PROYECTO FIN DE MÁSTER EN

SISTEMAS INTELIGENTES

CURSO 2010-2011

GESTIÓN INTELIGENTE DE LA DEMANDA EN LA CADENA DE SUMINISTRO

Fernando Turrado García

Director:

Luis Javier García Villalba

Departamento de Ingeniería del Software e Inteligencia Artificial

---

MÁSTER EN INVESTIGACIÓN EN INFORMÁTICA

FACULTAD DE INFORMÁTICA

UNIVERSIDAD COMPLUTENSE DE MADRID

El abajo firmante, matriculado en el Máster en Investigación en Informática de la Facultad de Informática, autoriza a la Universidad Complutense de Madrid (UCM) a difundir y utilizar con fines académicos, no comerciales y mencionando expresamente a su autor el presente Trabajo Fin de Máster: "***Gestión inteligente de la demanda en la cadena de suministro***", realizado durante el curso académico 2010-2011 bajo la dirección de Luis Javier García Villalba en el Departamento de Ingeniería del Software e Inteligencia Artificial, y a la Biblioteca de la UCM a depositarlo en el Archivo Institucional E-Prints Complutense con el objeto de incrementar la difusión, uso e impacto del trabajo en Internet y garantizar su preservación y acceso a largo plazo.

_____

Fernando Turrado García

Abstract

To be able of anticipate demand is a key factor for commercial success in the Supply-Chain sector. The benefits can be grouped around two main concepts: optimization of operations through the development of optimal strategies for procurement and stock reduction that reduces storage costs, handling ... Currently there are a variety of methods to make predictions, this methods vary from pure statistical methods such as exponential smoothing Holt-Winters or ARIMA models, to those based on artificial intelligence techniques like neural networks or fuzzy systems. However, despite being able to build accurate models, in managing the supply chain based on forecasts there is a problem known as "Forrester effect" regardless of the model chosen. Monitor the impact of this effect, given the volume of information handled in large corporations, is a process (often manual) very expensive for such corporations. Because it requires investigating issues such as the adequacy of the model, allocation of known models to the sales time series, discovery of new patterns of behavior ... This article proposes an intelligent system based on support vector machines to solve problems concerning the allocation and discovery of new models. With this focus in mind, the system objective is to build groups of time series that share the same forecasting model. For the identification of new models, the system will assign "virtual models" for those groups that do not have a predefined pattern.

Keywords

Time series, Classification, ARIMA, Support vector machines, intelligent systems.

Resumen

Disponer de mecanismos que permitan anticipar la demanda es un factor clave para éxito comercial de un distribuidor. Las ventajas se pueden agrupar en torno a dos grandes conceptos: optimización de operaciones mediante la elaboración de estrategias óptimas de aprovisionamiento y la reducción de stock que permite reducir costes de almacenamiento, manipulación … En la actualidad existe una gran variedad de métodos para realizar pronósticos, desde los métodos estadísticos puros como el alisado exponencial de Holt-Winters, modelos ARIMA hasta los basados en técnicas de inteligencia artificial como redes neuronales o sistemas borrosos. Sin embargo, a pesar de poder construir modelos precisos, en la gestión de la cadena de suministro basada en pronósticos existe el problema conocido como "efecto Forrester" con independencia del modelo escogido. Controlar el impacto de este efecto, dado el volumen de información que se maneja en las grandes corporaciones, es un proceso (muchas veces manual) muy costoso para dichas corporaciones ya que requiere investigar aspectos como la adecuación de los modelo, asignación de los modelos conocidos a las series temporales de venta, descubrimiento de nuevos patrones de comportamiento... En el presente trabajo se propone un sistema inteligente basado en máquinas que vectores de soporte para resolver los problemas relativos a la asignación de modelos y descubrimiento de nuevos modelos. Con este objetivo en mente, el sistema se encarga de construir grupos de series temporales que comparten modelo de pronósticos. Para la identificación de los nuevos modelos, el sistema asignará "modelos virtuales" a aquellos grupos que no tengan un modelo predefinido.

Palabras clave

Series temporales, Clasificación, ARIMA, Máquinas de vectores de soporte, sistemas inteligentes.

Agradecimientos

Son muchas las personas a quienes debería nombrar en estas líneas, pero me quedaré con las más trascendentales, con aquellas que no han bajado la guardia y siempre me han apoyado, tanto a lo largo del desarrollo de este Proyecto Fin de Máster como a lo largo de mi vida.

En especial a mi mujer Ana, por la paciencia y el apoyo recibidos sin los cuales no podría haber realizado el trabajo.

A mi hijo Jaime, fuente de alegría en los momentos difíciles.

A mis padres y mi tía Teresa, por haberme guiado en la vida.

A mis hermanos, por su ayuda y apoyo incondicional.

A Javier por toda la ayuda prestada, no sólo en el ámbito académico sino también en lo personal.

.

Lista de acrónimos

ACF	AutoCorrleation Function
ARIMA	AutoRegressive Integrated Moving Average model
DTW	Dynamic Time Warping
PACF	Partial AutoCorrelation Function
SARIMA	Seasonal AutoRegressive Integrated Moving Average model
SKU	Stock Keeping Unit
SVM	Support Vector Machine

ÍNDICE

1. INTRODUCCIÓN 2

1.1. OBJETO DE LA INVESTIGACIÓN 4
1.2. TRABAJOS RELACIONADOS 5
1.3. ESTRUCTURA DEL TRABAJO 8

2. SERIES TEMPORALES 9

2.1. PRESENTACIÓN DE UN CASO REAL 10
2.2. MODELOS DESCRIPTIVOS 13
2.3. MODELOS ARIMA 15
2.4. FUNCIONES DE AUTOCORRELACIÓN Y AUTOCORRELACIÓN PARCIAL 16
2.5. TÉCNICAS DE MEDICIÓN DE ERRORES EN PREVISIONES 18

3. CLASIFICACIÓN DE SERIES USANDO SVM 19

3.1. INTRODUCCIÓN A LAS MÁQUINAS DE VECTORES DE SOPORTE 19
3.2. DEFINICIÓN DE LAS CARACTERÍSTICAS DE LAS SERIES 21
3.3. ALGORITMO DEL CLASIFICADOR BASADO EN SVM 22
3.4. ALGORITMO DE GENERACIÓN DE LOS CONJUNTOS DE ENTRENAMIENTO. 24

4. SIMULACIÓN Y RESULTADOS 27

4.1. DESCRIPCIÓN DEL JUEGO DE DATOS 27
4.2. DETALLES DE LA IMPLEMENTACIÓN. 27
4.3. RESULTADOS OBTENIDOS CON UN CLASIFICADOR BASADO EN LA DISTANCIA EUCLÍDEA. 29
4.3.1. Resultados de la simulación con distancia máxima 3 30
4.3.2. Resultados de la simulación con distancia máxima 1,5 30
4.3.3. Resultados de la simulación con distancia máxima 1 31
4.3.4. Resultados de la simulación con distancia máxima 0,5 32
4.4. RESULTADOS OBTENIDOS CON EL CLASIFICADOR BASADO EN SVMS 33
4.4.1. Comentarios acerca de las principales categorías construidas. 34
4.4.1.1. Categoría 7073790000 35
4.4.1.2. Categoría 7964670000 35
4.4.1.3. Categoría 922670000 36
4.4.1.4. Categoría 1235480102 37
4.4.1.5. Categoría 864550000 38
4.4.1.6. Categoría 7346610000 39

5. CONCLUSIONES Y TRABAJO FUTURO 41

5.1. TRABAJO FUTURO 42

REFERENCIAS 44

ANEXO A: TABLAS DE DATOS 48

ÍNDICE DE TABLAS

2-1 Ejemplo de valores ACF / PACF 18
4-1 Resultados clas. euclídeo con distancia 1.5 31
4-2 Resultados clas. euclídeo con distancia 1 32
4-3 Resultados clas. euclídeo con distancia 0,5 33
4-4 Resultados clas. con SVMs 34
A-1 Datos ejemplo correspondientes a 2009 48
A-2 Datos ejemplo correspondientes a 2010 49

ÍNDICE DE FIGURAS

2-1 Datos correspondientes a 2 años 10
2-2 Sección correspondiente a 2 meses 11
2-3 Tres semanas consecutivas de venta 11
2-4 Venta agregada a nivel semana 12
2-5 Comparativa de la venta agregada a nivel semana 12
2-6 Gráfico de la función ACF 17
2-7 Gráfico de la función PACF 17
3-1 Ejemplo bidimiensional de clasificación 20
3-2 Ejemplo de proyección. 20
3-3 Vector de referencia de una serie 22
4-1 Relación entre componentes 29
4-2 Uníca catergoría con distancia euc. 3 30
4-3 Gráfica del vector 7073790000 35
4-4 Gráfica del vector 7964670000 36
4-5 Gráfica del vector 922670000 37
4-6 Gráfica del vector 1235480102 38
4-7 Gráfica del vector 864550000 39
4-8 Gráfica del vector 7346610000 40

1. INTRODUCCIÓN

Realizar previsiones de venta es una tarea muy compleja para la mayor parte de las grandes corporaciones dedicadas a la distribución de mercancías al por menor. Esta labor influye en gran medida en la cuenta de resultados ya que la mayor parte de las decisiones que se toman a nivel logístico están basadas en estas previsiones. Pongamos algunos ejemplos de los procesos de negocio que requieren de las previsiones de venta:

- Definición de la estrategia óptima de compra a los proveedores. Para minimizar costes de inventario, roturas… se compra la cantidad mínima necesaria para abastecer las tiendas / almacenes.

- Definición del stock de seguridad. En función de la venta de los SKUs se define un nivel mínimo de stock que deben tener las tiendas / almacenes para garantizar el servicio a los clientes (en caso de fallos en la cadena de suministro).

