connect the time evolution of a physical phenomenon with the kind of signal that it generates. Therefore, it is possible to speak about patterns. To analyse plasma properties, pattern search can be very helpful. However, an experimental database of a fusion device contains thousands of signals, so automated pattern recognition methods are required. Pattern recognition is a computational technique used to find patterns and develop classification schemes for data in very large data sets [1]. In this paper, we describe an automated technique to search for and retrieve similar time evolution signals to a reference waveform.

The procedure is divided in three stages. Firstly, the reference waveform is processed to extract signal features, i.e. a set of reduced properties to encode the

* Corresponding author. Tel.: +34 91 3987147; fax: +34 91 3988663.
E-mail address: gfarias@bec.uned.es (G. Farias).

0920-3796/$ – see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.fusengdes.2006.04.013

waveform. Secondly, a classification system performs a coarse filter to reduce the search space. Thirdly, similarity query methods are used to retrieve the waveforms most similar to the input signal. The technique has been applied to time evolution signals from different sensors on the TJ-II stellarator.

2. Feature extraction

A nuclear fusion environment is very hostile to experimental measurements from the electromagnetic point of view. Signals are digitized for the duration of a discharge; therefore they contain thousands of samples. Before applying classification and similarity procedures, it is necessary to reduce the noise and the dimensionality of the signals. Different methods can be applied, i.e. discrete Fourier transform (DFT) [2] and discrete wavelet transform (DWT) [3].

The wavelet transform is a very powerful computational tool whose application allows a high level of compression without losing information [4].

The use of the DWT makes possible to reach a desired decomposition level preserving signal information. Redundant information is minimized and the computational load is substantially cut down. Some of the properties that make wavelets so useful in pattern recognition are: the capacity of noise reduction, the signal enhancement and the detection of similarities. They allow the analysis of periodic and non-periodic signals. It is a time-scale approach, which shows results in the time–frequency plane [5].

Our procedure uses the DWT approximation for extraction of characteristics. When signals are processed by DWT a signal with 16,384 samples is reduced to 64 (Fig. 1).

3. Support vector machines

After feature extraction, the search process begins. The search procedure can be divided in two steps. The first one is used to narrow down the search space, i.e. to limit the search of the signals to a proper subset of the database. This is carried out by means of a classification system, which is previously trained to distinguish among different kind of signals. The training process is accomplished with TJ-II database signals whose features are the DWT coefficients.

We use the support vector machine (SVM) method for the classification system because the DWT + SVM combination has shown to be powerful enough for pattern recognition in the TJ-II environment [3].

The support vector machine is a universal constructive learning procedure based on the statistical learning theory [6]. The SVM maps input data into a high-dimensional space using a non-linear function.

Once input data are mapped into the high-dimensional space, linear functions with constraints on complexity (i.e. hyperplanes) are used to discriminate the inputs. A quadratic optimization problem must then be solved to determine the parameters of these functions. However, for high-dimensional feature spaces, the large number of parameters makes this problem intractable. For this reason, duality theory of optimization is used in SVM to make the parameter estimation in the high-dimensional feature space computationally affordable. The linear approximation function corresponding to the solution of the dual problem is given in the kernel representation, $K(x, x')$, and it is called the optimal separating hyperplane.

The solution in the kernel representation is written as a weighted sum of the support vectors. Fig. 2 shows the

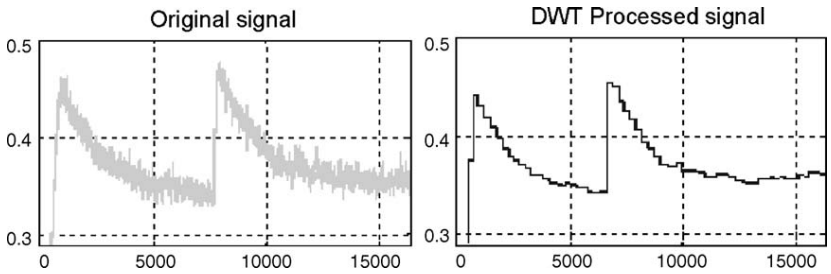


Fig. 1. Original and processed signal by DWT (the parameters of DWT are: Haar and app. coeff. at level eight.).

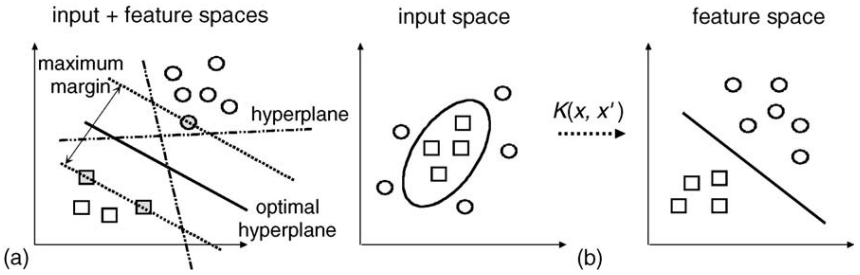


Fig. 2. The idea of SVMs: map the training data into a higher-dimensional feature space via K, and construct a separating hyperplane with maximum range there. This yields a non-linear decision boundary in input space. By the use of kernel functions, it is possible to compute the separating hyperplane without explicitly carrying out the map into the feature space: (a) linearly separable case and (b) non-linearly separable case.

SVM method, the data points at the margin (indicated in grey) are the support vectors.

4. Similarity query methods

Having reduced the search space by the use of the classification system, the last step consists of finding the signals most similar to the input signal. Two different methods can be applied for this purpose: Euclidean distance and bounding envelope.

4.1. Euclidean distance

A simple approach to determine possible similarities between two time series is to compute the Euclidean distance between them. Then, both series will be similar if the distance is less than some user-defined threshold. With our approach, waveform similarity is taken to mean similarity of characteristic values (features), and comparisons are made using these values. For example, DWT will be applied to extract the characteristics of waveform data and the first 0–k coefficients obtained are regarded as its characteristic values. So the multi-dimensional Euclidean space length L^2 between two points, X and Y, is a measurement of their similarity (Eq. (1)).

$$L^2(X - Y) = |X(0, \ldots, k) - Y(0, \ldots, k)|^2 \qquad (1)$$

4.2. Bounding envelope

A bounding envelope is based on the construction of two bounds around a signal (upper and lower bounds). A simple way to obtain these bounds consists in adding a user-defined threshold ($\pm\Delta y$) to each sample of the signal. To search for signals in the database most similar to the reference signal, we count the number of samples, which are outside the bounds. The signal with a minimum number of samples outside the bounds will be the most similar signal. Fig. 3 shows an example.

In our case, after SVM has performed a coarse filter to reduce the search space, feature extraction methods were used to apply the bounding envelope technique.

5. Application to the TJ-II database

This procedure has been applied to two different kinds of signals. In the first case, we have chosen raw data. The selected waveforms are described in Table 1.

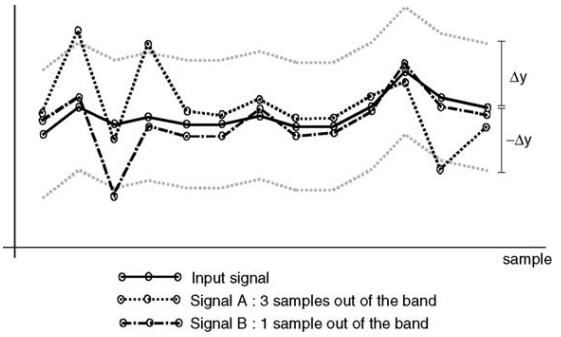


Fig. 3. An example of bounding envelope.

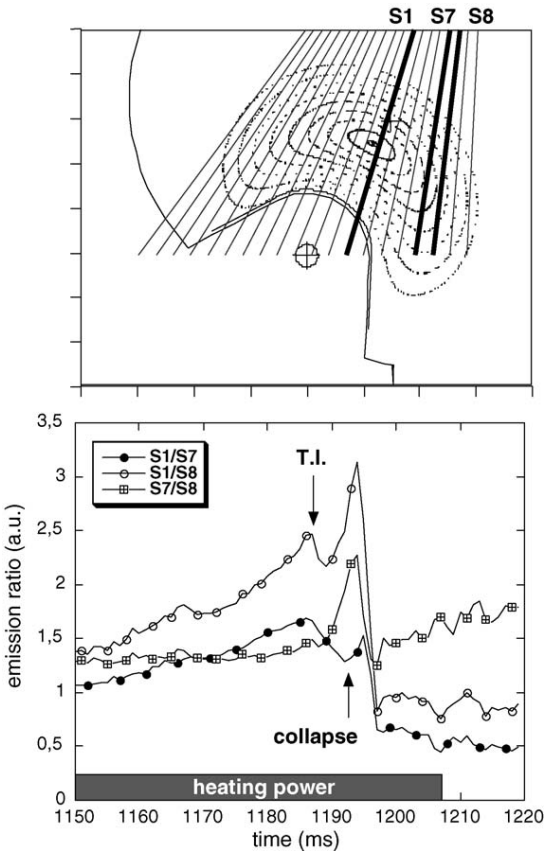


Fig. 4. An example of processed signals.

In the second case, processed signals are considered. In particular, we chose elaborated data connected with the thermal collapse, a universal behaviour in plasmas close to their density limit.

Table 1
Class of signals from TJ-II database

Class	Description
HALFAC3	Hα
DENSIDAD2_	Line averaged electron density
BOL5	Bolometer signal
RX306	Soft X-ray

The radiative collapse signatures can be recognised using a few pre-processed radiation signals: the ratios of three line integrated plasma emissions S1, S7 and S8 (Fig. 4).

The increase of the edge radiation and its propagation to the centre is seen as the decrease of the ratio S1/S8 followed by the decrease of S1/S7 together with the increase of S7/S8. The ratios are calculated from the raw signals.

6. Results

In order to show the overall procedure, the obtained results have been divided in two parts.

The first consists in validating SVM as a technique to reduce the search space. Fig. 5 shows the percentage of hits, misses and non-classifiable signals, with two different kernels, for the four types of raw signals described in Table 1.

This method has also been used with processed signals to classify collapse and non-collapse signals. In this case, the hit rate is about 80%.

Secondly, similarity query methods have been used to find the four most similar signals to a reference one

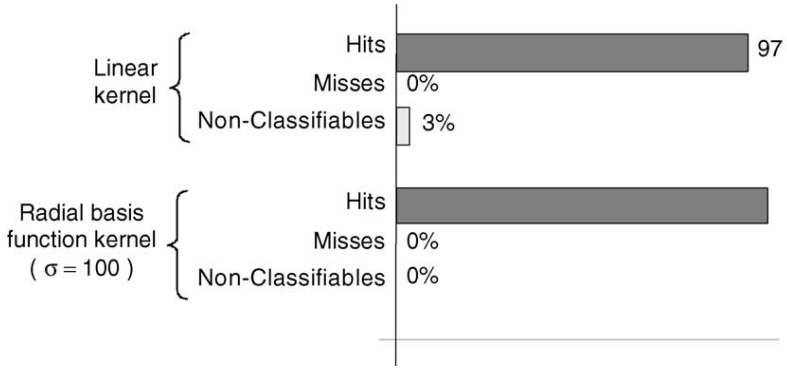


Fig. 5. SVM results.

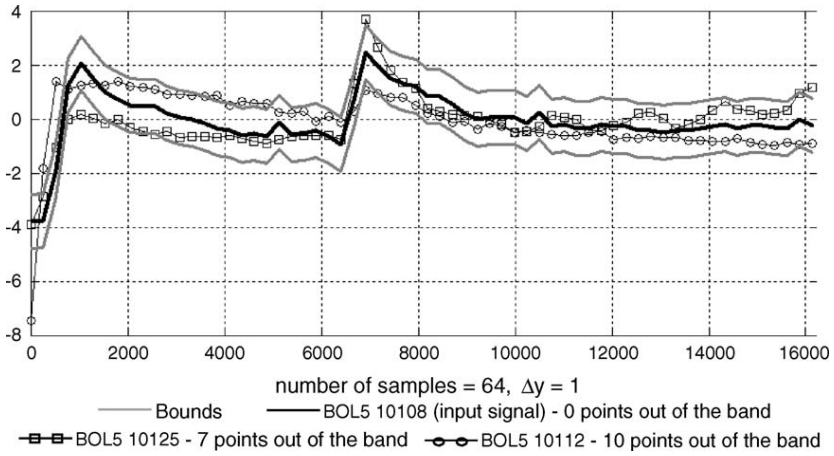


Fig. 6. Bounding envelope results.

from the database. Fig. 6 shows the results when the bounding envelope method is used.

7. Conclusions

In this paper, an automated technique to search for and to retrieve similar time evolution signals to a reference waveform is described. Specifically, it has been applied to time evolution signals of the TJ-II stellarator.

From the analysis of results, the bounding envelope method results to be a more robust technique than Euclidean distance. This is due to the accumulated error with Euclidean distance, in the weights it gives to more distant points. The bounding envelope method considers more distant points as outliers independent of their values.

References

[1] R. Duda, P. Hort, D. Stork, Pattern Classification, second ed., A Wiley-Interscience Publication, 2001.
[2] H. Nakanishi, T. Hochin, M. Kojima, LABCOM group, Search and retrieval methods of similar plasma waveforms, Fusion Eng. Des. 71 (2004) 189–193.
[3] S. Dormido-Canto, G. Farias, R. Dormido, J. Vega, J. Sánchez, M. Santos, The TJ-II Team, TJ-II wave forms analysis with wavelets and support vector machines, Rev. Sci. Instrum. 75 (10) (2004) 4254–4257.
[4] I. Daubechies, Ten lectures on wavelets, in: CBMS-NSF Regional Conference Series in Applied Mathematics, 1992, ISBN 0-89871-274-2.
[5] S. Mallat, A Wavelet Tour of Signal Processing, second ed., Academic Press, 2001.
[6] V. Vapnik, The Nature of Statistical Learning Theory, second ed., Springer, 2000.

Article 16

Automated clustering procedure for TJ-II experimental signals

16.1 Bibliographic Description

Title

Automated clustering procedure for TJ-II experimental signals.

Citation

N. Duro, J. Vega, R. Dormido, G. Farias, S. Dormido-Canto, J. Sánchez, M. Santos, G. Pajares (2006) Automated Clustering Procedure for TJ-II Experimental Signals, *Fusion Engineering and Design*, ISSN 0920-3796, Volume 81, Pages 1987-1991, Ed. Elsevier.

Abstract

Databases in fusion experiments are made up of thousands of signals. For this reason, data analysis must be simplified by developing automatic mechanisms for fast search and retrieval of specific data in the waveform database. In particular, a method for finding similar waveforms would be very helpful. The term similar implies the use of proximity measurements in order to quantify how close two signals are. In this way, it would be possible to define several categories (clusters) and to classify the waveforms according to them, where this classification can be a starting point for exploratory data analysis in large databases. The clustering process is divided in two stages. The first

one is feature extraction, i.e., to choose the set of properties that allow us to encode as much information as possible concerning a signal. The second one establishes the number of clusters according to a proximity measure.

References

H. Nakanishi, T. Hochin, M. Kojima (2004); G. Farias et al. (2004); S.G. Mallat (1999); K. Grochening, W.R. Madych (1992); S.C. Johnson (1967); B. MacQueen (1967); G. Carpenter, S. Grossberg (1987); W.L. Martínez, A.R. Martínez (2002).

Impact Factor

Fusion Engineering and Design has an impact factor of 1.49 according to Thomson Reuters Journal Citation Reports (2011).

Available online at www.sciencedirect.com
SCIENCE DIRECT®
Fusion Engineering and Design 81 (2006) 1987–1991
Fusion Engineering and Design
www.elsevier.com/locate/fusengdes

Automated clustering procedure for TJ-II experimental signals

N. Duro [a,*], J. Vega [b], R. Dormido [a], G. Farias [a], S. Dormido-Canto [a], J. Sánchez [a], M. Santos [c], G. Pajares [c]

[a] *Dpto. de Informática y Automática-UNED, 28040 Madrid, Spain*
[b] *Asociación EURATOM/CIEMAT para Fusión, 28040 Madrid, Spain*
[c] *Dpto. De Arquitectura de Computadores y Automática UCM, 28040 Madrid, Spain*

Available online 12 May 2006

Abstract

Databases in fusion experiments are made up of thousands of signals. For this reason, data analysis must be simplified by developing automatic mechanisms for fast search and retrieval of specific data in the waveform database. In particular, a method for finding similar waveforms would be very helpful. The term 'similar' implies the use of proximity measurements in order to quantify how close two signals are. In this way, it would be possible to define several categories (clusters) and to classify the waveforms according to them, where this classification can be a starting point for exploratory data analysis in large databases. The clustering process is divided in two stages. The first one is feature extraction, i.e., to choose the set of properties that allow us to encode as much information as possible concerning a signal. The second one establishes the number of clusters according to a proximity measure.

Keywords: Clustering; Feature extraction; TJ-II signals

1. Introduction

Diagnostics provide temporal evolution signals that translate plasma physical properties. In general, similar signals correspond to similar plasma behaviours and, therefore, it is possible to state the existence of patterns. Each kind of signal (density, temperature, soft X-rays, bolometry, etc.) allows the analysis of partial aspects of the plasma. Thus, a method for finding similar waveforms for each kind of signal would be very helpful to reveal, in an automated way, the set of discharges that show comparable behaviours. This is a pattern recognition problem. The goal of the problem is to classify waveforms into a number of categories (or clusters) and to apply proximity measures to evaluate the similarity between waveforms. The clustering method is responsible for 'revealing' the organisation of signals into 'sensible' clusters. This is called unsupervised clustering (UC).

* Corresponding author. Tel.: +34 913987169; fax: +34 91 3986697.
E-mail address: nduro@dia.uned.es (N. Duro).

0920-3796/$ – see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.fusengdes.2006.04.012

The application of pattern recognition techniques to fusion databases can help to build very useful tools for automated analysis and also for fast data search and retrieval [1,2]. The first reference is based on Fourier analysis and recognises slow varying waveforms or even waveforms with, at most, one major frequency component. The second reference shows a waveform pattern recognition technique based on wavelet transforms and support vector machines. The present article analyses several clustering criteria. It does not discuss similarity measures, however, we have used temporal evolution signals of the TJ-II stellarator. All computations were performed by developing several software tools based on MATLAB.

Section 2 describes the signal feature extraction; Section 3 explains the four methods that we have used; Section 4 shows results and, finally, some conclusions.

2. Feature extraction

A classification process begins choosing the set of characteristics to represent the signals (features). Temporal evolution signals in the TJ-II database have a very high number of samples (tens of thousands). In a first approximation we can use the samples themselves as features. However, to avoid such high dimensionality, which can lead to computational problems, some signal pre-processing is required.

Actually, the procedure consists of two phases: signal conditioning and feature extraction. The first one selects the same discharge interval for all waveforms and generates signals with the same number of samples and equal sampling period. This step is accomplished by using a cubic spline interpolation method. The second phase extracts only the important information of the signals while discarding noise and removing correlations. In this paper we have used two popular feature extraction techniques for time series: the Discrete Fourier Transform (DFT), with time complexity $O(n\log n)$, and the Discrete Wavelet Transform (DWT), with time complexity $O(n)$ [3].

A simple and commonly used wavelet is the Haar [4]. It has been chosen for the following reasons: (1) it allows good approximation with a subset of coefficients, (2) it can be computed quickly and easily, requiring linear time in the length of the signal and simple coding and (3) it preserves Euclidean distance.

3. Clustering methods

After feature extraction, the cluster analysis must group the signals into subsets (clusters). Two or more signals belong to the same cluster if they are 'close' according to a given similarity criteria, for instance, geometrical distance.

Our signals were analysed with the different techniques explained below.

3.1. Hierarchical

Given a set of N items to be clustered, and an $N \times N$ distance matrix, the basic process of hierarchical clustering [5] is:

1. Start by assigning each item to a cluster, so that if you have N items, you now have N clusters, each containing just one item. Let the distances (similarities) between the clusters be the same as the distance (similarities) between the distances they contain.
2. Find the closets (most similar) pair of clusters and merge them into a single cluster, so that now you have one cluster less.
3. Compute distances (similarities) between the new cluster and each of the old clusters.
4. Repeat steps 2 and 3 until all items are clustered into a single cluster of size N.

3.2. K-means

K-means [6] follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori.

The algorithm is composed of the following steps:

1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
2. Assign each object to the group that has the closest centroid.
3. When all objects have been assigned, recalculate the positions of the K centroids.

4. Repeat steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimised can be calculated.

3.3. Adaptive resonance theory

Adaptive resonance theory (ART) is applied to artificial neural network to develop a kind of competitive learning neural net. In this case when the information is presented to the input just one output neuron is activated. The idea is to *resonate* the input information with prototypes of classes those the net recognises. The algorithm for a neural network of type ART is described below [7]:

1. Input information is presented to net.
2. Input layer send to output layer through all connections.
3. All output neurons compete until just one is activated.
4. The neuron with the least Euclidean distance to the prototype of class will be the winner.
5. The prototype of class is compared with input information to obtain a similarity relation.
6. If similarity relation is less than a vigilance parameter (defined by the user) then the prototype of class is appropriate for the input information. Or else the input information will be a new prototype of class.

3.4. Grand tour

The grand tour (GT) of a multi-dimensional data set is an interactive visualisation technique for examining structure of high dimensional data using dynamic graphics. The idea is to project the *n*-dimensional data to a plane and to rotate the plane through all possible angles, searching for 'structure' in the data. 'Structure' is defined to be departure from normality [8] and includes such things as clusters, linear structures, holes and outliers.

4. Results

Unsupervised clustering techniques were applied to four different kinds of signals from the TJ-II database. We selected waveforms of 194 discharges corresponding to Hα emission, line average electron density, bolometer and soft X-ray signals. All the signals were pre-processed in order to be able to analyse the data within the same time window (258 ms) with identical sampling period (10 μs).

First of all, we tried to perform the classification process without feature extraction, i.e., by using all signal samples. However, computation time is extremely high in the GT method.

Feature extraction with DWT was carried out with the Haar wavelet and a decomposition level of 8, which generates a set of 64 coefficients. With the DFT, the set of characteristics was made up of the 24 first Fourier coefficients.

Therefore, each clustering process was initiated with 194 waveforms of the same signal (Hα, density, bolometer or X-ray signal) and 64 or 24 features each signal, depending on the characteristics extractor.

Each method (hierarchical, *K*-means, ART and GT) gives a set of clusters in each UC process. However, we considered only the clusters that included at least a 5% of the waveforms (10 signals). Clusters with less than 10 points are grouped together in a miscellaneous cluster.

Table 1 summarises the results of the above unsupervised clustering methods with feature extraction: wavelets (WT) and Fourier coefficients (FT). The table shows the number of clusters found and the percentage of signals included in each one. First of all, it must be pointed out that equivalent results are obtained without feature extraction. In particular, using Hierarchical, ART and GT the percentage obtained are very similar, not only in the main clusters but also in the miscellaneous cluster. It can be noted that at least the 50% of signals belong to the same cluster.

Our signals were also analysed using another clustering method called Projection Pursuit [8]. To determine the number of clusters using this method is a very difficult task. Nevertheless, to identify those signals which are very different from the rest can be accomplished with low effort.

The inspection of the *K*-means results shows that it produces a different behaviour: more number of clusters is generated. Besides, the number of signals in each cluster is smaller. Analyzing the signals that constitute these clusters it can be concluded that the signals for two or three clusters (depending on the experience) in

Table 1
A comparison of the unsupervised clustering methods with feature extraction

Signal type	Hierarchical	WT	FT	ART	WT	FT	G. tour	WT	FT	*K*-means	WT	FT
Bolometer	First cluster	60%	58%	First cluster	47%	54%	First cluster	60%	58%	First cluster	32%	22%
	Second cluster	12%	22%	Second cluster	33%	23%	Second cluster	23%	24%	Second cluster	23%	21%
	Third cluster	11%	12%	Third cluster	12%	10%	Third cluster	11%	12%	Third cluster	14%	17%
	Miscellan.	17%	8%	Miscellan.	8%	13%	Miscellan.	7%	6%	Fourth cluster	10%	15%
										Miscellan.	21%	25%
Density	First cluster	74%	78%	First cluster	50%	53%	First cluster	70%	68%	First cluster	24%	24%
	Second cluster	10%	8%	Second cluster	28%	36%	Second cluster	12%	19%	Second cluster	21%	21%
	Miscellan.	16%	14%	Third cluster	10%	–	Third cluster	8%	–	Third cluster	16%	20%
				Miscellan.	12%	11%	Miscellan.	10%	13%	Fourth cluster	16%	12%
										Miscellan.	23%	23%
Soft X-ray	First cluster	80%	89%	First cluster	82%	80%	First cluster	77%	81%	First cluster	36%	41%
	Second cluster	14%	5%	Second cluster	13%	15%	Second cluster	19%	7%	Second cluster	20%	16%
	Miscellan.	6%	6%	Miscellan.	5%	5%	Miscellan.	4%	12%	Third cluster	13%	15%
										Fourth cluster	10%	10%
										Miscellan.	21%	18%
Hα	First cluster	55%	60%	First cluster	50%	50%	First cluster	79%	72%	First cluster	24%	25%
	Second cluster	29%	29%	Second cluster	35%	39%	Second cluster	11%	14%	Second cluster	19%	17%
	Miscellan.	16%	11%	Miscellan.	15%	11%	Third cluster	8%	11%	Third cluster	15%	17%
							Miscellan.	2%	3%	Fourth cluster	14%	13%
										Miscellan.	28%	28%

the *K*-means method are integrated into a bigger cluster for the other three methods.

5. Conclusions

Clustering results in TJ-II show that, typically, most waveforms of a signal family are grouped into one big cluster, but there also appear a reduced number of clusters with few signals. The several methods group the same signals into the same clusters, independently on features. Thus, the problem for finding the most similar waveforms to a given one can be solved very efficiently: a pattern recognition system classifies the initial signal into one of the known clusters and the most similar signals can be obtained by means of proximity measures.

In addition, the results can simplify the search of interesting data in TJ-II. Roughly speaking, all families provide two clusters. Firstly, the big one symbolises that most signals translate an average physical behaviour of the measured plasma property. Secondly, the rest of the waveforms can be integrated into a single cluster. The latter includes non-average behaviours and, therefore, signals classified in this group reveal non-standard plasma properties. This fact helps diagnosticians because they can find, in an automated way, interesting data to be analysed, instead of having to search for them manually.

Acknowledgements

The authors wish to thank Prof. Sebastián Dormido Bencomo (UNED) and Prof. Jesús Manuel de la Cruz (UCM) for their constructive comments and invaluable guidance.

References

[1] H. Nakanishi, T. Hochin, M. Kojima, LABCOM group. Search and retrieval methods of similar plasma waveforms, Fus. Eng. Des. 71 (2004) 189.
[2] G. Farias, R. Dormido, J. Vega, J. Sánchez, M. Santos, TJ-II waveform analysis with wavelets and support vector machines, Rev. Sci. Ins. 75 (10) (2004) 4254–4257.
[3] S.G. Mallat, A Wavelet Tour of Signal Processing, Academic Press, 1999.
[4] K. Grochening, W.R. Madych, Multiresolution analysis, Haar bases, and self-similar tiling of r^n, IEEE Trans. Inf. Theory 38 (2) (1992) 556–568.