- Evaluación del rendimiento de las acciones comerciales (promociones, eventos…). Al disponer de una previsión precisa, se puede comparar la previsión realizada antes de la acción con la venta real. Esto permite evaluar objetivamente el rendimiento comercial de las diferentes promociones, campañas, anuncios …

Dada la importancia de esta tarea, estas empresas dedican muchos recursos al análisis, tratamiento y cálculo de previsiones de las series temporales de ventas. Esto es debido a que en el cálculo de las previsiones intervienen muchos factores de naturaleza muy diferente. Existen factores (dentro del análisis) que se pueden cuantificar con precisión como el tamaño medio de la cesta (tanto en número de SKUs como en unidades monetarias), frecuencia de compra, venta cruzada,… Sin embargo, también existen muchos factores que son de difícil

cuantificación como pueden ser los cambios de precio, la distancia que deben recorrer los consumidores hasta las tiendas, las acciones comerciales de la competencia…

A todos los factores anteriores, los cuales se centraban únicamente en la perspectiva metodológica del problema, hay que añadir los problemas computacionales derivados de los volúmenes de datos que se manejan en estas empresas. Esto es así, ya que en un establecimiento (gran superficie) existen a disposición del consumidor más de 10.000 SKUs diferentes, si la empresa dispone de 200 centros de ese tipo, existen más 2.000.000 de series temporales de venta. Partiendo del hecho que para poder determinar algunos comportamientos estadísticos, es necesario disponer de 2 años de datos por cada serie, será necesario almacenar más de 14.000.000.000 de registros. Llegados a este punto y aunque resulte obvio decirlo, la solución más fácil para el problema de la volumetría es la agregación de los SKUs en categorías compuestas por elementos con comportamiento similar.

Se podría pensar que el comportamiento de un SKU es parecido en todos los establecimientos, pero nada más lejos de la realidad, ya que la serie temporal de ventas de un mismo SKU puede llegar a ser completamente diferente en función de la ubicación geográfica. Este hecho resulta obvio si tomamos como ejemplo un refresco helado y analizamos su serie en dos tiendas, una en una región de montaña y otra situada cerca de una localidad turística con playa.

Otro tipo de categorización, sería el hecho por tipo de producto, el cual nos llevaría a considerar parecidos todos los refrescos de cola (por poner un ejemplo). Aunque con este enfoque, sea cierto que compartan algunos componentes como el estacional (los refrescos se venden más en verano), pero difiere enormemente en otros como la tendencia o las ventas anómalas. Continuando con el ejemplo de los refrescos de cola, la realidad es que un SKU de marca blanca no tiene la misma capacidad de ventas, ni recibe los mismos

estímulos comerciales (anuncios, promociones…) que los productos líderes del mercado, lo que hace necesario disponer de modelos diferentes para ellos.

Esta falta de clasificación, desde el punto de vista del cálculo de previsiones, representa otro problema que debe ser resuelto en el presente trabajo ya que ni se dispone de una categorización de referencia. Esto nos plantea dos dificultades añadidas: la primera, no se dispone de un criterio para validar si la categorización generada por el sistema propuesto es más eficiente que lo ya existente y por otro lado, esa ausencia implica que no existen juegos de datos sobre los que construir los conjuntos de entrenamiento de las SVMs.

1.1. Objeto de la investigación

El trabajo de la realización de previsiones de venta se centra en torno a dos temas diferenciados: la definición de modelos para el cálculo de las previsiones y la clasificación de los SKUs en categorías para posteriormente asignarles un modelo común de previsiones.

La clasificación de los SKUs, entendida como paso previo a la asignación del modelo, es una tarea que debe realizarse de forma periódica y con una frecuencia lo más baja posible. Esto garantiza que en todo momento cada SKU tiene asignado el mejor modelo de previsiones conocido. Por otro lado, la identificación de nuevos modelos, labor que requiere grandes esfuerzos en investigación, se debe realizar solamente cuando existan categorías que no tengan un modelo asociado.

Dadas las especiales características de las series que nos ocupan, hemos elegido los modelos estadísticos ARIMA para el cálculo de las previsiones.

Por tanto, el objetivo del presente trabajo es el de definir un sistema inteligente que se encargue de la clasificación de la series temporales de venta con el objetivo de minimizar el número de modelos de previsiones a construir.

Es decir, el sistema deberá agrupar en una misma categoría todas aquellas series que puedan compartir un mismo modelo ARIMA de cara a la realización de las previsiones.

1.2. Trabajos relacionados

El estudio de series temporales y la realización de previsiones sobre ellas son temas que han sido estudiados con profundidad, dando lugar a muchos trabajos de investigación de diferentes especialidades. Sin embargo, la gran mayoría se centra en uno de los dos aspectos del problema, o bien en la clasificación o bien en la generación de previsiones para las series.

Entre los trabajos relativos a la gestión de la cadena de suministro basada en previsión de demanda, destaca especialmente el trabajo realizado por Lee, Padmanabhan y Whang [1] en el cual se demuestra matemáticamente que en una cadena de suministro compuesta por una serie de empresas, cada una de ellas compra mercancía a la siguiente de la serie, la varianza de la variable aleatoria de los pedidos (que realizan las empresas intermediarias) es mayor que la varianza de las ventas que registra la primera empresa de la serie.

Esto implica que la propia metodología de trabajo genera ruido en el sistema, con independencia de los modelos elegidos realizar las previsiones. Cuantificar el impacto de este efecto, el cual fue esbozado por Forrester en los años 60 [4], ha sido objeto de estudio en numerosas ocasiones. Por citar algunas de ellas, en los trabajos de Mutters [2] o Zhen, Drezner y otros [3] se intenta medir su impacto en función de otros valores como el tiempo de entrega y cómo se verían afectadas las previsiones si los integrantes de la cadena de suministro compartiesen la información disponible. En su trabajo Mutters concluye que eliminando el efecto se puede obtener una mejora de los resultados que oscila entre el 10 y el 30%.

En la realización de previsiones con modelos ARIMA, Shukla y Jharkharia en

[5] proponen un modelo ARIMA(3,0,3) aplicado a la demanda de productos frescos (frutas y vegetales) con resultados aceptables. Aplicados a otro tipo de SKU totalmente diferente, Erdiger y Akar [6] utilizan modelos ARIMA para la previsión de demanda de combustible en Turquía.

Recientemente se han realizado trabajos en los que se compara el nivel de ajuste (de las previsiones) entre modelos estadísticos y modelos basados en técnicas de inteligencia artificial. Un caso de ello es [7] en el que Wang realiza una comparación entre un modelo ARIMA y un modelo basado en lógica borrosa para pronosticar las exportaciones en Taiwán. Otro trabajo anterior es el realizado por Shahrabi, Mousavi y Heydar [8] dónde compraran 5 técnicas (3 e las cuales son estadísticas): ARIMA, alisado exponencial, alisado exponencial con tendencia, redes neuronales y máquinas de vectores de soporte.

El problema de la clasificación de series temporales (usando técnicas de inteligencia artificial) ha sido abordado desde diferentes perspectivas. En [11] Ratanamahatana y Keogh proponen un clasificador utilizando la técnica conocida como Dynamic Time Warping (DTW), comparando los resultados obtenidos con un clasificador basado en la distancia euclídea.

Otra aproximación, para la clasificación de documentos manuscritos o paros cardíacos, la realizan Li Wei y Keogh en [10] usando técnicas de aprendizaje máquina sobre conjuntos de datos sin etiquetar. A esta técnica la denominaron aprendizaje máquina semi-supervisado. Un enfoque diferente, es el que propone Guerts en [12] donde aplica técnicas de reconocimiento de patrones a la clasificación de series temporales. Un trabajo que combina dos técnicas es el Orsernigo y Vercellis [9], en el que propusieron un sistema que combinaba máquinas de vectores de soporte con DTW, obteniendo mejoras en la precisión entre el 3 y 5 %.

Otro trabajo en el que se combinan varias técnicas de inteligencia artificial es el realizado por Martin Stepnicka y otros [24] en el que comparan una

implementación comercial de modelos ARIMA con dos sistemas híbridos compuestos por un componente común basado en lógica borrosa y una red neuronal o una máquina de vectores de soporte. En sus simulaciones obtienen resultados comparables a los obtenidos con la implementación comercial.

En otros ámbitos, como el sector financiero, las máquinas de vectores de soporte también han sido usadas para realizar previsiones. Un ejemplo de ello, es el trabajo de Cao [25] en el que propone un sistema que combina un mapa auto-organizado de características y SVMs para la realización las previsiones. Es sus simulaciones, el uso combinado de ambas técnicas mejora los resultados obtenidos únicamente con SVMs.

Otro estudio como el realzado por KIM [26] en 2003, también dentro del sector financiero, muestra cómo se pueden utilizar las SVMs para realizar previsiones de las series temporales compuestas por los precios de las acciones en el mercado de valores. En sus simulaciones, los resultados obtenidos con las SVMs son mejores que los obtenidos con redes neuronales (basadas en retro-propagación) o sistemas de razonamiento basado en casos.

En un trabajo previo (del año 2001), Cao y Tay [27] obtuvieron resultados parecidos al comparar las previsiones realizas por SVMs con las previsiones calculadas por una red neuronal alimentada por retro-propagación. Para medir la calidad de las previsiones utilizaron varios indicadores estadísticos, como el error absoluto medio, obteniendo siempre los mejores valores con las máquinas de vectores de soporte.

Otro enfoque, parecido al que se propone en este trabajo, es el realizado por Da-yong Zhang, Hong-wei Song y Pu Chen [28] en el que se construye un estimador combinando las previsiones realizadas por un modelo ARIMA y una máquina de vectores de soporte. Este estudio se centra en la series temporales compuestas por los valores de índices S&P 500 y Nikkei 225, obteniendo mejores resultados que los modelos ARIMA o las máquinas de vectores de

forma individual.

En este trabajo se propone un sistema basado en dos técnicas conocidas, SVM para la clasificación de series y los modelos estadísticos ARIMA para el cálculo de previsiones.

1.3. Estructura del trabajo

El resto del trabajo está organizado en 4 capítulos con la estructura que se comenta a continuación.

El Capítulo 2 contiene una introducción a las series temporales y a algunos métodos estadísticos para el cálculo de previsiones. Además se profundiza en los métodos para identificar y construir modelos ARIMA en función de los resultados de las funciones de autocorrelación y autocorrelación parcial (ACF y PACF respectivamente).