[5] S.C. Johnson, Hierarchical Clustering Schemes, vol. 2, Psychometrika, 1967, pp. 241–254.
[6] B. MacQueen, Some methods for classification and analysis of multivariate observations, in: Proceedings of Fifth Berkeley Symp. Math. Statistics and Probability, vol. 1, Berkeley, Univ. California Press, 1967, pp. 281–297.
[7] G. Carpenter, S. Grossberg, ART2: self-organization of stable category recognition codes for analog input patterns, in: Proceedings of IEEE First International Conference on Neural Networks, vol. II, 1987, p. 727.
[8] W.L. Martinez, A.R. Martínez, Computational Statistics Handbook with Matlab, Chapman & Hall, CRC, 2002.

Article 17

Information retrieval with wavelets and support vector machines

17.1 Bibliographic Description

Title

Information retrieval and classification with wavelets and support vector machines.

Citation

S. Dormido Canto, J. Vega, Sánchez, G. Farias (2005) Information Retrieval and Classification with Wavelets and Support Vector Machines, *Lecture Notes in Computer Science*, ISSN 0302-9743, Volume 3562, Part 2, Pages 548-557, Springer-Verlag.

Abstract

Since fusion plasma experiment generates hundreds of signals. In analyzing these signals it is important to have automatic mechanisms for searching similarities and retrieving of specific data in the waveform database. Wavelet transform (WT) is a transformation that allows to map signals to spaces of lower dimensionality, that is, a smoothed and compressed version of the original signal. Support vector machine (SVM) is a very effective method for general purpose pattern recognition. Given a set of input vectors

which belong to two different classes, the SVM maps the inputs into a high-dimensional feature space through some non-linear mapping, where an optimal separating hyperplane is constructed. This hyperplane minimizes the risk of misclassification and it is determined by a subset of points of the two classes, named support vectors (SV). In this work, the combined use of WT and SVM is proposed for searching and retrieving similar waveforms in the TJ-II database. In a first stage, plasma signals will be preprocessed by WT in order to reduce the dimensionality of the problem and to extract their main features. In the next stage, and using the new smoothed signals produced by the WT, SVM will be applied to show up the efficency of the proposed method to deal with the problem of sorting out thousands of fusion plasma signals.

References

D. Radiei, A. Mendelzon (1998); H. Nakanishi, T. Hochin, M. Kojima (2003); S. Mallat (2001); M. Vetterli (2000); V. Vapnik (1995); V. Vapnik (1998); R.O. Duda, P.E Hart, D.G. Stork (2001); J.D. Sebald, J.A. Bucklew (2001); C. Alejaldre, et al. (1999); A. Haar (1910); K. Grochening, W.R. Madych (1992).

Impact Factor

Lectures Notes on Computer Science has a factor of 0.33 according to SCImago Journal Rank (SJR) (2011).

Information Retrieval and Classification with Wavelets and Support Vector Machines

S. Dormido-Canto[1], J. Vega[2], J. Sánchez[1], and G. Farias[1]

[1] Dpto. de Informática y Automática -
E.T.S.I. Informática - U.N.E.D. Madrid 28040, Spain
{sebas, jsanchez}@dia.uned.es, gfarias@bec.uned.es
[2] Asociación EURATOM/CIEMAT, Madrid 28040, Spain
jesus.vega@ciemat.es

Abstract. Since fusion plasma experiment generates hundreds of signals. In analyzing these signals it is important to have automatic mechanisms for searching similarities and retrieving of specific data in the waveform database. Wavelet transform (WT) is a transformation that allows to map signals to spaces of lower dimensionality, that is, a smoothed and compressed version of the original signal. Support vector machine (SVM) is a very effective method for general purpose pattern recognition. Given a set of input vectors which belong to two different classes, the SVM maps the inputs into a high-dimensional feature space through some non-linear mapping, where an optimal separating hyperplane is constructed. This hyperplane minimizes the risk of misclassification and it is determined by a subset of points of the two classes, named support vectors (SV). In this work, the combined use of WT and SVM is proposed for searching and retrieving similar waveforms in the TJ-II database. In a first stage, plasma signals will be preprocessed by WT in order to reduce the dimensionality of the problem and to extract their main features. In the next stage, and using the new smoothed signals produced by the WT, SVM will be applied to show up the efficency of the proposed method to deal with the problem of sorting out thousands of fusion plasma signals.

1 Introduction

Databases in nuclear fusion experiments are made up of thousands of signals. For this reason, data analysis must be simplified by developing automatic mechanisms for fast search and retrieval of specific data in the waveform database. In particular, a method for finding similar waveforms would be very helpful.

In [1] a method is proposed to find similar time sequences using *Discrete Fourier Transformation* (DFT) to reduce the dimensionality of the feature vectors, that is, to minimize the computation time for indexing and comparing signals. In [2], the previous DFT based method is used to search similar phenomena in waveform databases but just it is applied with slowly varying signals. However, the DFT has difficulties when used with fast varying waveforms since

J. Mira and J.R. Álvarez (Eds.): IWINAC 2005, LNCS 3562, pp. 548–557, 2005.

time information is lost when transforming to the frequency domain and non-stationary or transitory characteristics can not be detected. The *Short Time Fourier Transform* (STFT) can obtain the non-stationary characteristics using an analysis window. However, the precision is determined by the analysis window that is the same for all frequencies. *Wavelets* (WT) offers an efficient alternative to data processing and provides many advantages: 1) data compression, 2) computing efficiency, and 3) simultaneous time and frequency representation. Because of these characteristics, wavelets have a growing impact on signal processing applications [3, 4].

Support Vector Machines (SVM) is a very effective method for general purpose pattern recognition [5, 6, 7]. In a few words, given a set of input vectors which belong to two different classes, the SVM maps the inputs into a high-dimensional feature space through some non-linear mapping, where an optimal separating hyperplane is constructed in order to minimize the risk of misclassification. The hyperplane is determined by a subset of points of the two classes, named *Support Vectors* (SV). Several methods had been proposed to cope with multi-category classification [8].

In this work, preliminary results are shown when using WT techniques for characterizing the signals and SVM as the technique for pattern recognition and information retrieval. The proposed method has been applied to the TJ-II stellarator database. The TJ-II is a stellarator device [9] (heliac type, $B(0) \leq 1.2T$, $R(0) = 1.5m$, $\langle a \rangle \leq 0.22m$) located at CIEMAT (Madrid, Spain) that can explore a wide rotational transform range ($0.9 \leq iota/2\pi \leq 2.2$). TJ-II plasmas are produced and heated with ECRH (2 gyrotrons, 300 kW each, 53.2 GHz, 2nd harmonic, X-mode polarization) and NBI (300 kW). At present, 928 digitization channels are available for experimental measurements in the TJ-II.

2 Wavelet Transform

WT are basis functions used in representing data or other functions. Wavelet algorithms process data at different resolutions or decomposition levels in contrast with DFT where only frequency components are considered. The construction of the first orthonormal system by Haar [10] is an important milestone since the Haar basis is still a foundation of modern wavelet theory. In this work, the use of the Haar wavelets in the problem of extracting characteristic of the plasma signals will be considered.

The motivation for using the WT is to have a decomposition method that is fast to compute and requires little data storage for each signal. The Haar wavelet is chosen for many advantages: (1) it allows good approximation with a subset of coefficients, (2) it can be computed quickly and easily, requiring linear time in the length of the signal and simple coding, and (3) it preserves Euclidean distance. Concrete mathematical foundations can be found in [11]. In the WT with Haar base, there are two kinds of functions called *approximation function and difference function*. The approximation function generates a sequence of the averages between two adjacent values of the input sequence, that is, the sampled

signal. The difference function generates a sequence of the differences between two consecutive data in the current approximation sequence. These functions are applied recursively until the number of the elements in the difference sequence is one. That is, the ith approximation sequence A_i and difference sequence D_i are defined as follow:

$$A_i = \{\frac{A_{i-1}(1)+A_{i-1}(2)}{2}, \frac{A_{i-1}(3)+A_{i-1}(4)}{2}, \ldots, \frac{A_{i-1}(m-1)+A_{i-1}(m)}{2}\}$$

$$D_i = \{\frac{A_{i-1}(1)-A_{i-1}(2)}{2}, \frac{A_{i-1}(3)-A_{i-1}(4)}{2}, \ldots, \frac{A_{i-1}(m-1)-A_{i-1}(m)}{2}\}$$

where A_i(j) is the j-th element in the sequence A_i and m is the number of the elements in the sequence A_{i-1}. Next, a brief example of the Haar transformation of a discrete sequence $\overrightarrow{X} = \{9, 7, 4, 8, 5, 3, 8, 8\}$ is shown (Table 2).

Table 1. Example of the Haar transformation

Approximation	Averages	Coefficients (Differences)
8	{9, 7, 4, 8, 5, 3, 8, 8}	-
4	{8, 6, 4, 8}	{1, -2, 1, 0}
2	{7, 6}	{1, -2}
1	{6.5}	{0.5}

Approximation 8 is the full resolution of the discrete sequence. In approximation 4, {8, 6, 4, 8} are obtained by taking the average of {9, 7}, {4, 8}, {5, 3} and {8, 8} at resolution 8 respectively. The coefficients {1, -2, 1, 0} at resolution 4 are the differences of {9, 7}, {4, 8}, {5, 3} and {8, 8} divided by two respectively. This process is continued until an approximation of 1 is reached. The Haar transform $H(\overrightarrow{x}) = \{c, d_0^0\, d_0^1, d_1^1, d_0^2, d_1^2, d_2^2, d_3^2\}$ is obtained which is composed of the last average value 6.5 and the coefficients found on the right most column. It should be pointed out that c is the overall average value of the whole time sequence.

The reason of using Haar transform to replace DFT is based on several evidences. The first reason is on the pruning power. The nature of the Euclidean distance preserved by Haar transform and DFT are different. In DFT, comparison of two time sequences is based on their low frequency components, where most energy is presumed to be concentrated on. On the other hand, the comparison of Haar coefficients is matching a gradually refined resolution of the two time sequences. Another reason is the complexity consideration. The complexity of Haar transform is *O(n)* whilst *O(n log n)* computation is required for DFT. Both impose restriction on the length of time sequences which must be an integral power of 2. Although these computations are all involved in pre-processing stage, the complexity of the transformation can be a concern especially when the database is large, as happens in our case. Another advantage of using WT is the multi-resolution representation of signals since it has the time-frequency localization property. Thus, WT is able to give locations in both time and frequency.

Therefore, wavelet representations of signals bear more information than that of DFT, in which only frequencies are considered. While DFT extracts the lower harmonics which represent the general shape of a time sequence, WT encodes a coarser resolution of the original time sequence with its preceding coefficients.

Fig. 1 shows the WT is applied to the original signals in order to compute a few coefficients for each signal in a fast way.

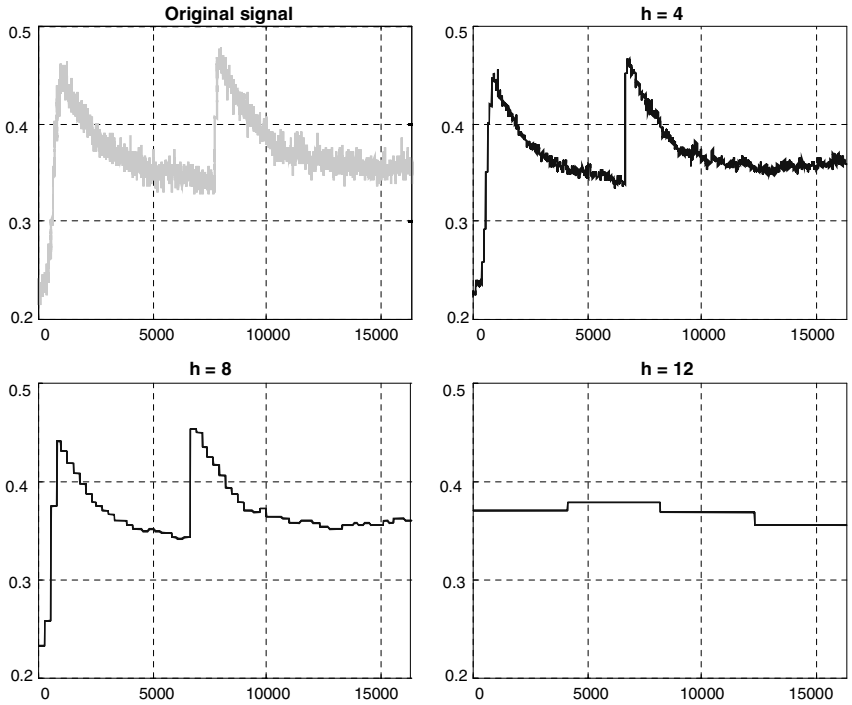


Fig. 1. Original signal and its Wavelet transform approximations with three different decomposition levels (h=4, 8, 12)

3 Support Vector Machines for Classification

The support vector machine (SVM) is a universal constructive learning procedure based on the statistical learning theory [5]. The SVM maps input data into a high-dimensional space using a non-linear function. Once input data are mapped into the high-dimensional space, linear functions with constraints on complexity (i.e., hyperplanes) are used to discriminate the inputs, and a quadratic optimization problem must be solved to determine the parameters of these functions. Nevertheless for high-dimensional feature spaces, the large number of parameters makes this problem intractable. For this reason, duality theory of optimization is used in SVM to make the estimation of parameters in the high-dimensional feature space computationally affordable. The linear approximation function corresponding to the solution of the dual problem is given in the kernel representation and it is called the optimal separating hyperplane. The solution in the kernel

representation is written as a weighted sum of the support vectors, that is, a subset of the training data. Let's explain how to obtain the optimal separating hyperplane.

A separating hyperplane is a linear function that is capable of separating the training data without error. Consider the problem of separating the set of training vectors belonging to two separate classes (a binary classifier),

$$\{(x_1, y_1), \ldots, (x_n, y_n)\}, \quad x \in \mathbb{R}, \quad y \in \{+1, -1\}$$

with a hyperplane decision function $D(x)$,

$$D(x) = < w, x >$$

where $< \cdot, \cdot >$ denotes inner product. In linearly separable cases, SVM constructs a hyperplane which separates the training data without error. The hyperplane is constructed by finding another vector w and a parameter b that minimizes $\|w\|^2$ and satisfies the following conditions:

$$y_i = [< w, x > +b] \geq 1, \quad i = 1, \ldots, n$$

where w is a normal weight vector to the hyperplane, $|b|/\|w\|$ is the perpendicular distance from the hyperplane to the origin, and $\|w\|^2$ is the Euclidean norm of w. After the determination of w and b, a given vector x can be classified by:

$$sign(< w, x > +b) \ . \tag{1}$$

In non-linearly separable cases, SVM can map the input vectors into a high dimensional feature space. By selecting a non-linear mapping a priori, SVM constructs and optimal separating hyperplane in this higher dimensional space. A kernel function $K(x, x')$ performs the non-linear mapping into feature space [7], and the original constrains are the same. In this way, the evaluation of the inner products among the vectors in a high-dimensional feature space is done indirectly via the evaluation of the kernel between support vectors and vectors in the input space (Fig. 2).

This provides a way of addressing the technical problem of evaluating inner products in a high-dimensional feature space. Examples of kernel functions are shown in Table 2.

Linear support vector machine is applied to this feature space and then the decision function is given by Eq. 2:

$$f(x) = sign(\sum_{i \in SV} \alpha_i y_i K(x_i, x) + b) \ . \tag{2}$$

where the coefficients α_i and b are determined by maximizing the following Langrangian expression:

$\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j)$, where: $\alpha_i \geq 0$ and $\sum_{i=1}^{n} \alpha_i y_i = 0$

A positive or negative value from Eq. 1 or Eq. 2 indicates that the vector x belongs or not to class 1. The data samples for which the are nonzero are the support vectors. The parameter b is given by:

$$b = y_s \sum_{i \in SV} \alpha_i y_i K(x_s, x_i)$$

where (x_s, y_s) is any one of the support vectors.

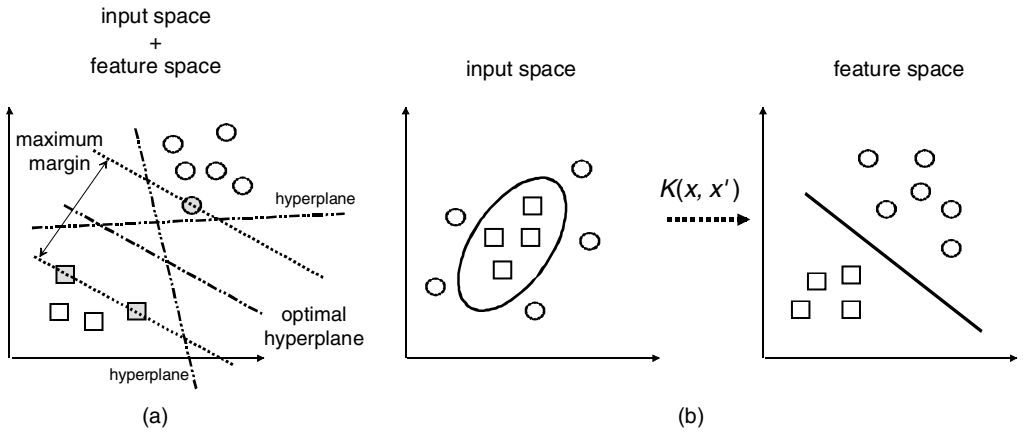


Fig. 2. The idea of SVMs: map the training data into a higher-dimensional feature space via K, and construct a separating hyperplane with maximum range there. This yields a nonlinear decision boundary in input space. By the use of kernel functions, it is possible to compute the separating hyperplane without explicity carrying out the map into the feature space. (a) Linearly separable case. (b) Non-linearly separable case

Table 2. Kernel functions extensively used

Kernel Function	Description
Inner product	$K(x, x') =< x, x' >$
Polynomial of degree d	$K(x, x') = (< x, x' > +1)^d$
Gaussian Radial Basis Function	$K(x, x') = \exp\{-\|x - x'\|^2/2\sigma^2\}$
Exponential Gaussian Radial Basis Function	$K(x, x') = \exp\{-\sqrt{\|x - x'\|^2}/2\sigma^2\}$

4 Performance Evaluation

Some preliminary results of our pattern classification approach based on wavelets and SVM are presented in this Section. We have focused the attention in showing the method validity instead of looking for a specific application. Our proof was based on classifying and recognizing temporal evolution signals from the TJ-II database. It is accomplished in a two-step process. A first step provides signal conditioning (Fig. 3), to ensure the same sampling period and number of samples.

This requirement is a consequence of the fact that signals could have been collected with different acquisition parameters. A second step is devoted to perform,

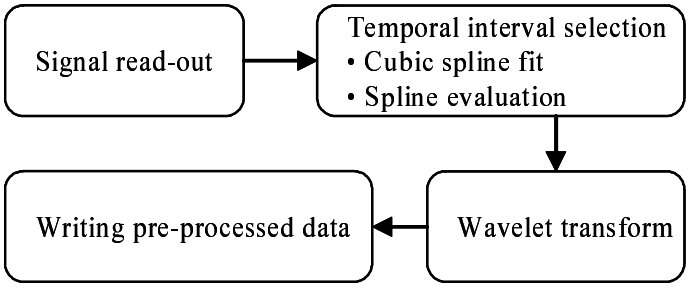


Fig. 3. Signal conditioning data flow

firstly, the learning process with SVM and some of the pre-processed data. Secondly, classification tasks are carried out. All processes have been programmed from the MATLAB software package. In order to evaluate the approach, two experiments have been carried out to classify signals stored in the TJ-II database. These signals belong to one of the classes shown in Table 3.

Table 3. Classes of signals of the TJ-II database

Classes	Description
BOL5	Bolometer signal
ECE7	Electron ciclotron emission
RX306	Soft x-ray
ACTON275	Espectroscopic signal (CV)
HALFAC3	Hα
Densidad2	Line averaged electron density

In the first stage of our approach, the signals are preprocessed in both of our experiments by Haar transform (with a decomposition level of 8) to reduce the dimensionality of the problem. In the second stage, the test signals are classified using SVM.

The method applied is one versus the rest, that allows to get multi-class classifiers. For that reason, we construct a set of binary classifiers as it is explained in Section III. Each classifier is trained to separate one class from the rest, and to combine them by doing the multi-class classification according to the maximal output before applying the sign function (Eq. 1). Next, two experiments are shown to demonstrate the viability of the proposed approach.

In the first experiment, 4 classes have been considered: ECE7, BOL5, RX306, and Densidad2. The training set is composed by 40 signals and the test set by 32 signals obtained from the TJ-II database.

The Fig. 4 displays the positive support vectors for each class using a linear kernel, the training signal corresponding to the original signal in TJ-II, and the wavelet approach which is the signal resampled to 16384 samples after the wavelet transform.

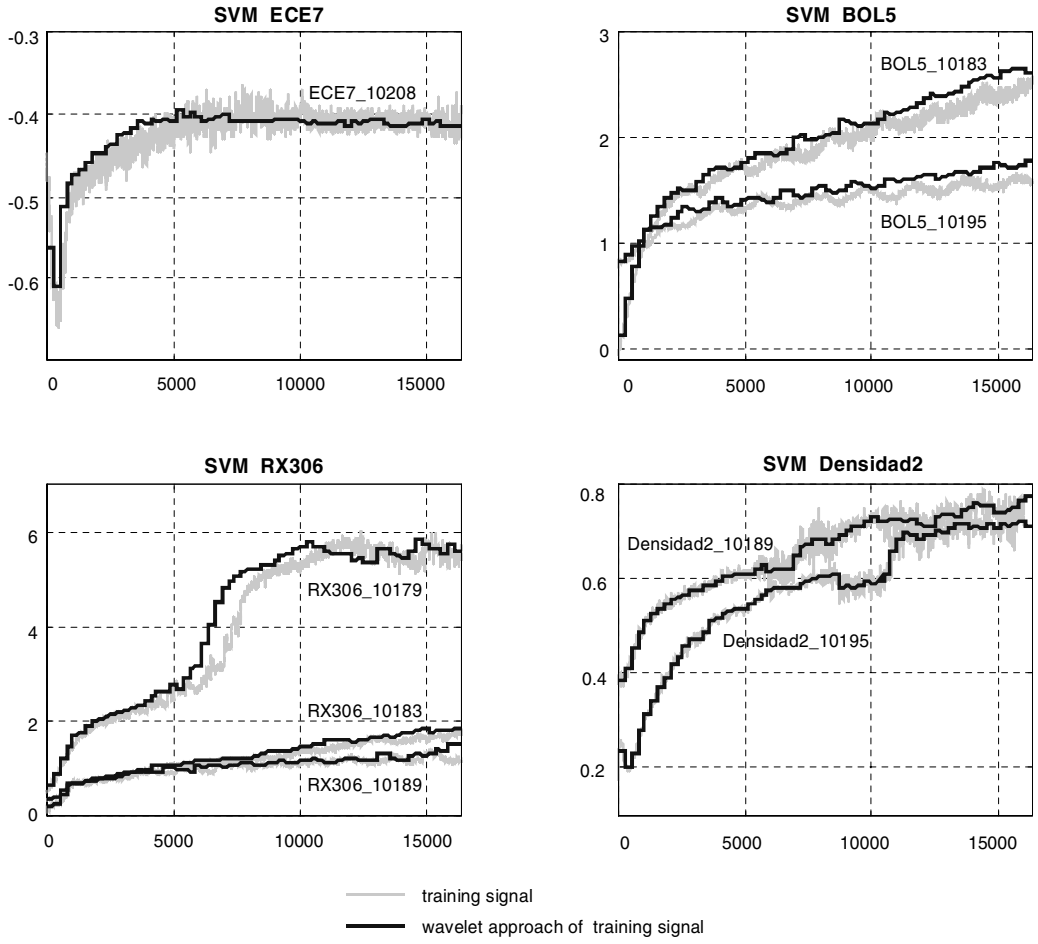


Fig. 4. Positive support vector for every class in the experiment 1

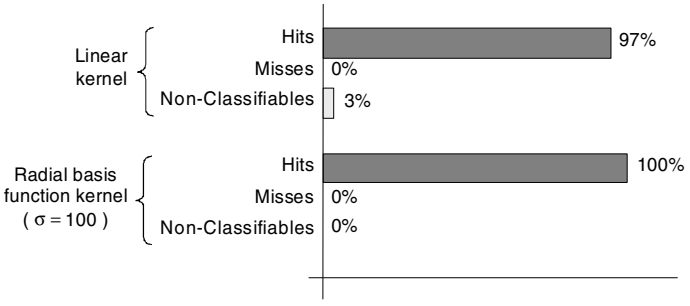


Fig. 5. Results of the experiment 1

The percentage of hits, misses, and non-classifiable signals are illustrated in Fig. 5.

In a second experiment, the training and test sets are composed by 60 and 48 signals and the number of classes was 6, respectively. Fig. 6 shows the results.

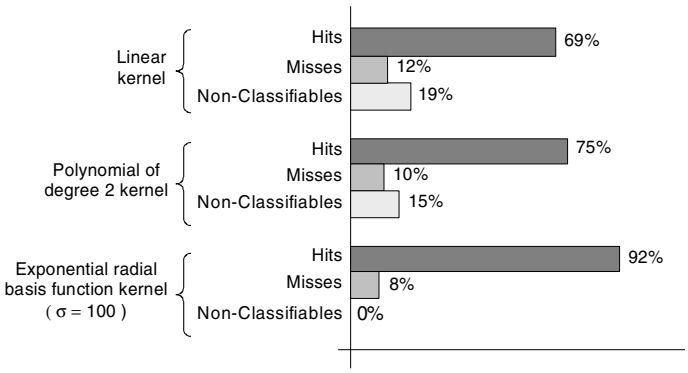


Fig. 6. Results of the experiment 2

5 Conclusions

In this paper, we present a new approach for the classification of plasma experiments signals. The method proposed here contains two processing stages, preprocessing of the original signals by Wavelet Transform (WT) and multi-class classification by Support Vector Machines (SVM). In the first stage, wavelet transformations are applied to signals to reduce the number of dimensions of the feature vectors. After that, a SVM-based multi-class classifier is constructed using the preprocessed signals as input space.

From observation of several experiments, our WT+SVM method is very viable and efficient time, and the results seem promising. However, we have further work to do. We have to finish the development of a Matlab toolbox for WT+SVM processing and to include new relevant features in the SVM inputs to improve the technique, even developing new kernel functions. We have also to make a better preprocessing of the input signals and to study the performance of other generic and self-custom kernels.

References

1. Radiei, D., Mendelzon, A.: Efficient Retrieval of Similar Time Sequences Using DFT. Proc. 5th Inter. Conf. on Foundations of Data Organization (FODO'98), (1998), 249-257
2. Nakanishi, H., Hochin, T., Kojima, M. and LABCOM group: Search and Retrieval Methods of Similar Plasma Waveforms. 4th IAEA TCM on Control, Data Acquisition, and Remote Participation for Fusion Research. San Diego. July 21-23, (2003)
3. Mallat, S.: A Wavelet Tour of Signal Processing. 2 Edition, Academic Press, (2001)
4. Vetterli, M.: Wavelets, Approximation and Compression. IEEE Signal Processing Magazine, (2000), 59-73
5. Vapnik, V.: The Nature of Statistical Learning Theory. Springer, (1995)
6. Vapnik, V.: Statistical Learning Theory. John Wiley & Sons, INC, (1998)

7. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification. 2 Edition, John Wiley, (2001)
8. Sebald, J.D., Bucklew, J.A.: Support Vector Machines and the Multiple Hypothesis Test Problems. IEEE Trans. on Signal Processing, vol. 49, no. 11, (2001), 2865-2872
9. Alejaldre, C. et al.: Plasma Phys. Controlled Fusion 41, 1 (1999), pp. A539
10. Haar A.: Theorie der orthogonalen funktionen-systeme. Mathematische Annalen, (1910) 69:331-371
11. Grochening, K., Madych, W.R.: Multiresolution analysis, Haar bases, and self-similar tilings of rn. IEEE Trans. of Information Theory, vol. 38, no 2, (1992), 556-568

Article 18

Image classifier for the TJ-II Thomson Scattering diagnostic

18.1 Bibliographic Description

Title

Image classifier for the TJ-II Thomson Scattering diagnostic: Evaluation with a feed forward neural network.