El Capítulo 3 presenta un breve estado del arte referente a la clasificación de series temporales, así como otra introducción a la técnica elegida (máquinas de vectores de soporte), el algoritmo para la construcción de los juegos de datos de entrenamiento y el algoritmo del clasificador.

El Capítulo 4 contiene los resultados obtenidos de las simulaciones realizadas.

Por último, el Capítulo 5 muestra las principales conclusiones extraídas de este trabajo así como algunas líneas futuras de trabajo.

2. SERIES TEMPORALES

Una serie temporal o cronológica es una secuencia de datos, observaciones o valores, medidos en determinados momentos del tiempo, ordenados cronológicamente y, normalmente, espaciados entre sí de manera uniforme. El análisis de series temporales comprende métodos que ayudan a interpretar este tipo de datos, extrayendo información representativa, tanto referente a los orígenes o relaciones subyacentes como a la posibilidad de extrapolar y predecir su comportamiento futuro.

El análisis más clásico de las series temporales se basa en la suposición de que los valores que toma la variable de observación es la consecuencia de cuatro componentes, cuya actuación conjunta da como resultado los valores medidos, estos componentes son:

- Tendencia, indica la marcha general y persistente del fenómeno observado, es una componente de la serie que refleja la evolución a largo plazo. Por ejemplo, la tendencia creciente del consumo de productos de marca blanca.

- Variación estacional. Es el movimiento periódico de corto plazo. Se trata de una componente causal debida a la influencia de ciertos fenómenos que se repiten de manera periódica en un año (las estaciones), una semana (los fines de semana) o un día (las horas puntas) o cualquier otro periodo. Recoge las oscilaciones que se producen en esos períodos de repetición.

- Variación cíclica. Es el componente de la serie que recoge las oscilaciones periódicas de amplitud superior a un año. movimientos normalmente irregulares alrededor de la tendencia, en las que a diferencia de las variaciones estacionales, tiene un período y amplitud variables,

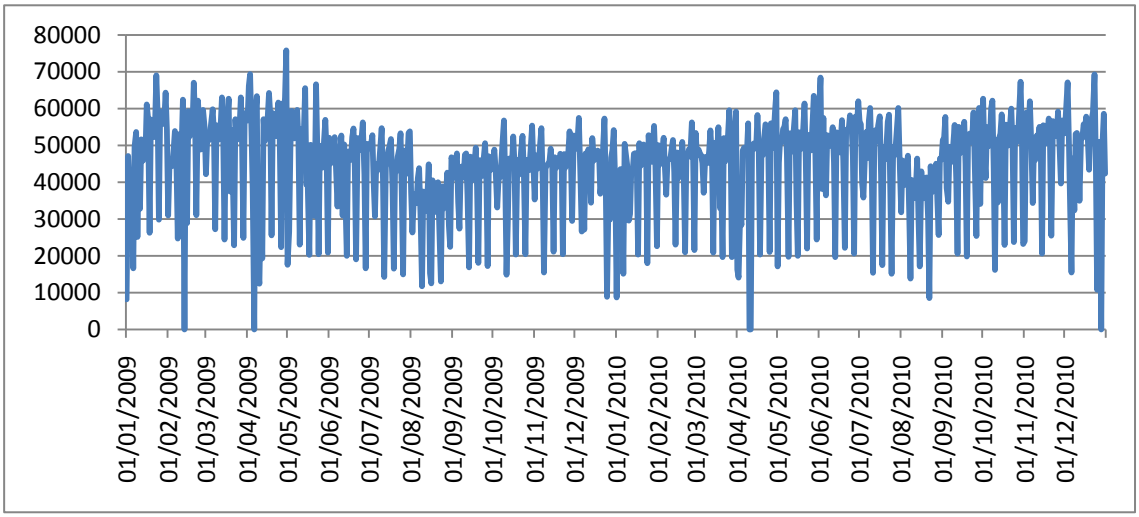
pudiendo clasificarse como cíclicos, cuasi-cíclicos o recurrentes.

- Variación aleatoria, accidental, de carácter errático, también denominada residuo, no muestran ninguna regularidad, debidos a fenómenos de carácter ocasional como pueden huelgas de transportes, eventos deportivos (Mundial de futbol) etc.

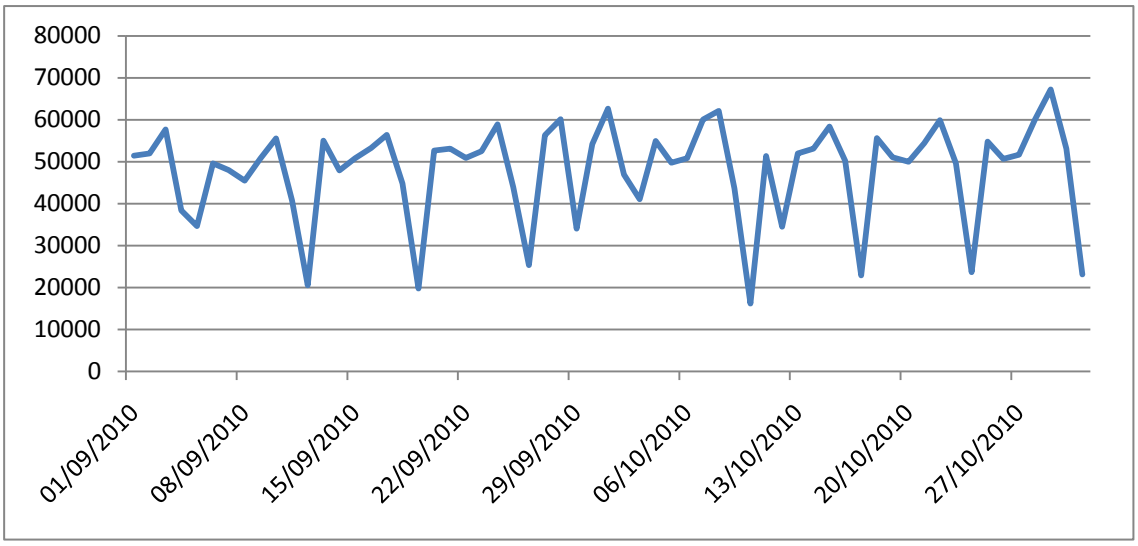
El resto del capítulo se estructura de la siguiente forma: En el apartado 2.1 se hace una presentación de una serie temporal con datos reales. En el apartado 2.2 se presentan algunos modelos estadísticos descriptivos. El apartado 2.3 se centra en la identificación y cálculo de los modelos ARIMA, definiendo además las características de la serie sobre las que va a trabajar el clasificador.

2.1. Presentación de un caso real

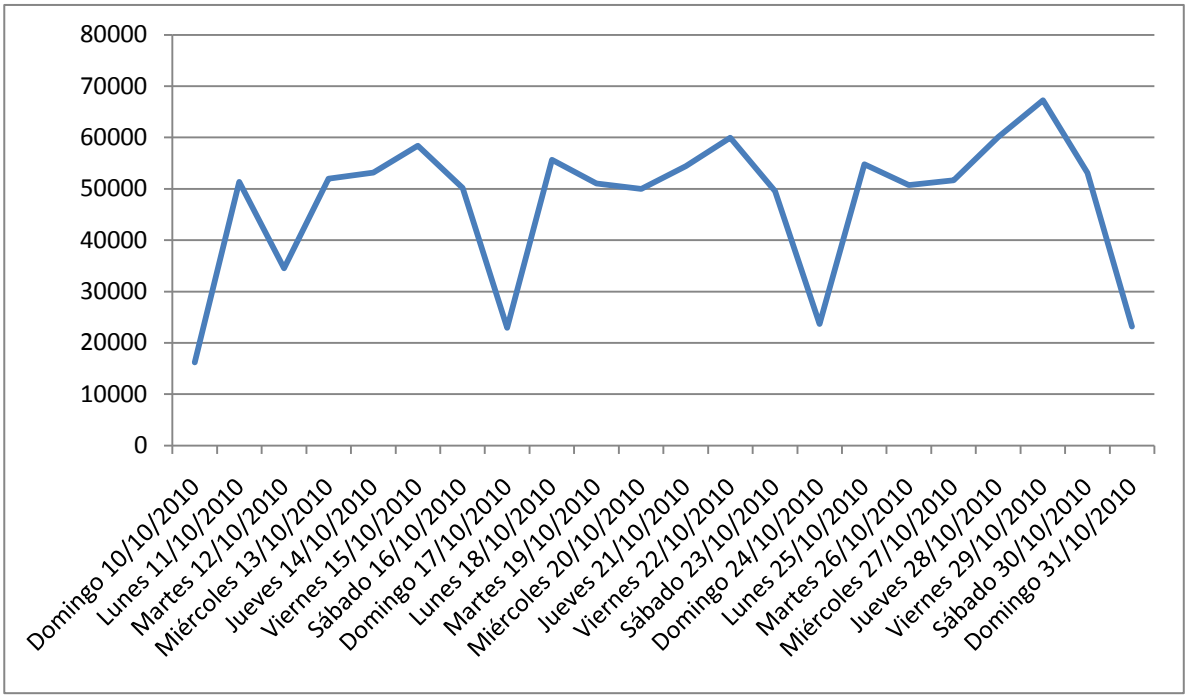
A continuación se muestran gráficamente los valores de la serie temporal correspondiente a un SKU (combustible fósil, perteneciente a una gasolinera de un centro comercial) durante 2 años consecutivos, una sección de la misma compuesta por 2 meses, otra sección de 3 semanas incluidas en los 2 meses anteriores, las ventas agregadas a nivel semana y una comparativa de la venta agregada. El juego de datos a nivel día está disponible en el Anexo A.