Citation

G. Farias, R. Dormido, M. Santos, N. Duro (2005) Image Classifier for the TJ-II Thomson Scattering Diagnostic: Evaluation with a Feed Forward Neural Network, *Lecture Notes in Computer Science*, ISSN 0302-9743, Volume 3562, Part 2, Pages 604-612, Springer-Verlag.

Abstract

There are two big stages to implement in a signal classification process: features extraction and signal classification. The present work shows up the development of an automated classifier based on the use of the Wavelet Transform to extract signal characteristics, and Neural Networks (Feed Forward type) to obtain decision rules. The classifier has been applied to the nuclear fusion environment (TJ-II stellarator), specifically to the Thomson Scattering diagnostic, which is devoted to measure density and

temperature radial profiles. The aim of this work is to achieve an automated profile reconstruction from raw data without human intervention. Raw data processing depends on the image pattern obtained in the measurement and, therefore, an image classifier is required. The method reduces the 221.760 original features to only 900, being the success mean rate over 90% .This classifier has been programmed in MATLAB.

References

C. Alejaldre et al. (1999); R. Duda, P. Hort, D. Stork (2001); H. Nakanishi, T. Hochin, M. Kojima (2003); S. Dormido-Canto, et al. (2004); I. Daubechies (1992); S. Mallat (2001); M. Misiti et al. (1998); G. Farias et al. (2004); J.R. Hilera, V.J. Martínez (1995); J. Freeman, D. Skapura (1993); H. Demuth, M. Beale (1998).

Impact Factor

Lectures Notes on Computer Science has a factor of 0.33 according to SCImago Journal Rank (SJR) (2011).

Image Classifier for the TJ-II Thomson Scattering Diagnostic: Evaluation with a Feed Forward Neural Network

G. Farias[1], R. Dormido[1], M. Santos[2], and N. Duro[1]

[1] Dpto. de Informática y Automática -
E.T.S.I. Informática - U.N.E.D. Madrid 28040, Spain
gfarias@bec.uned.es, {raquel, nduro}@dia.uned.es
[2] Dpto. de Arquitectura de Computadores y Automática -
Facultad de CC. Fisicas - Universidad Complutense, Madrid 28040, Spain
msantos@dacya.ucm.es

Abstract. There are two big stages to implement in a signal classification process: features extraction and signal classification. The present work shows up the development of an automated classifier based on the use of the Wavelet Transform to extract signal characteristics, and Neural Networks (Feed Forward type) to obtain decision rules. The classifier has been applied to the nuclear fusion environment (TJ-II stellarator), specifically to the Thomson Scattering diagnostic, which is devoted to measure density and temperature radial profiles. The aim of this work is to achieve an automated profile reconstruction from raw data without human intervention. Raw data processing depends on the image pattern obtained in the measurement and, therefore, an image classifier is required. The method reduces the 221.760 original features to only 900, being the success mean rate over 90%. This classifier has been programmed in MATLAB.

1 Introduction

The TJ-II is a medium-size stellarator (heliac type) [1] located at CIEMAT (Spain). The Thomson Scattering (TS) in plasmas consists in the re-emission of incident radiation (from very powerful lasers) by free electrons. Electron velocity distribution generates a spectral broadening of the scattered light (by Doppler effect) related to the electronic temperature. The total number of scattered photons is proportional to the electronic density.

Every laser shot produces a bi-dimensional image from which is possible to obtain radial profiles of temperature and density. Only a restricted number of pattern images appear in the TJ-II. They represent different physical situations related to either the plasma heating or the system calibration. Depending on the pattern obtained, data are processed in different ways. Therefore, to perform an automated data analysis, a computerized classification system must provide the kind of pattern obtained in order to execute the proper analysis routines.

J. Mira and J.R. Álvarez (Eds.): IWINAC 2005, LNCS 3562, pp. 604–612, 2005.

As in any classification process, Thomson Scattering images need to be preprocessed in a suitable way [2]. Most of the analyses try to extract either unique or common signal features, thereby allowing identification of patterns that reflect similar experimental conditions [3, 4].

The present work shows up the development of an automated classifier (programmed in MATLAB) made up of two phases. The first one (feature extraction) uses the Wavelet Transform, and the second one (classification) makes use of Multilayer Neural Networks (Feed Forward type).

1.1 Image Patterns

The TJ-II Thomson Scattering images can be grouped under five different classes (Fig. 1).

Table 1 shows a brief description corresponding to every pattern.

As it can be seen in Fig. 1, all the patterns except BKGND correspond to images with, at least, four important features: an empty zone in the middle, two central vertical components, and a thin line on the right. Without giving details about the physical meaning of these characteristics, the differences among the patterns are consequence of the light intensity: very high in the central

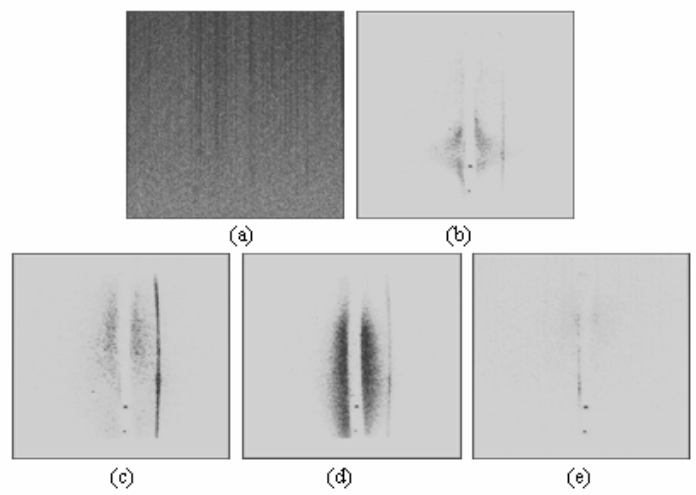


Fig. 1. Image patterns: (a) BKGND (b) COFF (c) ECH (d) NBI (e) STRAY

Table 1. Description of TJ-II Thomson Scattering patterns

Pattern	Description
BKGND	CCD Camera background
COFF	Reached cut off density for plasmas heated by Electron Cyclotron Resonant Heating
ECH	Plasmas heated by Electron Cyclotron Resonant Heating
NBI	Plasmas heated by Neutral Beam Injectors
STRAY	Measurement of parasitic light without plasma

components for the NBI case, low for the ECH case (although with a very intense thin line), central components grouped at the bottom for the COFF case, and practically null for the STRAY case.

2 Procedures for Data Mining and Information Retrieval

In this section effective feature extraction and classification methods for images are briefly illustrated. Firstly, a short review of the Wavelet Transform and its application to the signals is presented. Secondly, the Neural Networks technique used in the signal classification process is described. Finally, training and validation procedures of the classification method are commented.

2.1 Wavelet Transform

In many cases the signals present pseudo-periodic behaviour, oscillating around a fixed frequency. The most widely used analysis tool for periodic signals is the Fourier Transform, which allows to study time depended signals in the frequency domain. However, there are many other signals that can present a non-periodical behaviour, whose principal features must be obtained from a temporal analysis. For these signals, the Fourier Transform is unsuitable.

To analyze signals which present periodic and non-periodic behaviour is necessary to make use of transforms in the time-frequency plane. For such a reason the Wavelet Transform (WT), that overcomes the drawbacks of the Fourier Transform, is used. In fact, as it is shown in Fig. 2, it is a Time-Scale approach, which allows understanding the results in the time-frequency plane. Note that the scale is inversely proportional to the frequency.

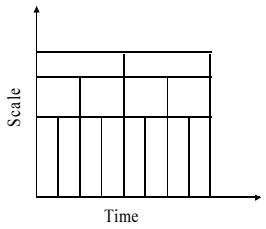


Fig. 2. Time-Frequency relation of the Wavelet analysis

Wavelet Transform Processing. The Wavelet Transform compares the original signal with the so-called Mother Wavelet. The Mother Wavelet is a wavelet prototype function, which can be modified to scale and to shift the signal as needed. Fig. 3 displays two types of Mother Wavelet function belonging to the Daubechies and Haar types. The correlation coefficients can be obtained from the comparison between the different Mother Wavelet functions and the original signal. From these coefficients it is possible to reconstruct the original signal using the inverse of the Wavelet Transform.

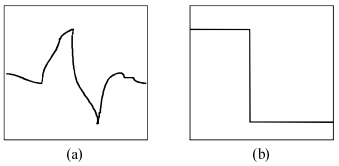


Fig. 3. Mother Wavelets, (a) Daubechies 2 and (b) Haar

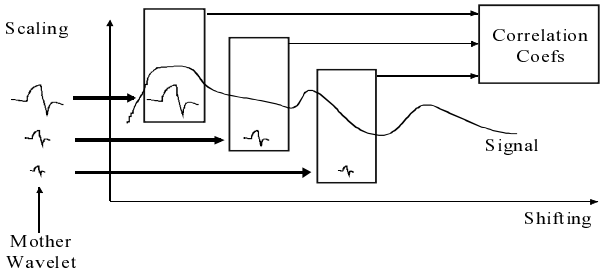


Fig. 4. Wavelet Transform processing

Fig. 4 shows the described process.

Once the Wavelet Transform is performed for all-possible scales (or levels), many characteristics of the signal are obtained. However, to select the most interesting scales and shifts for a concrete signal is a very difficult task. For such a reason, it is very common to analyze the signal by the Discrete Wavelet Transform. This transform consists in choosing only the scales and shifts based on powers of two. Thanks to this choosing, the redundant information is minimized, and so the computational load is substantially cut down [5, 6, 7].

Application of the Wavelet Transform to Images. Analysis of bidimensional signals is getting great improvements by using Wavelet based methods. For this problem Wavelet analysis technique makes use of regions with variable size. This technique allows not only to analyze regions of considerable size where information associated to low frequencies can be found (nearly homogeneous regions), but also small regions where information related to high frequencies is contained (vertices regions, edges or colours changes).

It is possible to characterize an image as a series of approximations and sets of finer details. The Wavelet Transform provides such a representation. The WT descomposition is multi-scale: it consists of a set of Approximation coefficients and three sets of Detail coefficients (Horizontal, Vertical and Diagonals). The Approximation coefficients represent coarse image information (they contain the most part of the image's energy), while the Details are close to zero, but the information they represent can be relevant in a particular context.

Fig. 5 illustrates this point. It has been obtained applying the WT to the image of a signal belonging to COFF class, using a Mother Wavelet Haar at level 2.

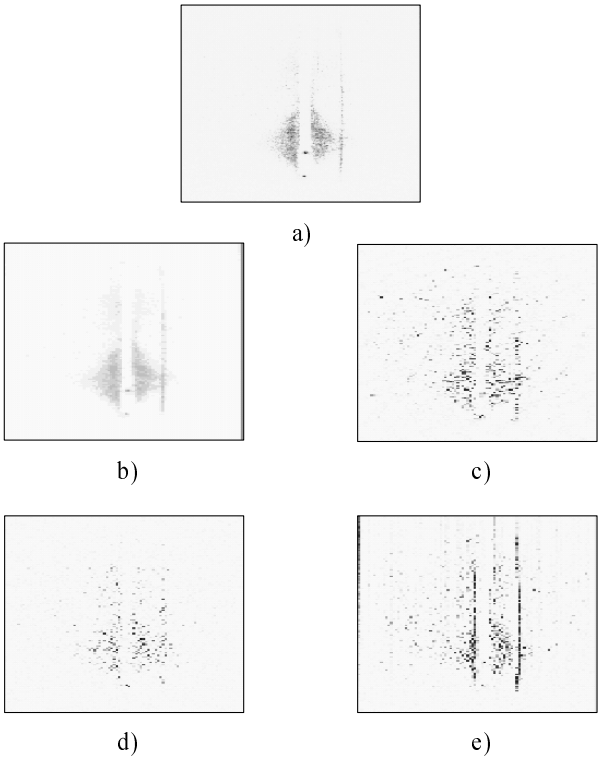


Fig. 5. DWT application to a signal of class COFF: (a) Original Signal, (b) Approximation, (c) Horizontal Detail, (d) Diagonal Detail, and (e) Vertical Detail

The family of Mother Wavelet functions used plays an important role in the final results. Its choice comes from experimentation with the work signals. Other important properties to be considered are the features of the wavelet coefficients and the decomposition level of the transform.

In relation to the TS signals, it has been found [8] that the best coefficient to characterize the images is the Vertical Detail, when the selected Mother Wavelet is the Haar at level 4. When applying the mentioned Wavelet Transform to the signals of the TS, the attributes are reduced from 221.760 to 900. So, the obtained attributes with the Wavelet Transform represent the 0.4% of the complete original signal.

2.2 Neural Networks: Feed Forward Multilayer

Neuronal Networks (NN) have been used successfully in great number of problems and applications. There is a diversity of types of NN, each one with a different structure according to the intentions of designer. However, in every case is possible to find the basic elements that define them. The NNs consist of elements of processing called neurons, which are grouped in layers and connected by synapses with an associate weight [9, 10].

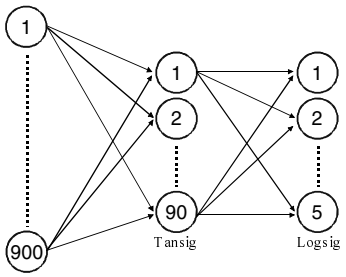


Fig. 6. Structure of the proposed NN

In the present work, a NN Feed Forward has been used. One of the possibilities of this type of NN is to use it for the supervised learning, where it is necessary to train the NN indicating to the input layer the attributes of a signal and the wished values to the output layer.