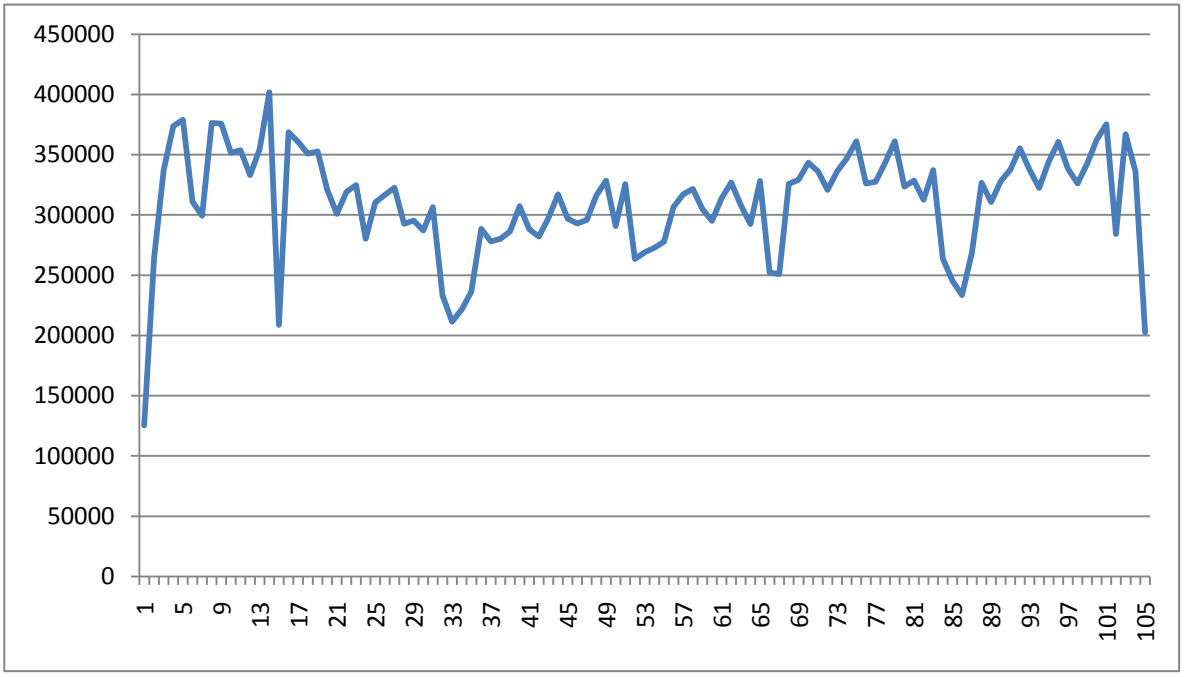
2-1 Datos correspondientes a 2 años



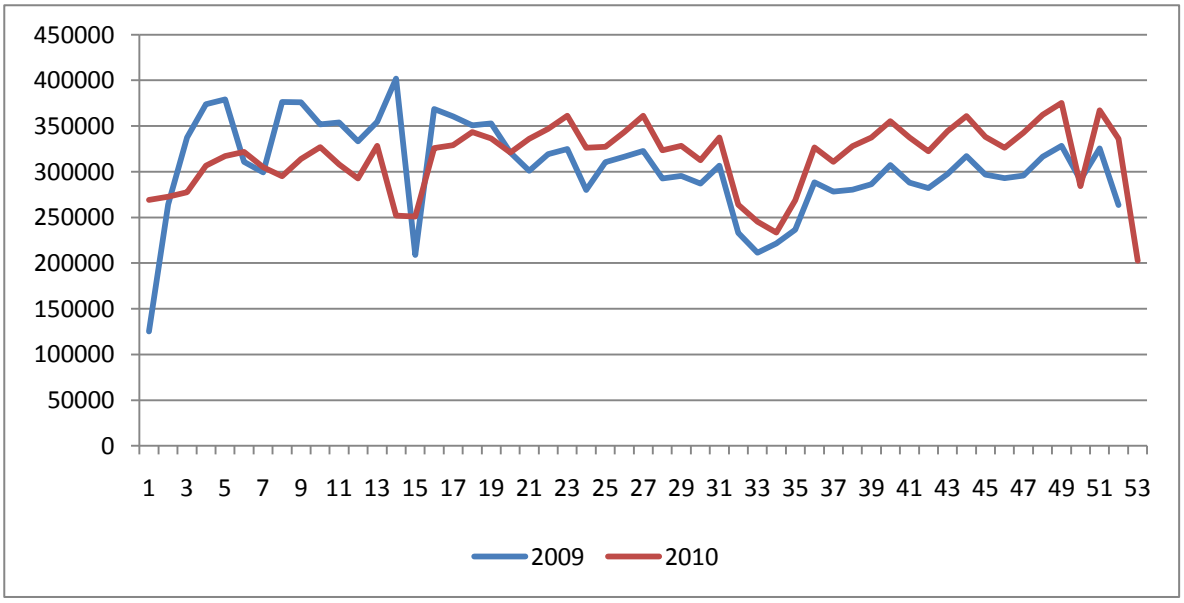
2-2 Sección correspondiente a 2 meses



2-3 Tres semanas consecutivas de venta



2-4 Venta agregada a nivel semana



2-5 Comparativa de la venta agregada a nivel semana

En el estudio de la serie se observan ciertas particularidades de este tipo de series:

- Dentro del ciclo anual de ventas, se observa una reducción

significativa de las ventas en los meses de agosto. También se observa una ligera tendencia al alza en las ventas del año 2010.

- En el ciclo mensual de ventas se observan semanas con un comportamiento diferente. Esto es debido a la apertura de extraordinaria de algunos domingos, lo que provoca un incremento de las ventas.

- Dentro de los ciclos semanales de venta existe una distribución claramente definida de las ventas a nivel día, siendo los viernes los días con mayor venta dentro de la semana. También se aprecia el efecto de los días festivos en las ventas (en la figura 2.1.3 está incluido el día del Pilar, 12 de Octubre) que se deben considerar como ventas anómalas.

En el estudio y análisis de las series temporales se realizan una serie de tratamientos estadísticos como la detección de datos anómalos, alisamiento … previos a la realización de previsiones. Este tipo de tratamientos previos está fuera del alcance del presente trabajo y se presupone realizado sobre las series temporales.

2.2. Modelos descriptivos

En este apartado realizaremos una pequeña introducción a los métodos estadísticos descriptivos, cuyo objetivo es explicar la evolución pasada de la serie para predecir sus valores futuros.

En general, los métodos descriptivos se pueden agrupar en torno a dos categorías: los modelos aditivos y los modelos multiplicativos en función de la operación que relaciona los términos (suma o producto). Es decir, si T representa el término de la tendencia, E el correspondiente a la estacionalidad, C el referente al componente cíclico y A el componente aleatorio, las fórmulas

generales serían:

- Modelo aditivo: $X_t = T_t + E_t + C_t + A_t$
- Modelo multiplicativo: $X_t = T_t * E_t * C_t * A_t$

En algunos casos, como es el que nos ocupa, es necesario dar mayor importancia a los datos recientes frente a los más antiguos. Esto se consigue utilizando modelos con parámetros variables (los cuales son estimados en función de unos valores iniciales). Estos métodos, que se denominan de alisado exponencial, son los más apropiados para las series temporales que presentan estacionalidad y tendencia.

El método más representativo es el de alisado exponencial de Holt-Winters [13,14], cuya formulación en el caso multiplicativo (siendo alpha, gamma y beta los parámetros de alisado) sería:

$$S_t = \alpha * \frac{y_t}{I_{t-L}} + (1-\alpha) * (S_{t-1} + b_{t-1}) \quad con\ 0 < \alpha < 1$$

$$b_t = \gamma * (S_t - S_{t-1}) + (1-\gamma) * b_{t-1} \quad con\ 0 < \gamma < 1$$

$$I_t = \beta * \left(\frac{y_t}{S_t}\right) + (1-\beta) * I_{t-L+m} \quad con\ 0 < \beta < 1$$

$$F_{t+m} = (S_t + m * b_t) * I_{t-L+m}$$

Dónde:

- y representa los valores observados.
- S representa los valores observados alisados.
- b representa la tendencia.
- I es el componente estacional.
- F es la predicción de la serie para m períodos futuros.

- L es el número de periodos que componen un ciclo.

2.3. Modelos ARIMA

Los modelos ARIMA, también conocidos como modelos de Box-Jenkins, representan un papel importante en el campo de las series temporales. Son capaces de recoger la tendencia y la estacionalidad de los datos pero, a diferencia del método de Holt-Winters, su filosofía no se basa en la descomposición de las series en tales factores.

Antes de definir propiamente estos modelos es necesario fijar unos conceptos previos:

- Series estacionarias. Se llaman estacionarias aquellas series cuyos valores fluctúan alrededor de una media constante y el grado de dispersión no varía en el tiempo. La combinación lineal de series estacionarias es una serie estacionaria.

- Operador retardo. Definimos el operador retardo B como la función que desfasa la serie un término. O lo que es lo mismo: $By_t = y_{t-1}$

- Operador diferencia. El operador diferencia de orden 1 se define como: $\nabla y_t = (1 - B)y_t = y_t - y_{t-1}$. En general, una diferencia de orden d se puede escribir como: $\nabla^d y_t = (1 - B)^d y_t$

- Se dice que una serie temporal es un proceso autorregresivo de orden p si el término n se puede expresar como una combinación lineal de los p anteriores y un término de error. Su formulación sería:

$$y_t = \theta_1 * y_{t-1} + \theta_2 * y_{t-2} + \cdots + \theta_p * y_{t-p} + Z_t$$

- Se dice que una serie temporal es un proceso de medias móviles de orden q si el término n se puede expresar como una suma ponderada de los últimos q errores. Su formulación sería:

$$y_t = Z_t + \mu_1 * Z_{t-1} + \mu_2 * Z_{t-2} + \cdots + \mu_q * Z_{t-q}$$

Así, un modelo ARIMA(p, d, q) se define como una combinación de un proceso autorregresivo de orden p y un proceso de medias móviles de orden q donde la serie original ha sido diferenciada d veces para eliminar la tendencia. Así, el modelo se podría expresar como:

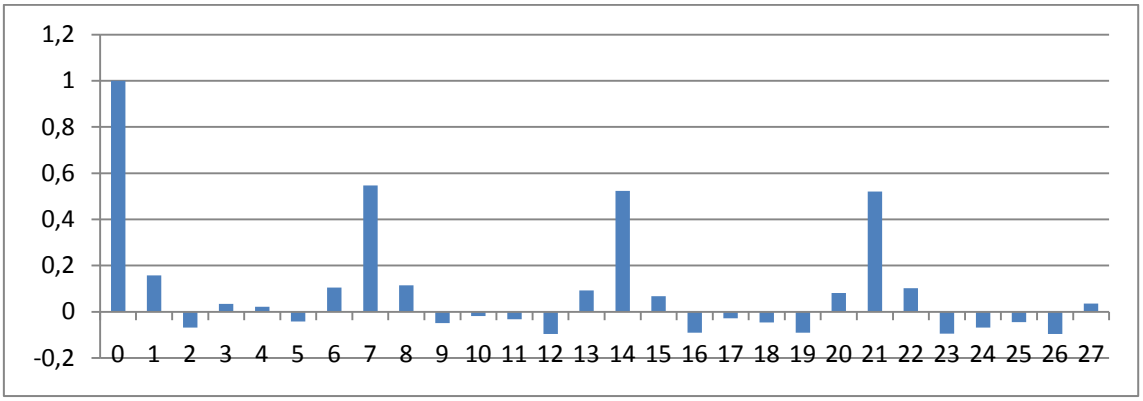
$$\left(1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_p B^p\right)(1 - B)^d Y_t = \left(1 - \mu_1 B_1 - \mu_2 B^2 - \cdots - \mu_q B^q\right) Z_t$$

Estos modelos están pensados para series sin componente estacional. Para tenerlo en cuenta, se desarrollaron los modelos S-ARIMA (Seasonal ARIMA) que son resultado de combinar dos modelos ARIMA, uno para la parte regular y otro para la estacional. Se los denota como $ARIMA(p, d, q) \times (P, D, Q)_S$ dónde (p, d, q) son los parámetros del "modelo regular", (P, D, Q) son los parámetros del "modelo estacional" y S es periodo estacional.

2.4. Funciones de autocorrelación y autocorrelación parcial

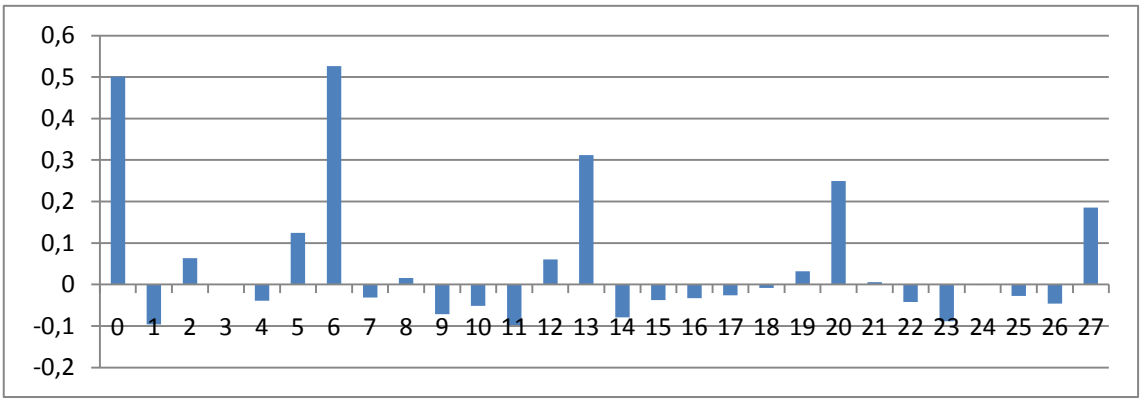
Para la facilitar la estimación de los parámetros (p,d,q) de un modelo ARIMA, Box y Jenkins [15] definieron un par funciones conocidas como autocorrelación (ACF) y autocorrelación parcial (PACF).

La función ACF representa el grado de similitud entre una serie de tiempo dada, y una versión de sí misma más retrasada k intervalos de tiempo sucesivos. Es el mismo que el cálculo de la correlación entre dos series de tiempo diferentes, salvo que la serie de tiempo se utiliza el mismo dos veces: una vez en su forma original y una vez retrasados uno o más períodos de tiempo. La siguiente gráfica muestra su cálculo para la serie de ejemplo:



2-6 Gráfico de la función ACF

La función PACF se calcula entre dos términos y_t e y_{t+k} como la autocorrelación entre ellos sin tener en cuenta la dependencia lineal entre ellos. Con los valores que calcula se puede estimar el parámetro p del modelo ARIMA (términos de autorregresión). A continuación se muestra gráficamente los resultados que genera para la serie de ejemplo:



2-7 Gráfico de la función PACF

Los datos numéricos de las gráficas anteriores son:

ACF		PACF	
Posición	**Valor**	**Posición**	**Valor**
0	1	**0**	0,5005745
1	0,1578773	**1**	-0,09594242
2	-0,0686258	**2**	0,06324612
3	0,03408096	**3**	-0,00163266
4	0,02185287	**4**	-0,03890517
5	-0,04178672	**5**	0,1247569
6	0,1056074	**6**	0,5265353

7	0,5474993	**7**	-0,03128314
8	0,1147253	**8**	0,0153725
9	-0,0492512	**9**	-0,07117078
10	-0,0190471	**10**	-0,05115186
11	-0,03232445	**11**	-0,09778934
12	-0,0959074	**12**	0,06055041
13	0,09202108	**13**	0,3121339
14	0,5233232	**14**	-0,07946364
15	0,06781492	**15**	-0,03763084
16	-0,0897178	**16**	-0,032858
17	-0,0275454	**17**	-0,02563726
18	-0,04595095	**18**	-0,00796526
19	-0,08980832	**19**	0,03227846
20	0,08070604	**20**	0,2491688
21	0,5206176	**21**	0,00569842
22	0,1022317	**22**	-0,04181279
23	-0,09429995	**23**	-0,08814423
24	-0,06836006	**24**	-0,00161713
25	-0,04537733	**25**	-0,0277175
26	-0,09564992	**26**	-0,04638965
27	0,03500869	**27**	0,1856221

2-1 Ejemplo de valores ACF / PACF

2.5. Técnicas de medición de errores en previsiones

Debido a la componente aleatoria de las series, de cara a poder mediar la calidad de las previsiones, es necesario definir indicadores estadísticos que midan el error cometido. A continuación se enumeran algunos de los más usados.

- Error cuadrático medio (ECM): $M = \frac{1}{N}\sum_{i=1}^{N}(x_i - f_i)^2$
- Error absoluto medio (MAE): $M = \frac{1}{N}\sum_{i=1}^{N}|x_i - f_i|$
- Error porcentual medio (MAPE): $M = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{x_i - f_i}{f_i}\right|$

Dónde x son los valores reales y f las predicciones realizadas por el modelo.

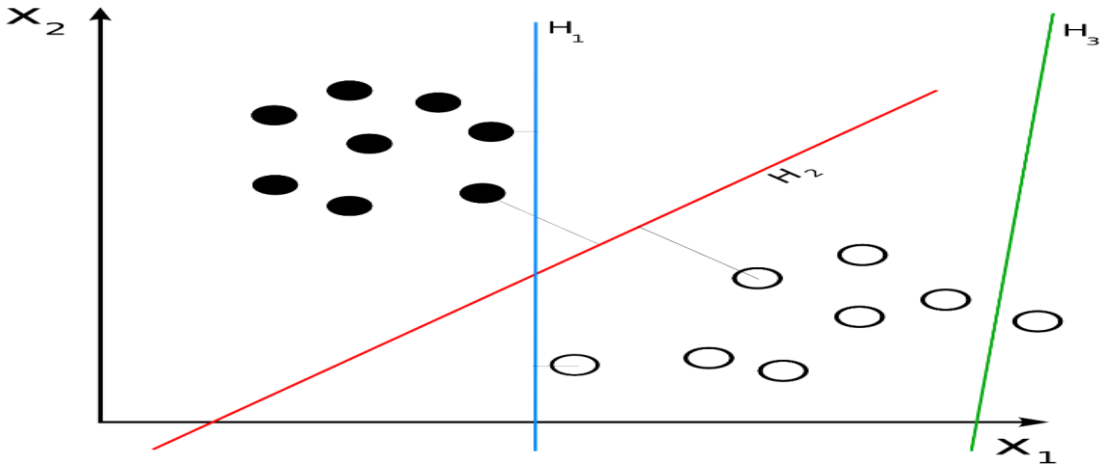
3. CLASIFICACIÓN DE SERIES USANDO SVM

Una vez establecidos los modelos estadísticos para las previsiones, es necesario definir la estructura del clasificador. Para ello, se hará una breve introducción a la técnica de inteligencia artificial escogida (SVM) y se definirán las características sobre las que trabajará el clasificador. Además, como no se dispone de un juego de datos de entrenamiento, será necesario desarrollar un algoritmo de construcción de los datos de entrenamiento.

3.1. Introducción a las máquinas de vectores de soporte

Las máquinas de vectores de soporte [23] son un conjunto de algoritmos de aprendizaje supervisado. Estos métodos están propiamente relacionados con problemas de clasificación y regresión, dado un conjunto de ejemplos de entrenamiento (de muestras) se pueden etiquetar las categorías y entrenar una SVM para construir un modelo que prediga las categorías de una nueva muestra. Fueron definidas por primera vez en 1992 por Boser, Guyon y Vapnik en [21], aunque fueron revisadas por Cortes y Vapnik en [22] tres años después.