Fig. 6 shows the NN that has generated the best results. Note that the NN has an input layer of 900 attributes which come from the previous processing stage (generated by WT). The hidden layer used 90 with functions of activation Tansig, whereas the output layer has 5 neurons with functions of activation Logsig. After the training of the NN, every signal is associated with its class through of the activation from its output neuron and resetting the remaining ones.

The functions of activation are defined in Eq.1 and Eq.2.

$$Tansig(n) = \frac{2}{1 + \exp^{-2n}} - 1 \ . \tag{1}$$

$$Logsig(n) = \frac{1}{1 + \exp^{-n}} \ . \tag{2}$$

2.3 Train, Classify, and Testing Process

The training processing of the NN implies to obtain a set of weight through a Back-Propagation algorithm, which produces the minimal error between the values of the output layer and the wished values.

The classification process implies to use the trained NN to decide whether a signal presented to the NN belongs to a class or another.

To validate the efficiency of the classifier, the signals were divided randomly in training and testing sets. In this case, the training set was composed of 60% of the all signals. Later on, testing set was used to compare the obtained results with the wished values. To obtain an average performance, the procedure was realized for 100 different training and testing sets.

3 The Classifier

To implement the previous ideas in the present work, an application named Thomson Scattering Classifier has been designed in MATLAB [7, 11]. This ap-

plication allows to manipulate a set of labeled signals, whose main function is to evaluate the performance of the different classifiers. These classifiers can be obtained easily by modifying some parameters as: the Mother Wavelet, the decomposition level, the number of the NN layers, the activation functions, etc.

Fig. 7 presents the graphical user interface of the application.

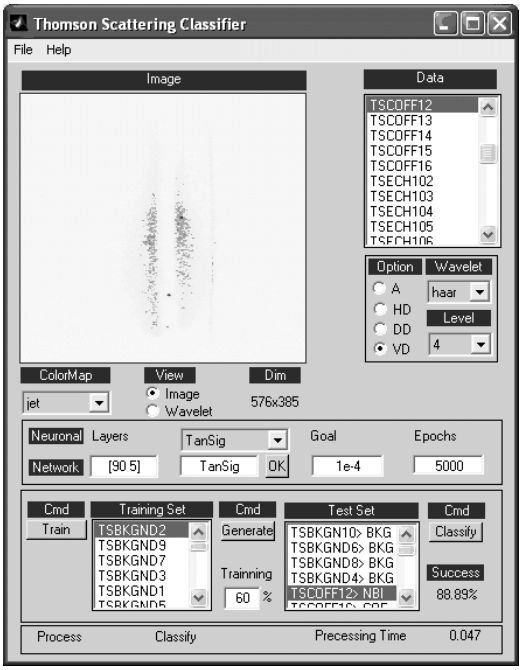


Fig. 7. Thomson Scattering Signals Classifier

3.1 Description of the Application

A brief description of the capabilities and available options of the developed classifier is given in the following sections.

Signal Image. The image of the signals can be displayed in the Image window at the left side. To select the signal to be displayed, an item of the Data list has to be clicked.

Wavelet Transform Configuration. In the application, it is possible to specify the different parameters associated with the WT. So, if to set up a decomposition level for the signal is wanted, there is to select an option from the popup Level menu. In this application, it is also possible to define the Mother Wavelet and in addition the set of obtained wavelet coefficients, that is, the Approximation (A) or the Detail (D).

Wavelet Transform Image. Once the type of the WT has been selected, the application displays the Wavelet Transform image of a particular signal. For this

purpose, it is necessary to select the signal from the Data list and then to click on the Wavelet option in the View section.

Random Generation of Training and Testing Sets. When pressing the Generate button, two sets of signals are randomly obtained for training and testing purposes. The proportion of signals that compose the training set is defined by the user.

Neuronal Network Parameters. The application allows to set up the NN parameters to specify: the number of layers, the number of neurons in every layer, the functions of activation, the required goal, and the training epochs.

Neuronal Network Training. After the application has generated the training and testing sets, it is necessary to train the NN. For this aim, it is only necessary to press the Train button. The NN is now ready to classify, only just if the required goal has been reached.

Testing Signals Classification. To evaluate the testing set, it is necessary to press the Classify button. So, the classifier will make the predictions for every signal of the testing set. Automatically, the classifier will also compare the obtained results with the labels of each one of the signals, identifying therefore, the percentage of hits.

4 Results and Conclusions

Thomson Scattering Classifier allows to do many different kind of classifiers, due to fact that NN or WT parameters can be changed according to user requirements.

We have selected a Mother Wavelet Haar at level 4 with vertical details, while the selected Neuronal Network has the same structure that the NN proposed in Fig. 6.

To test the designed classifier, many experiments were made. It is necessary to indicate that the number of signals available at the moment to do the experiments was 46.

We made 100 experiments, where every experiment was composed of two signal groups, that is, training and testing set, which were randomly generated. Fig. 8 shows the results for each one of the classes, being the average percentage of hits of 90.89%.

The previous classifier, using Wavelet Transform and Neuronal Network, constitutes an alternative for automatic classification of Thomson Scattering signals.

The development of Thomson Scattering Classifier to experiment with the described techniques, allows to observe the effect of each one in the classification, to reduce evaluation time, and to search satisfactory parameters.

It is necessary to consider that probably better results can be obtained if knowledge about the problem context is added. However the presented results are a start point for new analysis.

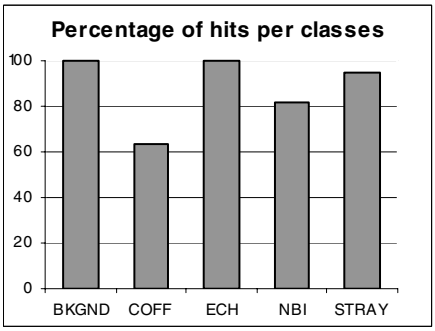


Fig. 8. Results for Each One of the Classes

Finally, it is necessary to emphasize that the training of the classifier has been done by a limited set of signals. So if the number of signal increase, it will be possible to improve the Percentage of hits obtained.

References

1. Alejaldre, C. et al.: Plasma Phys. Controlled Fusion 41, 1 (1999), pag. A539
2. Duda, R., Hort, P., Stork, D.: Pattern Classification, Second Edition, A Wiley-Interscience Publication (2001)
3. Nakanishi, H.; Hochin, T.; Kojima, M. and LABCOM group: Search and Retrieval Methods of Similar Plasma Waveforms. 4th IAEA TCM on Control, Data Acquisition, and Remote Participation for Fusion Research. San Diego, USA (July 21-23, 2003). (To appear in Fusion Engineering and Design)
4. Dormido-Canto, S., Farias, G., Dormido, R., Vega, J., Snchez, J., Santos, M. and The TJ-II Team: TJ-II Wave Forms Analysis with Wavelets and Support Vector Machines. Review Scientific Instruments, vol. 75, no. 10, (2004), 4254-4257
5. Daubechies, I.: Ten Lectures on Wavelets, SIAM (1992)
6. Mallat, S.: A Wavelt Tour of signal Processing, 2 Edition, Academia Press (2001)
7. Misiti, M., Oppenheim, G., Poggi, J., Misita, Y.: Wavelet Toolbox User's Guide (V.1). The MathWorks, Inc. (1995-98)
8. Farias, G., Santos, M., Marrn, J. L., Dormido-Canto, S.: Determinacin de Parmetros de la Transformada Wavelets para la Clasificacin de Seales del Diagnstico Scattering Thomson. XXV Jornadas de Automtica, Ciudad Real (Espaa), ISBN: 84-688-7460-4 (2004)
9. Hilera, J.R., Martnez, V.J.: Redes Neuronales Artificiales. Fundamentos, modelos y aplicaciones. Ed. Rama (1995)
10. Freeman, J., Skapura, D. : Redes Neuronales, Algoritmos, aplicaciones y tcnicas de programacin. Addison-Wesley/Diaz de Santos (1993)
11. Demuth, H., Beale, M.: Neural Network Toolbox User's Guide (V.3). The MathWorks Inc. (1992-1998)

Article 19

TJ-II wave forms analysis

19.1 Bibliographic Description

Title

TJ-II wave forms analysis with wavelets and support vector machines.

Citation

S. Dormido-Canto, G. Farias, R. Dormido, J. Vega, J. Sánchez, M. Santos and The TJ-II Team (2004) TJ-II Wave Forms Analysis with Wavelets and Support Vector Machines, *Review of Scientific Instruments*, ISSN 0034-6748, Volume 75, Pages 4254-4257.

Abstract

Since the fusion plasma experiment generates hundreds of signals, it is essential to have automatic mechanisms for searching similarities and retrieving of specific data in the wave form database. Wavelet transform (WT) is a transformation that allows one to map signals to spaces of lower dimensionality. Support vector machine (SVM) is a very effective method for general purpose pattern recognition. Given a set of input vectors which belong to two different classes, the SVM maps the inputs into a high-dimensional feature space through some nonlinear mapping, where an optimal separating hyperplane is constructed. In this work, the combined use of WT and SVM is proposed for searching and retrieving similar wave forms in the TJ-II database. In a first stage, plasma signals will be preprocessed by WT to reduce their dimensionality and to extract their main features. In the next stage, and using the smoothed signals produced by the WT,

SVM will be applied to show up the efficiency of the proposed method to deal with the problem of sorting out thousands of fusion plasma signals.From observation of several experiments, our WT+SVM method is very viable, and the results seems promising. However, we have further work to do. We have to finish the development of a Matlab toolbox for WT+SVM processing and to include new relevant features in the SVM inputs to improve the technique. We have also to make a better preprocessing of the input signals and to study the performance of other generic and self custom kernels. To reach it, and since the preprocessing stages are very time consuming, we are going to study the viability of using DSPs, RPGAs or parallel programming techniques to reduce the execution time.

References

D. Radiei, A. Mendelzon(1998); Nakanishi et al. (2003); V. Vapnik (1995); J. D. Sebald, J. A. Bucklew (2001); C. Alejaldre et al. (1999); S. Mallat (2001); B. Schölkopt, A. J. Smola (2002).

Impact Factor

Review Of Scientific Instruments has an impact factor of 1.367 according to Thomson Reuters Journal Citation Reports (2011).

REVIEW OF SCIENTIFIC INSTRUMENTS VOLUME 75, NUMBER 10 OCTOBER 2004

TJ-II wave forms analysis with wavelets and support vector machines

S. Dormido-Canto[a)]
Departarmento Informática y Automática-UNED. C/Juan del Rosal 16, 5. 28040 Madrid, Spain

G. Farias
Departarmento Arquitectura de Computadores y Automática-UCM. Ciudad Universitaria, 28040 Madrid, Spain

R. Dormido
Departarmento Informática y Automática-UNED, C/Juan del Rosal 16, 5, 28040 Madrid, Spain

J. Vega
Asociación EURATOM/CIEMAT para FUSIÓN. Ada, Complutense 22. 28040 Madrid, Spain

J. Sánchez
Departarmento Informática y Automática-UNED, C/Juan del Rosal 16, 5, 28040 Madrid, Spain

M. Santos
Departarmento Arquitectura de Computadores y Automática-UCM. Ciudad Universitaria, 28040 Madrid, Spain

The TJ-II Team

(Presented on 22 April 2004; published 18 October 2004)

Since the fusion plasma experiment generates hundreds of signals, it is essential to have automatic mechanisms for searching similarities and retrieving of specific data in the wave form database. Wavelet transform (WT) is a transformation that allows one to map signals to spaces of lower dimensionality. Support vector machine (SVM) is a very effective method for general purpose pattern recognition. Given a set of input vectors which belong to two different classes, the SVM maps the inputs into a high-dimensional feature space through some nonlinear mapping, where an optimal separating hyperplane is constructed. In this work, the combined use of WT and SVM is proposed for searching and retrieving similar wave forms in the TJ-II database. In a first stage, plasma signals will be preprocessed by WT to reduce their dimensionality and to extract their main features. In the next stage, and using the smoothed signals produced by the WT, SVM will be applied to show up the efficiency of the proposed method to deal with the problem of sorting out thousands of fusion plasma signals.From observation of several experiments, our WT+SVM method is very viable, and the results seems promising. However, we have further work to do. We have to finish the development of a Matlab toolbox for WT+SVM processing and to include new relevant features in the SVM inputs to improve the technique. We have also to make a better preprocessing of the input signals and to study the performance of other generic and self custom kernels. To reach it, and since the preprocessing stages are very time consuming, we are going to study the viability of using DSPs, RPGAs or parallel programming techniques to reduce the execution time. © *2004 American Institute of Physics.* [DOI: 10.1063/1.1787611]

I. INTRODUCTION

Databases in nuclear fusion experiments are made up of thousands of signals. For this reason, data analysis must be simplified by developing automatic mechanisms for fast search and retrieval of specific data in the wave form database. In particular, a method for fast similarity search in the database would be very helpful.

In Ref. 1 a method to find similar time sequences using discrete Fourier transformation (DFT) to reduce the dimensionality of the feature vectors is proposed. In Ref. 2, the previous DFT-based method is used to search similar phenomena in wave form databases but it is just applied with slowly varying signals. However, the DFT has difficulties when used with fast varying wave forms since time information is lost when transforming to the frequency domain. Wave transform (WT) offers an efficient alternative to data processing and provides many advantages: (1) data compression, (2) computing efficiency, and (3) simultaneous time and frequency representation.

Support vector machine (SVM) is a very effective method for general purpose pattern recognition.[3,4] In a few words, given a set of input vectors which belong to two different classes, the SVM maps the inputs into a high-dimensional feature space through some nonlinear mapping, where an optimal separating hyperplane is constructed in order to minimize the risk of misclassification. The hyperplane is determined by a subset of points of the two classes, named Support vectors.