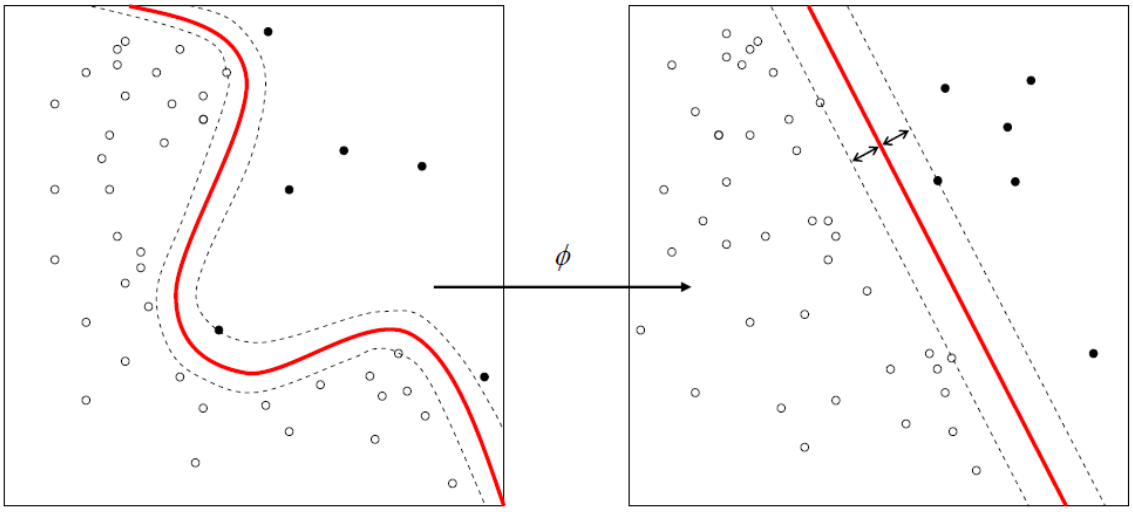
Intuitivamente, una SVM es un modelo que representa a los puntos de muestra en el espacio, separando las categorías por un espacio lo más amplio posible. Cuando las nuevas muestras se ponen en correspondencia con dicho modelo, en función de su proximidad pueden ser clasificadas a una u otra categoría.



3-1 Ejemplo bidimiensional de clasificación

En el ejemplo anterior H3 no separa los puntos, H1 realiza una separación completa y H2 hace la separación maximizando el margen entre las categorías.

La labor de clasificación la realiza la SVM mediante la búsqueda de un hiperplano que separe de forma óptima a los puntos de una clase de la de otra, que eventualmente han podido ser previamente proyectados a un espacio de dimensionalidad superior (mediante una función que denotaremos φ). Al tratarse de un proceso de aprendizaje supervisado es necesario proporcionar un juego de entrenamiento compuesto por un conjunto de vectores cuyos elementos tienen asignados unas etiquetas (las cuales indican la categoría a la que pertenece el vector).



3-2 Ejemplo de proyección.

Una vez definido el conjunto de datos de entrenamiento $(y_i, x_i)\ con\ i\epsilon[1, l], x_i \in R^n,\ y_i \in \{-1,1\}$, para obtener la SVM es necesario resolver el siguiente problema de optimización [21,22]:

$$\min_{w,b,\delta} \frac{1}{2} w^T w + C \sum_{i=1}^{l} \delta_i$$

$$\text{Sujeto a:} \quad y_i(w^T \varphi(x_i) + b\ ) \geq 1 - \ \delta_i\ , \delta_i \geq 0$$

Dónde C es la constante de penalización para los errores.

También se define la función núcleo de la SVM como $K(x_i, x_j) = \varphi(x_i)^T \varphi(x_j)$. Las funciones núcleo más comunes son:

- Lineal: $K(x_i, x_j) = {x_i}^T x_j$
- Polinomial: $K(x_i, x_j) = (\gamma {x_i}^T x_j + r)^d\ , con\ \gamma > 0$
- Base radial gaussiana: $K(x_i, x_j) = exp(-\gamma \|x_i - x_j\|^2)\ , con\ \gamma > 0$
- Sigmoide: $K(x_i, x_j) = \tanh\ (\gamma {x_i}^T x_j + r)$

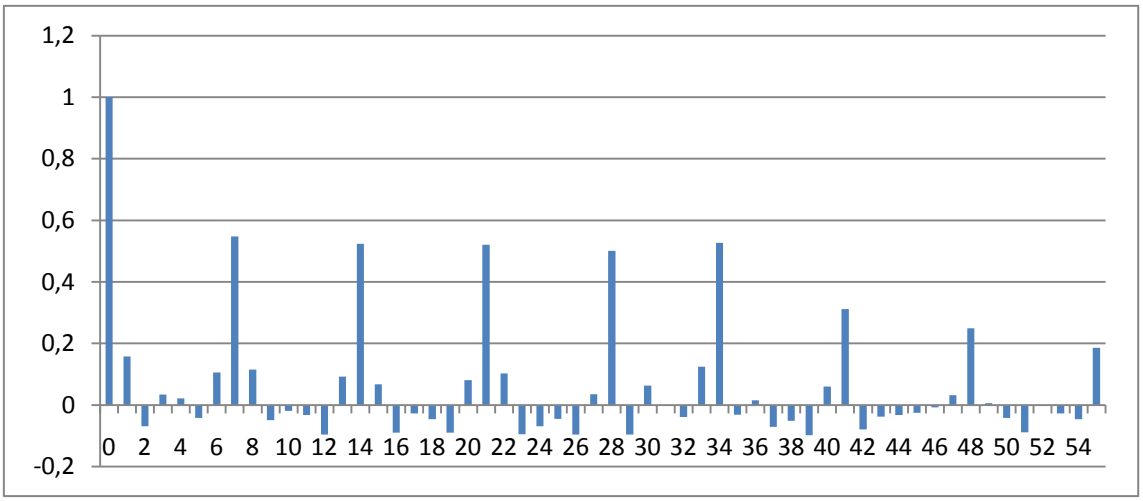
3.2. Definición de las características de las series

Siguiendo las directrices dadas por Box y Jenkins en [15], la identificación de los modelos ARIMA se basa en torno a las funciones definidas como autocorrelación y autocorrelación parcial). Los valores de dichas funciones se calculan un número determinado de términos de la serie (llamemos N a este número). Dicho esto, a dos series temporales se les podrá asignar el mismo modelo ARIMA si las funciones de autocorrelación y autocorrelación parcial dan resultados parecidos en los mismos N primeros términos.

De esta forma, definimos las características de la series temporales (de cara a la clasificación) como los resultados de las anteriores funciones sobre el número de términos regresivos. Por lo que, de aquí en adelante, podremos representarlas mediante un vector de 2N posiciones. Las primeras N posiciones representan los resultados de la función de autocorrelación (la posición 0 siempre toma valor 1) y los N últimos los resultados de la función de autocorrelación parcial.

Un hecho importante a destacar es que los valores calculados por las funciones anteriormente mencionadas toman sus valores en el intervalo [-1,1]. Esto facilita el tratamiento de las series, ya que no es necesario realizar conversiones o ajustes en los datos de entrada de las máquinas de vectores [20].

A continuación se muestra el vector del ejemplo usado en el apartado anterior:



3-3 Vector de referencia de una serie

3.3. Algoritmo del clasificador basado en SVM

Una vez definidas las características de las series sobre las que se va realizar la clasificación, es el momento de definir dicho algoritmo. El proceso de clasificación se plantea como un proceso iterativo en el que se utiliza una SVM diferente para cada categoría. De esta forma, cada categoría se puede

representar mediante una terna <R,svm(R),S(R)> compuesta por un vector de referencia, la SVM que reconoce los elementos pertenecientes y el listado de elementos asignados a ella.

1. Sea C el conjunto de categorías. $C = \emptyset$
2. Sea D el conjunto de vectores a clasificar.
3. Para cada vector d perteneciente a D hacer:
    a. Por cada elemento <X,svm(X),S(X)> de C hacer:
        i. Si svm(X) reconoce d como perteneciente: Añadir d a S(X), ir al paso 3. En caso contrario, continuar.
    b. Generar un conjunto de datos de entrenamiento usando d como vector de referencia y los parámetros necesarios (ver apartado 3.4).
    c. Entrenar una nueva SVM usando el conjunto anterior.
    d. Añadir $< d, svm(d), \{d\} >$ a C.

En el diseño del proceso se podía haber utilizado una única SVM para reconocer todas las categorías. Esto requería que cada vez que apareciese un nuevo vector de referencia, fuese necesario ampliar el conjunto de datos de entrenamiento y se volviese a entrenar la SVM . Este enfoque fue descartado atendiendo a dos razones:

- Rendimiento del sistema. Con la implementación elegida, y tras varias pruebas, parecía que la labor de entrenamiento de la SVM sobre conjuntos grandes (+10000 casos) era muy costosa.
- Paralelización de las tareas de evaluación de las categorías. Contar con una única SVM limitaba una posible mejora del sistema, por la cual podría procesar varias series en paralelo.

3.4. Algoritmo de generación de los conjuntos de entrenamiento.

Dado que no se dispone de un conjunto etiquetado de datos (para cada serie) que permita entrenar una SVM, es necesario definir un algoritmo que sea capaz de generar dichos datos.

Partiendo de un vector de referencia, y con la idea de que vectores parecidos deben pertenecer a la misma categoría, un buen método de partida sería comparar los vectores mediante la distancia euclídea. Es decir, dos vectores pertenecen a la misma categoría si distan menos de un ε dado. Sin embargo, este método no es todo lo preciso que debería ya que basta con que sólo una de las coordenadas del vector diste más de ε para que el vector sea catalogado como diferente.

Para solventar este problema, se propone un método basado en distancia pero que tiene parametrizado un número máximo de diferencias. Es decir, con este método dos vectores X e Y son considerados parecidos si y sólo si $|X_i - Y_i| < \varepsilon \; \forall i \epsilon [1,2N]$ excepto quizás en una cantidad D de coordenadas.

Una vez definido el método de comparación, sólo queda por definir el algoritmo completo.

El algoritmo recibe los siguientes parámetros:

- Sea R el vector de referencia de la categoría a identificar.
- Sea α la diferencia máxima aceptable en cada una de las coordenadas vector generado respecto del vector de referencia
- Sea β el número máximo de coordenadas en las que la diferencia sea

mayor que α permitidas

- Sea M el número casos de entrenamiento a construir que pertenecen a la categoría.
- Sea P el número casos de entrenamiento a construir que no pertenecen al clúster.

El algoritmo consta de los siguientes pasos:

i. Sea I el número de casos construidos no pertenecientes a la categoría. I=0.

ii. Sea J el número de casos construidos pertenecientes a la categoría. J=0

iii. Mientras I < M hacer:

   - Generar aleatoriamente un vector X en el que X(j) esté en [-1,1], j en [1,2N].
   - Sea D = número de coordenadas tal que abs(X(i)-R(i))>alpha.
   - Si D > beta -> I=I+1; Marcar X como no perteneciente a la categoría. En caso contrario J=J+1; Marcar X como perteneciente a la categoría

iv. Mientras J < P hacer:

   - Generar un vector X en el que X(j) =0, j en [1,2N].
   - Por cada k en [1,2N+1] hacer:
     - Generar Y en [-1,1] de forma aleatoria.
     - X(k)=R(k)*(1+Y*alpha)
   - J=J+1; Marcar X como perteneciente

4. SIMULACIÓN Y RESULTADOS

4.1. Descripción del juego de datos

En este apartado describiremos el conjunto de datos sobre el que se ha realizado el trabajo. Como conjunto inicial, se obtuvo el listado completo de series temporales de un establecimiento con las ventas de un año.

Disponer de más datos de ventas sería necesario para una correcta identificación de los modelos ARIMA, ya que con sólo un año no se pueden determinar los comportamientos anuales de compra. Sin embargo, de cara a la clasificación, es más que suficiente ya que disponemos de una cantidad de valores (365 para ser exactos) que permiten el cálculo de las funciones ACF y PACF.

El listado original estaba compuesto por más de 250.000 SKUs diferentes, pero al realizar un primer estudio de los mismos se observó que había una gran cantidad de SKUs con venta esporádica (se considera venta irregular, no predecible). Para filtrar estas series, se estableció un filtro basado en la venta media, de tal forma que todas las series con venta media menor a 1 unidad fueron descartadas. Así, tras filtrar las series, el conjunto de SKUs reducido consta en torno a 14.000 elementos.

4.2. Detalles de la implementación.

Para el desarrollo de un prototipo ha sido necesario el desarrollo de una serie de programas java y el uso de un par de librerías de código abierto:

- Para el cálculo de las funciones ACF /PACF se ha utilizado la librería estadística de código abierto **R** [16].

- Para la construcción de las máquinas de vectores de soporte se ha

utilizado la librería **LIBSVM** [17].

Las clases java desarrolladas han sido:

- AbstractClassifier: Implementación abstracta del algoritmo del clasificador sobre el tipo abstracto Cluster. Delega el método de construcción de Cluster permitiendo las dos especializaciones necesarias.

- SVMClassifier: Implementación concreta de un clasificador que utiliza objetos de tipo SVMCluster para la representación de las categorías.

- EuclideanClassifier: Implementación concreta de un clasificador que utiliza objetos de tipo EuclideanCluster para la representación de las categorías.

- Cluster: Implementación de una categoría genérica. Delega el método que determina si un vector pertenece a la categoría que representa.

- EuclideanCluster: Implementación concreta de un Cluster que utiliza como función de pertenencia la distancia euclídea.

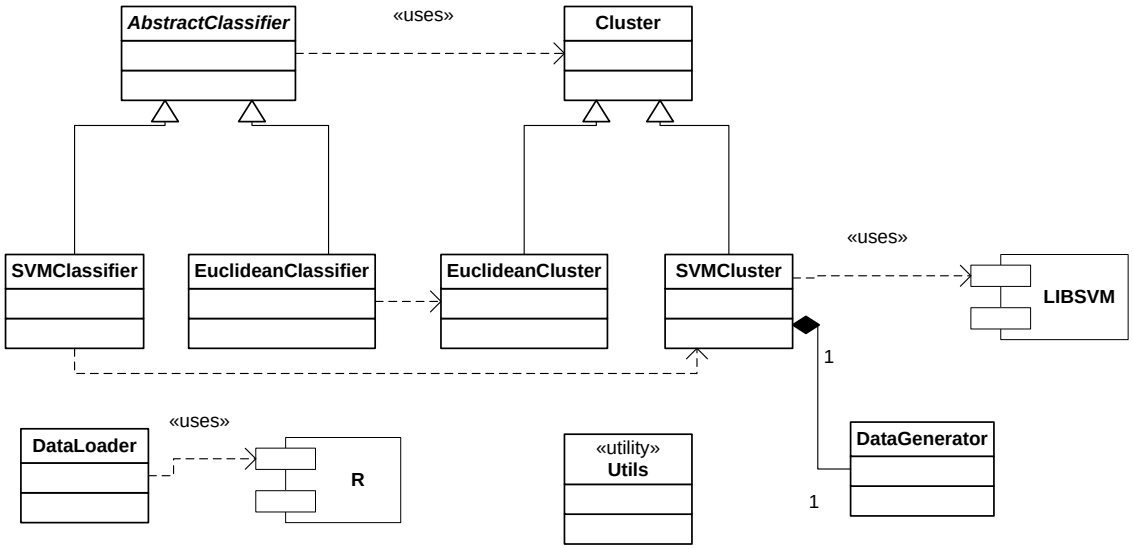
- SVMCluster: Implementación concreta de un Cluster que utiliza una SVM para reconocer los elementos pertenecientes a una categoría.

- DataGenerator: Clase de utilidad que genera los ficheros de entrenamiento de las SVMs

- DataLoader: Clase de utilidad que filtra las series temporales y calcula los valores de las funciones ACF / PACF.

- Utils: Clase de utilidad encarga del acceso a datos (lectura / escritura de ficheros …)

El siguiente diagrama muestra cómo se relacionan los diferentes elementos de software:



4-1 Relación entre componentes

En la parametrización y definición de las SVMs se ha hecho teniendo en cuenta los resultados obtenidos por Schölkopf en [18], por lo cual se han usado máquinas de tipo nu-SVC (cuyo modelo teórico define Schölkopf y otros en [19]) con núcleo basado en funciones radiales, usando los parámetros $C = 10000, \gamma = 3$.

Los ficheros de entrenamiento se han generado mediante un programa Java, con los parámetros alpha=0.2, beta=3, N=2000, P=2000.

4.3. Resultados obtenidos con un clasificador basado en la distancia euclídea.

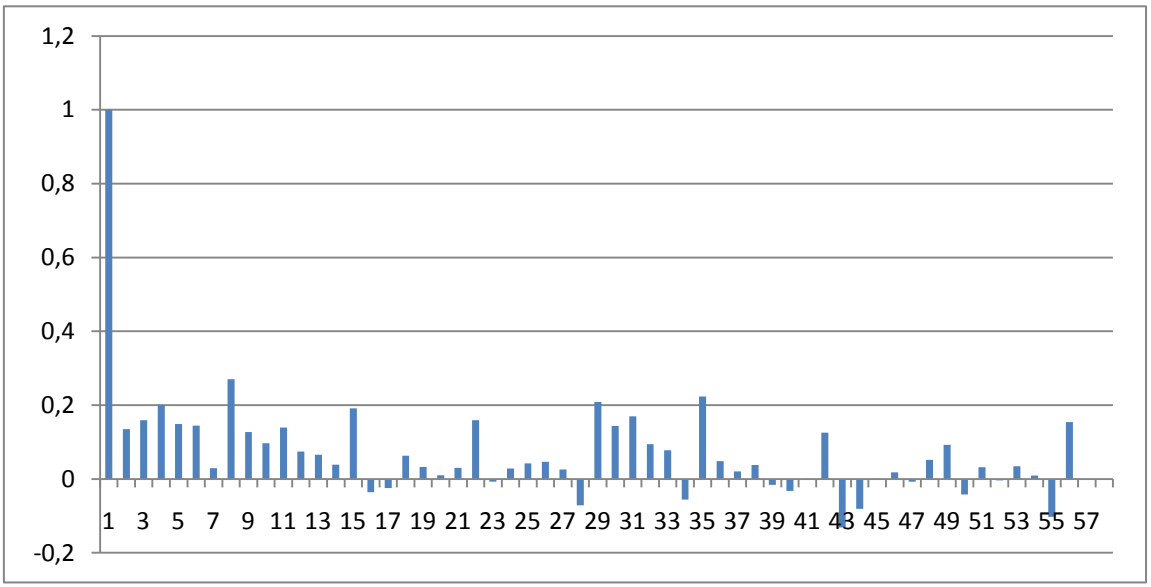
Para comparar los resultados obtenidos por el clasificador, y teniendo en cuenta que el criterio de semejanza está basado en la una función de distancia, se ha modificado el algoritmo del clasificador para que compare los vectores según la

distancia euclídea (en función de una distancia máxima dada) y se han realizado varias simulaciones sobre varios parámetros.

Antes de pasar a los resultados de las simulaciones, es necesario recalcar que al encontrarse los vectores en el espacio $[-1,1]^{56}$, la distancia máxima entre dos vectores es de $\sqrt{\sum_{i=1}^{56}(-1-1)^2} = 2 * \sqrt{56} = 14.99$

4.3.1. Resultados de la simulación con distancia máxima 3

En este caso se genera una única categoría que agrupa a todos los elementos. Obviamente, esto indica que todos los elementos se encuentran dentro de una esfera de radio 3 en torno al representante de la categoría. En esta simulación, le representante elegido es:



4-2 Única categoría con distancia euc. 3

4.3.2. Resultados de la simulación con distancia máxima 1,5

En esta ocasión, al haber reducido la distancia a la mitad el clasificador genera 6 categorías con la siguiente distribución.