[a)]Author to whom correspondence should be addressed; electronic mail: sebas@dia.uned.es

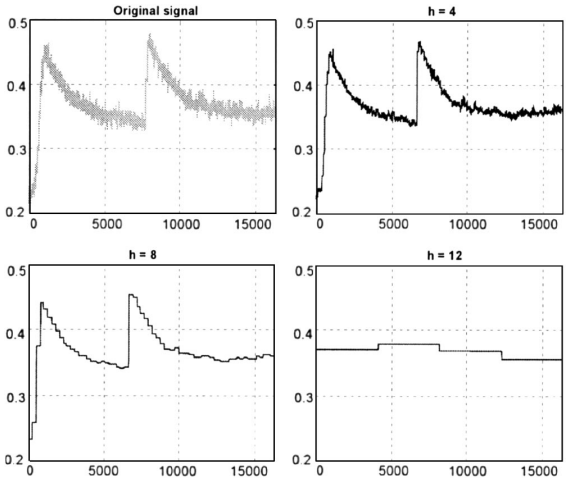


FIG. 1. Original signal and its wavelet transform approximations with three different decomposition levels ($h=4,8,12$).

In this work, preliminary results are shown when using WT techniques for characterizing the signals and SVM as the technique for pattern recognition and information retrieval. The proposed method has been applied to the TJ-II database. The TJ-II is a stellarator device[5] (heliac type, $B(0) \leqslant 1.2\ T, R(0)=1.5$ m, $<a> \leqslant 0.22$ m) located at CIEMAT (Madrid, Spain) that can explore a wide rotational transform range ($0.9 \leqslant \text{iota}/2p \leqslant 2.2$). At present, 940 digitization channels are available for experimental measurements in the TJ-II.

II. WAVELET TRANSFORM

Wavelet algorithms process data at different resolutions or decomposition levels in contrast with DFT where only frequency components are considered.[6] The WT is applied to the original signals in order to compute a few coefficients for each signal in a fast way (Fig. 1).

The WT, in particular the Haar transform is used in this work, is chosen for many advantages. First, WT has been used for data compression. Second, it can be computed quickly, requiring linear time in the length of the signal and simple coding. The complexity of Haar transform is $O(n)$ while $O(n \log n)$ computation is required for DFT. Although these computations are all involved in the preprocessing stage, the complexity of the transformation can be a concern, especially when the database is large, as it happens in our case. Another advantage is that Haar wavelet representations of signals bear more information than that of DFT. While DFT extracts the lower harmonics, which represent the general shape of a time sequence, WT encodes a coarser resolution of the original time sequence with its preceding coefficients.

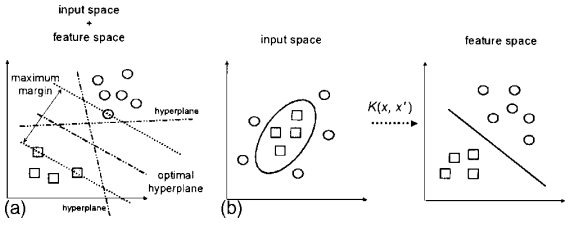


FIG. 2. The idea of SVMs: map the training data into a higher-dimensional feature space via K, and construct a separating hyperplane with maximum range there. This yields a nonlinear decision boundary in input space. By the use of kernel functions, it is possible to compute the separating hyperplane without explicitly carrying out the map into the feature space. (a) Linearly separable case. (b) Nonlinearly separable case.

TABLE I. Kernel functions extensively used.

Inner product	$K(x,x')=\langle x,x'\rangle$
Polynomial of degree d	$K(x,x')=(\langle x,x'\rangle+1)^d$
Gaussian radial basis function	$K(x,x')=\exp\{-\|x-x'\|^2/2\sigma^2\}$

III. SUPPORT VECTOR MACHINES FOR CLASSIFICATION

SVM is a universal constructive learning procedure based on the statistical learning theory.[3] Let us consider the problem of separating a set of training vectors belonging to two separate classes (a binary classifier)

$$\{(x_1,y_1),\ldots,(x_n,y_n)\}, x \in R'', y \in \{-1,+1\}$$

with a hyperplane decision function $D(x)$

$$D(x) = \langle w,x\rangle + b,$$

where $\langle,\rangle$ denotes inner product. In linearly separable cases, SVM constructs a hyperplane which separates the training data without error. The hyperplane is constructed by finding another vector w and a parameter b that minimizes $\|w\|^2$ and satisfies the following conditions:

$$y_i[\langle w,x_i\rangle + b] \geqslant 1, \quad i=1,\ldots,n,$$

where w is a normal weight vector to the hyperplane, $|b|/\|w\|$ is the perpendicular distance from the hyperplane to the origin, and $\|w\|^2$ is the Euclidean norm of w. After the determination of w and b, a given vector x can be classified by

$$\text{sgn}(\langle w,x\rangle + b). \tag{1}$$

In nonlinearly separable cases, SVM can map the input vectors into a high-dimensional feature space. By selecting a nonlinear mapping *a priori*, SVM constructs an optimal separating hyperplane in this higher-dimensional space. A

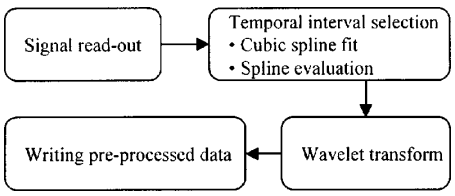


FIG. 3. Signal conditioning data flow.

TABLE II. Classes of signals of the TJ-II database.

Classes of signals	Description
BOL5	Bolometer signal
ECE7	Electron cyclotron emission
RX306	Soft x ray
ACTON275	Espectroscopic signal (*CV*)
HALFAC3	$H\alpha$
Densidad2	Line averaged electron density

kernel function $K(x,x')$ performs the nonlinear mapping into feature space,[7] and the original constrains are the same. In this way, the evaluation of the inner products among the vectors in a high-dimensional feature space is done indirectly via the evaluation of the kernel $K(x,x')$ between support vectors and vectors in the input space (Fig. 2). This provides a way of addressing the technical problem of evaluating inner products in a high-dimensional feature space. Examples of kernel functions are shown in Table I.

Linear support vector machine is applied to this feature space and then the decision function is given by

$$f(x) = \mathrm{sgn}\Big(\sum_{i\in \mathrm{SVs}} \alpha_i y_i K(x_i,x) + b\Big), \tag{2}$$

where the coefficients α_i and b are determined by maximizing the following Langrangian expression:

$$\sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\alpha_j y_i y_j K(x_i,x_j)$$

under conditions

$$\alpha_i \geqslant 0 \quad \text{and} \quad \sum_{i=1}^{n} \alpha_i y_i = 0.$$

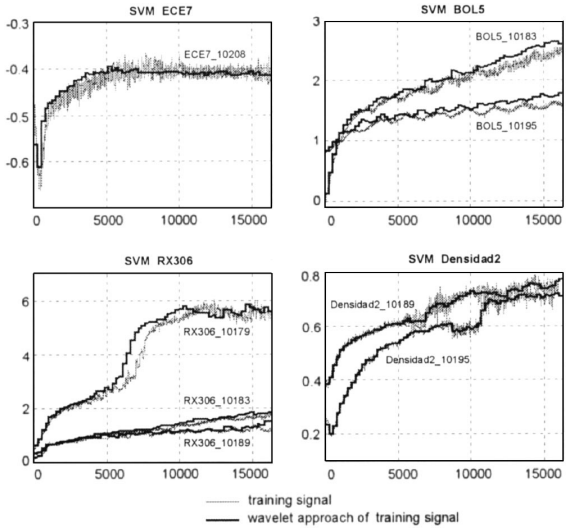


FIG. 4. Positive support vector for every class in the experiment 1.

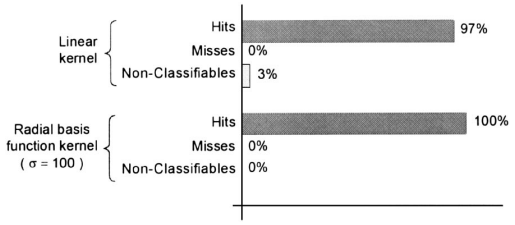


FIG. 5. Results of the experiment 1.

A positive or negative value from Eqs. (1) or (2) indicates that the vector x belongs or not to class 1. The data samples for which the α_i are nonzero are the support vectors. The parameter b is given by

$$b = y_s \sum_{i\in \mathrm{SVs}} \alpha_i y_i K(x_s,x_i),$$

where (x_s,y_s) is any one of the support vectors.

IV. PERFORMANCE EVALUATION

Some preliminary results are presented in this section. Our proof was based on classifying and recognizing temporal evolution signals from the TJ-II database. It is accomplished in a two-step process. A first step provides signal conditioning (Fig. 3), to ensure the same sampling period and number of samples.

This requirement is a consequence of the fact that signals could have been collected with different acquisition parameters. A second step is devoted to perform, first, the learning process with SVM and some of the preprocessed data. Second, classification tasks are carried out. All processes have been programmed from the MATLAB software package.

In order to evaluate the approach, two experiments have been carried out to classify signals stored in the TJ-II database. These signals belong to one of the classes shown in Table II.

In the first stage of our approach, the signals are preprocessed in both of our experiments by Haar transform (with a decomposition level of 8) to reduce the dimensionality of the problem. In the second stage, the test signals are classified using SVM. The method applied is *one versus the rest*, that allows one to get multi-class classifiers. For that reason, we construct a set of binary classifiers as is explained in Sec. III.

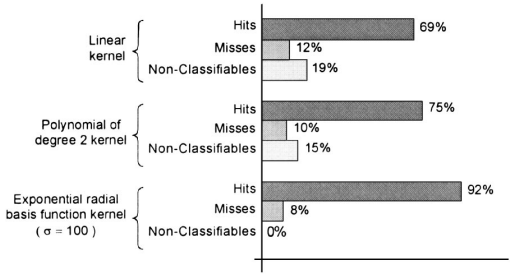


FIG. 6. Results of the experiment 2.

Each classifier is trained to separate one class from the rest, and to combine them by doing the multi-class classification according to the maximal output before applying the *sign* function (Eq. (1)). Next, two experiments are shown to demonstrate the viability of the proposed approach.

In the first experiment, four classes have been considered: ECE7, BOL5,RX306, and Densidad2. The training set is composed by 40 signals and the test set by 32 signals obtained from the TJ-JI database. Figure 4 displays the positive support vectors for each class using a linear kernel, the training signal corresponding to the original signal in TJ-II, and the wavelet approach which is the signal re-sampled to 16384 samples after the wavelet transform. The percentages of hits, misses, and nonclassifiable signals are illustrated in Fig. 5. In a second experiment, the training and test sets are composed by 60 and 48 signals and the number of classes was 6, respectively. Figure 6 shows the results.

ACKNOWLEDGMENTS

The authors wish to thank Professor S. Dormido Bencomo (UNED) and Professor J. M. de la Cruz (UCM) for their constructive comments and invaluable guidance.

[1] D. Radiei and A. Mendelzon, *Fifth International Conference on Foundations of Data Organization*, Kobe, Japan, 1998, p. 249 (unpublished).
[2] Nakanishi *et al.*, *Fourth IAEA TCM on Control, Data Acquisition, and Remote Participation for Fusion Research*, San Diego, CA, 2003 (unpublished).
[3] V. Vapnik, *The Nature of Statistical Learning Theory* (Springer, Berlin, 1995).
[4] J. D. Sebald and J. A. Bucklew, IEEE Trans. Signal Process. **49**, 2865 (2001).
[5] C. Alejaldre *et al.*, Plasma Phys. Controlled Fusion **41**, 539 (1999).
[6] S. Mallat, *A Wavelet Tour of Signal Processing*, 2nd ed. (Academic press, New York, 2001).
[7] B. Schölkopt and A. J. Smola, *Learning with Kernels* (MIT Press, Combridge, MA, 2002).

Bibliography

Alejaldre, C., Alonso, J., Almoguera, L., Ascasíbar, E., Baciero, A., Balbín, R., Blaumoser, M., Botija, J., Brañas, B., De la Cal, E. et al. (1999), 'First plasmas in the TJ-II flexible heliac', *Plasma physics and controlled fusion* **41**(3A), A539.

Bellizio, T., Albanese, R., Ambrosino, G., Ariola, M., Artaserse, G., Crisanti, F., Coccorese, V., De Tommasi, G., Lomas, P. J., Maviglia, F. et al. (2011), 'Control of elongated plasma in presence of ELMs in the JET tokamak', *Nuclear Science, IEEE Transactions on* **58**(4), 1497–1502.

Breault, R. P. (1995), 'Control of stray light', *Handbook of Optics* **1**, 38–1.

Cannas, B., Fanni, A., Marongiu, E. & Sonato, P. (2003), 'Disruption forecasting at JET using neural networks', *Nuclear fusion* **44**(1), 68.

Cannas, B., Fanni, A., Pautasso, G., Sias, G., Sonato, P. & Zedda, M. (2006), Disruption prediction at ASDEX upgrade using neural networks, *in* '33rd EPS Conference on Plasma Physics. Contributed Papers,(Eds.) F. De Marco, G. Flad. ECA', Vol. 30.

Carpenter, G. A. & Grossberg, S. (1993), 'Art 2: Self-organization of stable category recognition codes for analog input patterns applied optics 26: 4919ˆ 930', *Neurocomputing* **2**(4919), 151.

Cherkassky, V. & Mulier, F. M. (2007), *Learning from data: concepts, theory, and methods*, Wiley-IEEE Press.

Chi, M., Feng, R. & Bruzzone, L. (2008), 'Classification of hyperspectral remote-sensing data with primal svm for small-sized training dataset problem', *Advances in space research* **41**(11), 1793–1799.

Daubechies, I. (1992), 'Ten lectures on wavelets (cbms-nsf regional conference series in applied mathematics) author: Ingrid daubechies, publishe'.

Dormido-Canto, S., Farias, G., Dormido, R., Vega, J., Sánchez, J. & Santos, M. (2004), 'TJ-II wave forms analysis with wavelets and support vector machines', *Review of scientific instruments* **75**(10), 4254–4257.

Dormido-Canto, S., Farias, G., Vega, J., Dormido, R., Sánchez, J., Duro, N., Santos, M., Martín, J. & Pajares, G. (2006), 'Search and retrieval of plasma wave forms: Structural pattern recognition approach', *Review of scientific instruments* **77**(10), 10F514–10F514.

Dormido-Canto, S., Farias, G., Vega, J. & Pastor, I. (2012), 'Image processing methods for noise reduction in the TJ-II Thomson Scattering diagnostic', *Fusion Engineering and Design* **87**(12), 2170–2173.

Dormido-Canto, S., Farias, G. et al. (2008*a*), 'Classifier based on support vector machine for JET plasma configurations', *Review of Scientific Instruments* **79**(10), 10F326–10F326.

Dormido-Canto, S., Farias, G. et al. (2008*b*), 'Structural pattern recognition methods based on string comparison for fusion databases', *Fusion Engineering and Design* **83**(2), 421–424.

Dormido-Canto, S., Vega, J., Sánchez, J. & Farias, G. (2005), 'Information retrieval and classification with wavelets and support vector machines', *Artificial Intelligence and Knowledge Engineering Applications: A Bioinspired Approach* pp. 548–557.

Duda, R. O., Hart, P. E. & Stork, D. G. (2001), 'Pattern classification', *New York: John Wiley, Section* **10**, l.

Duro, N., Dormido, R., Vega, J., Dormido-Canto, S., Farias, G., Sánchez, J., Vargas, H., Murari, A. & Contributors, J. (2009), 'Automated recognition system for ELM classification in JET', *Fusion Engineering and Design* **84**(2), 712–715.

Duro, N., Vega, J., Dormido, R., Farias, G., Dormido-Canto, S., Sánchez, J., Santos, M. & Pajares, G. (2006), 'Automated clustering procedure for TJ-II experimental signals', *Fusion engineering and design* **81**(15), 1987–1991.

EFDA (2013*a*), 'The Joint European Torus (JET) homepage'. [Online; accessed 31-March-2013].
URL: *https://www.efda.org/jet/*

EFDA (2013*b*), 'The Joint European Torus (JET) homepage: ELM Classification'. [Online; accessed 31-March-2013].
URL: *https://www.efda.org/fusion/focus-on/edge-localised-modes/elm-classification/*

Farias, G., Dormido-Canto, S., Vega, J., Pastor, I. & Santos, M. (2013), 'Application and validation of image processing algorithms to reduce the stray light on the TJ-II Thomson Scattering diagnostic', *Fusion Science and Technology* **63**(1), 20–25.

Farias, G., Dormido-Canto, S., Vega, J., Sánchez, J., Duro, N., Dormido, R., Ochando, M., Santos, M. & Pajares, G. (2006), 'Searching for patterns in TJ-II time evolution signals', *Fusion engineering and design* **81**(15), 1993–1997.

Farias, G., Dormido, R., Santos, M. & Duro, N. (2005), 'Image classifier for the TJ-II Thomson Scattering diagnostic: evaluation with a feed forward neural network', *Artificial Intelligence and Knowledge Engineering Applications: A Bioinspired Approach* pp. 362–381.

Farias, G. & Santos, M. (2005), 'Analysis of the wavelet transform parameters in images processing', *LNCCS* **2**, 51–54.

Farias, G. & Santos, M. (2007), 'A computational fusion of wavelets and neural networks in a classifier for biomedical applications', *Lecture Series on Computer and Computational Sciences* **8**, 66–70.

Farias, G., Santos, M. & López, V. (2010), 'Making decisions on brain tumor diagnosis by soft computing techniques', *Soft Computing-A Fusion of Foundations, Methodologies and Applications* **14**(12), 1287–1296.

Farias, G., Santos, M., Marrón, J. & Dormido-Canto, S. (2004), Determinación de parámetros de la transformada wavelets para la clasificación de señales del diagnóstico scattering thomson, *in* 'XXV Jornadas de Automática, Ciudad Real (España)'.

Farias, G., Vega, J., González, S., Pereira, A., Lee, X., Schissel, D. & Gohil, P. (2012), 'Automatic determination of L/H transition times in DIII-D through a collaborative distributed environment', *Fusion Engineering and Design* **87**(12), 2081–2083.

Feddema, H. (2001), *Microsoft Access version 2002 inside out*, Microsoft Press.

Forman, G. & Cohen, I. (2004), 'Learning from little: Comparison of classifiers given little training', *Knowledge Discovery in Databases: PKDD 2004* pp. 161–172.

Freeman, J. & Skapura, D. (1991), 'Neural networks, algorithms, applications, and programming techniques'.

Fu, K. S. & Albus, J. E. (1982), *Syntactic pattern recognition and applications*, Vol. 4, Prentice-Hall New York.

Fulkerson, B., Vedaldi, A. & Soatto, S. (2009), Class segmentation and object localization with superpixel neighborhoods, *in* 'Computer Vision, 2009 IEEE 12th International Conference on', IEEE, pp. 670–677.

General Atomics (2013), 'The DIII-D homepage'. [Online; accessed 31-March-2013]. **URL:** *https://fusion.gat.com/global/DIII-D*

González, S., Vega, J., Murari, A., Pereira, A., Dormido-Canto, S. & Ramírez, J. (2012), 'H/L transition time estimation in JET using conformal predictors', *Fusion Engineering and Design* **87**(12), 2084–2086.

González, S., Vega, J., Murari, A., Pereira, A., Ramírez, J., Dormido-Canto, S. & Contributors, J. (2010), 'Support vector machine-based feature extractor for L/H transitions in JET', *Review of Scientific Instruments* **81**, 10E123.

Gu, C., Lim, J. J., Arbeláez, P. & Malik, J. (2009), Recognition using regions, *in* 'Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on', IEEE, pp. 1030–1037.

Haykin, S. (2004), 'A comprehensive foundation', *Neural Networks* **2**.

Hearst, M. A., Dumais, S., Osman, E., Platt, J. & Scholkopf, B. (1998), 'Support vector machines', *Intelligent Systems and their Applications, IEEE* **13**(4), 18–28.

Hilera, J. & Martínez, V. (1995), 'Redes neuronales artificiales: fundamentos, modelos y aplicaciones', *Madrid: Ra-ma* .

Jang, J. (1993), 'ANFIS: Adaptive-network-based fuzzy inference system', *Systems, Man and Cybernetics, IEEE Transactions on* **23**(3), 665–685.

Johnson, S. C. (1967), 'Hierarchical clustering schemes', *Psychometrika* **32**(3), 241–254.

Lawson, J. D. (2002), 'Some criteria for a power producing thermonuclear reactor', *Proceedings of the Physical Society. Section B* **70**(1), 6.

Liang, Y. (2011), 'Overview of edge-localized mode control in tokamak plasmas', *Fusion Science and Technology* **59**(3), 586.

Lister, J. B., Hofmann, F., Moret, J.-M., Bühlmann, F., Dutch, M. J., Fasel, D., Favre, A., Isoz, P.-F., Marletaz, B., Marmillod, P. et al. (1997), 'The control of tokamak configuration variable plasmas: Plasma control issues for tokamaks', *Fusion Technology* **32**(3), 321–373.

MacQueen, J. et al. (1967), Some methods for classification and analysis of multivariate observations, *in* 'Proceedings of the fifth Berkeley symposium on mathematical statistics and probability', Vol. 1, California, USA, pp. 281–297.

Makili, L., Vega, J., Dormido-Canto, S., Pastor, I., Pereira, A., Farias, G., Portas, A., Pérez-Risco, D., Rodríguez-Fernández, M. & Busch, P. (2010), 'Upgrade of the automatic analysis system in the TJ-II Thomson Scattering diagnostic: New image recognition classifier and fault condition detection', *Fusion Engineering and Design* **85**(3), 415–418.

Mallat, S. (2008), *A wavelet tour of signal processing: the sparse way*, Academic press.

Martín, J., Santos, M., Farias, G., Duro, N., Sanchez, J., Dormido, R., Dormido-Canto, S. & Vega, J. (2007), Dynamic clustering and neuro-fuzzy identification for the analysis of fusion plasma signals, *in* 'Intelligent Signal Processing, 2007. WISP 2007. IEEE International Symposium on', IEEE, pp. 1–6.

Martín, J., Santos, M., Farias, G., Duro, N., Sanchez, J., Dormido, R., Dormido-Canto, S., Vega, J. & Vargas, H. (2009), 'Dynamic clustering and modeling approaches for

fusion plasma signals', *Instrumentation and Measurement, IEEE Transactions on* **58**(9), 2969–2978.

Martinez, W. L. & Martinez, A. R. (2001), *Computational statistics handbook with MATLAB*, Vol. 2, Chapman & Hall/CRC.

Misiti, M., Misiti, Y., Oppenheim, G. & Poggi, J.-M. (2004), 'Matlab wavelet toolbox user's guide. version 3.'.

Murari, A., Vagliasindi, G., Arena, P., Fortuna, L., Barana, O., Johnson, M. & Contributors, J. (2008), 'Prototype of an adaptive disruption predictor for JET based on fuzzy logic and regression trees', *Nuclear Fusion* **48**(3), 035010.

Murari, A., Vega, J., Rattá, G., Vagliasindi, G., Johnson, M., Hong, S. & Contributors, J. (2009), 'Unbiased and non-supervised learning methods for disruption prediction at JET', *Nuclear Fusion* **49**(5), 055028.

Nakanishi, H., Hochin, T. & Kojima, M. (2004), 'Search and retrieval method of similar plasma waveforms', *Fusion engineering and design* **71**(1), 189–193.

Nakanishi, H., Hochin, T. & Kojima, M. (2006), 'Similar pattern search for time-sectional oscillation in huge plasma waveform database', *Fusion engineering and design* **81**(15), 2003–2007.

Ongena, J. (2006), 'JET's contribution to fusion science and iter', *Physica Scripta* **2006**(T123), 14.

Rafiei, D. & Mendelzon, A. (1998), 'Efficient retrieval of similar time sequences using dft', *arXiv preprint cs/9809033* .

Rattá, G. (2012), Técnicas de minería de datos aplicadas a fusión nuclear: predicción en tiempo real y clasificación, PhD thesis, Universidad Nacional de Educación a Distancia (España). Escuela Técnica Superior de Ingeniería Informática.

Rattá, G., Vega, J., Pereira, A., Portas, A., De la Luna, E., Dormido-Canto, S., Farias, G., Dormido, R., Sánchez, J., Duro, N., Vargas, H., Santos, M., Pajares, G., Murari, A. & Contributors, J. (2008), 'First applications of structural pattern recognition methods to the investigation of specific physical phenomena at JET', *Fusion Engineering and Design* **83**(2), 467–470.

Reitz, J. & Milford, F. (1996), 'Fundamentos de la teoría electromagnética'.

Rojas, R. (1996), *Neural networks: a systematic introduction*, Springer.

Ruiz, M., Vega, J., Rattá, G., Barrera, E., Murari, A., López, J. & Arcas, G. (2010), 'Real time plasma disruptions detection in JET implemented with the itms platform using fpga based idaq'.

Saibene, G., Horton, L., Sartori, R., Balet, B., Clement, S., Conway, G., Cordey, J., De Esch, H., Ingesson, L., Lingertat, J. et al. (2002), 'The influence of isotope mass, edge magnetic shear and input power on high density ELMy H modes in JET', *Nuclear Fusion* **39**(9), 1133.

Santos, M. & Farias, G. (2010), 'Laboratorios virtuales de procesamiento de señales', *Revista Iberoamericana de Automática e Informática Industrial RIAI* **7**(1), 91–100.

Schölkopf, B. & Smola, A. J. (2001), *Learning with kernels: Support vector machines, regularization, optimization, and beyond*, MIT press.

Schuller, F. (1999), 'Disruptions in tokamaks', *Plasma Physics and Controlled Fusion* **37**(11A), A135.

Sebald, D. J. & Bucklew, J. A. (2000), 'Support vector machine techniques for nonlinear equalization', *Signal Processing, IEEE Transactions on* **48**(11), 3217–3226.

Sheffield, J. (1994), 'The physics of magnetic fusion reactors', *Reviews of Modern Physics* **66**(3), 1015.

Vapnik, V. (1999), *The nature of statistical learning theory*, Springer.

Vega, J., Murari, A., Vagliasindi, G., Rattá, G. & Contributors, J. (2009), 'Automated estimation of L/H transition times at JET by combining bayesian statistics and support vector machines', *Nuclear Fusion* **49**(8), 085023.

Vega, J., Pereira, A., Portas, A., Dormido-Canto, S., Farias, G., Dormido, R., Sánchez, J., Duro, N., Santos, M., Sánchez, E. & Pajares, G. (2008), 'Data mining technique for fast retrieval of similar waveforms in fusion massive databases', *Fusion Engineering and Design* **83**(1), 132–139.

Vega, J., Rattá, G., Murari, A., Castro, P., Dormido-Canto, S., Dormido, R., Farias, G., Pereira, A., Portas, A., De la Luna, E., Pastor, I., Sánchez, J., Duro, N., Castro, R., Santos, M. & Vargas, H. (2007), Recent results on structural pattern recognition for fusion massive databases, *in* 'Intelligent Signal Processing, 2007. WISP 2007. IEEE International Symposium on', IEEE, pp. 1–6.

Vovk, V., Gammerman, A. & Shafer, G. (2005), *Algorithmic learning in a random world*, Springer Science+ Business Media.

Wagner, F., Becker, G., Behringer, K., Campbell, D., Eberhagen, A., Engelhardt, W., Fussmann, G., Gehre, O., Gernhardt, J., Gierke, G. v. et al. (1982), 'Regime of improved confinement and high beta in neutral-beam-heated divertor discharges of the ASDEX tokamak', *Physical Review Letters* **49**(19), 1408–1412.

Wakatani, M. (1998), *Stellarator and Heliotron devices*, Vol. 95, Oxford University Press, USA.

Weston, J., Mukherjee, S., Chapelle, O., Pontil, M., Poggio, T. & Vapnik, V. (2001), 'Feature selection for svms', *Advances in neural information processing systems* pp. 668–674.