Categoría	N. Elem.	%	Categoría	N. Elem.	%
7964670000	11408	79,98	7471690000	83	0,58

922670000	2577	18,07	3463790000	6	0,04
1235480102	185	1,3	1282260000	4	0,03

4-1 Resultados clas. euclídeo con distancia 1.5

4.3.3. Resultados de la simulación con distancia máxima 1

Al reducir aún más la distancia, el número de categorías aumenta hasta la 74. Cabe destacar que las seis primeras categorías agrupan al 95% de la muestra. Los vectores se distribuyen de la siguiente forma:

Categoría	N. Elem.	%	Categoría	N. Elem.	%
7073790000	6342	44,46	583360000	4	0,03
7964670000	3552	24,9	1643620000	3	0,02
922670000	2395	16,79	7951110000	3	0,02
1235480102	510	3,58	1488590000	3	0,02
864550000	438	3,07	5276720000	3	0,02
7346610000	329	2,31	5797540000	3	0,02
1425630000	130	0,91	6368510000	3	0,02
2896410000	87	0,61	2115210000	2	0,01
1308990000	67	0,47	5821040000	2	0,01
2354850000	54	0,38	1418860000	2	0,01
2448570000	32	0,22	2006730000	2	0,01
5606700000	28	0,2	7912370000	2	0,01
1282150000	23	0,16	2214730000	2	0,01
6058530000	22	0,15	7160420000	2	0,01
7968400000	19	0,13	1033050000	2	0,01
1282220000	17	0,12	2925740000	2	0,01
3772150000	15	0,11	1560690000	2	0,01
7928170000	11	0,08	503550000	2	0,01
825790000	10	0,07	1145640000	1	0,01
7610730000	10	0,07	797950000	1	0,01
5951890000	10	0,07	5632290000	1	0,01
7936600000	9	0,06	29120000	1	0,01
31550000	9	0,06	4116740000	1	0,01
1015370000	8	0,06	2550970000	1	0,01
1078810000	8	0,06	1356910000	1	0,01
4742010000	7	0,05	2556610000	1	0,01
6657420000	7	0,05	88770000	1	0,01
2617220000	6	0,04	6952890000	1	0,01
5568570000	6	0,04	6840330000	1	0,01
1186590000	6	0,04	764330000	1	0,01

1282260000	5	0,04	5621990000	1	0,01
829710000	5	0,04	7432570000	1	0,01
2923480000	5	0,04	4538470000	1	0,01
7127200000	5	0,04	2827900000	1	0,01
2910720000	5	0,04	967610000	1	0,01
7962570000	4	0,03	2352970000	1	0,01
1235470502	4	0,03	460000	1	0,01

4-2 Resultados clas. euclídeo con distancia 1

4.3.4. Resultados de la simulación con distancia máxima 0,5

Al reducir hasta la distancia máxima hasta el medio punto se consigue que el clasificador sea demasiado estricto, por lo que genera un total de 3194 categorías. En esta ocasión, las 100 primeras categorías sólo agrupan cerca del 55% de los vectores. La distribución de estas 100 primeras categorías es la siguiente:

Categoría	N. Elem.	%	Categoría	N. Elem.	%
7073790000	851	5,99	3767260000	37	0,26
1141850000	430	3,03	5332830000	36	0,25
2315720000	317	2,23	7333080000	36	0,25
6868720000	302	2,12	6509320000	35	0,25
4562980000	292	2,05	3990460000	33	0,23
5344380000	266	1,87	7466630000	33	0,23
7903990000	240	1,69	3578410000	32	0,23
6953410000	227	1,6	430000	31	0,22
294810000	209	1,47	328060000	30	0,21
6437700000	193	1,36	5818930000	30	0,21
5362400000	182	1,28	7108000000	30	0,21
7346610000	161	1,13	7369360000	30	0,21
21850000	148	1,04	7908880000	30	0,21
7938240000	146	1,03	923300000	29	0,2
1980000	128	0,9	5246410000	29	0,2
7409030000	115	0,81	5526800000	29	0,2
1504400000	107	0,75	6710040000	29	0,2
1360000	106	0,75	2039030000	28	0,2
3578740000	105	0,74	2318970000	28	0,2
7531540000	105	0,74	2348430000	28	0,2
2367790000	96	0,68	7380530000	28	0,2
1935420000	94	0,66	7905470000	28	0,2

2304460000	94	0,66	7928530000	28	0,2
7480530000	92	0,65	7945750000	28	0,2
5756530000	85	0,6	335480000	27	0,19
7913290000	79	0,56	3009900000	27	0,19
6938300000	76	0,53	5106560000	27	0,19
4121810000	72	0,51	7964670000	27	0,19
6536950000	71	0,5	44820000	26	0,18
7082610000	68	0,48	7111650000	26	0,18
2306920000	67	0,47	1830270000	25	0,18
7236220000	67	0,47	2339940000	25	0,18
3726810000	62	0,44	2648050000	25	0,18
34160000	60	0,42	698500000	24	0,17
864550000	60	0,42	7949430000	24	0,17
1423550000	58	0,41	26960000	23	0,16
2870330000	53	0,37	872980000	23	0,16
7605250000	53	0,37	6473720000	23	0,16
5085040000	51	0,36	919790000	22	0,15
523710000	50	0,35	1542400000	22	0,15
1975550000	50	0,35	2315140000	22	0,15
6796960000	49	0,34	4970300000	22	0,15
5438660000	48	0,34	7945490000	22	0,15
922670000	47	0,33	34580000	20	0,14
5150520000	43	0,3	4987890000	20	0,14
1611270000	41	0,29	6651300000	20	0,14
5114480000	41	0,29	7950570000	20	0,14
6836770000	41	0,29	7971390000	20	0,14
942400000	39	0,27	26760000	19	0,13
1968980000	39	0,27	306130000	19	0,13

4-3 Resultados clas. euclídeo con distancia 0,5

4.4. Resultados obtenidos con el clasificador basado en SVMs

El clasificador basado en SVMs cataloga el conjunto de vectores en torno a 69 categorías, con la siguiente distribución:

Categoría	N. Elem.	%	Categoría	N. Elem.	%
7073790000	6240	43,75	583360000	4	0,03
7964670000	3628	25,44	6368510000	4	0,03
922670000	2474	17,35	5821040000	3	0,02
1235480102	516	3,62	7962570000	3	0,02
864550000	433	3,04	2925740000	3	0,02

7346610000	316	2,22	1246820000	3	0,02
1425630000	131	0,92	2115210000	2	0,01
2896410000	78	0,55	2345710000	2	0,01
1308990000	67	0,47	1418860000	2	0,01
2354850000	55	0,39	7951110000	2	0,01
2448570000	31	0,22	2006730000	2	0,01
5606700000	26	0,18	2214730000	2	0,01
6058530000	24	0,17	7160420000	2	0,01
1282150000	23	0,16	5276720000	2	0,01
1282220000	20	0,14	2971340000	2	0,01
7928170000	12	0,08	6296260000	2	0,01
3772150000	11	0,08	1033050000	2	0,01
7610730000	10	0,07	1560690000	2	0,01
5951890000	10	0,07	1418870000	2	0,01
825790000	9	0,06	5797540000	2	0,01
7936600000	8	0,06	503550000	2	0,01
1015370000	7	0,05	2890330000	2	0,01
266300000	7	0,05	1145640000	1	0,01
1078810000	7	0,05	29120000	1	0,01
2910720000	7	0,05	4116740000	1	0,01
6657420000	6	0,04	1356910000	1	0,01
7127200000	6	0,04	88770000	1	0,01
1282260000	5	0,04	6952890000	1	0,01
5606790000	5	0,04	1130420000	1	0,01
5568570000	5	0,04	5621990000	1	0,01
2923480000	5	0,04	2827900000	1	0,01
1186590000	5	0,04	967610000	1	0,01
7912370000	4	0,03	2352970000	1	0,01
1488590000	4	0,03	460000	1	0,01
6733000000	4	0,03			

4-4 Resultados clas. con SVMs

En esta categorización, destaca el hecho que las 6 primeras categorías agregan al 95% de los vectores y que las 12 primeras agregan al 98% de los mismos.

4.4.1. Comentarios acerca de las principales categorías construidas.

En este apartado realizaremos unos breves comentarios acerca de las principales categorías construidas. Nos centraremos en las seis primeras

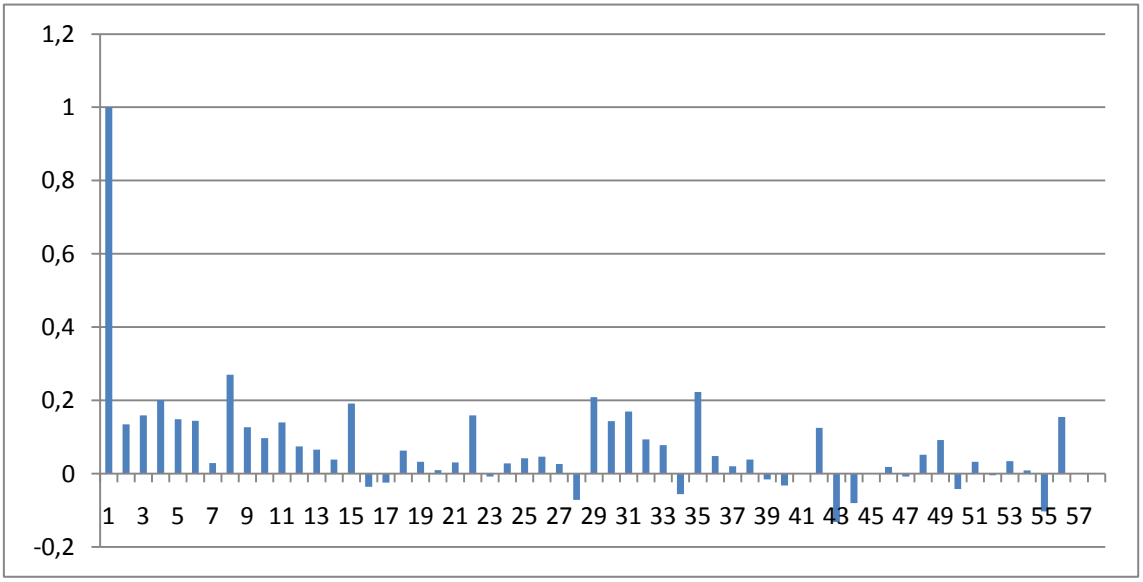
(7073790000, 7964670000, 922670000, 1235480102, 864550000, 7346610000), dado que representan al 95% del total.

4.4.1.1. Categoría 7073790000

Destaca el caso de la categoría con mayor número de elementos, el representado por la serie denominada 7073790000 que agrupa al 43,75% de las series, sea la categoría de las series que no admiten un modelo ARIMA (los valores de vector son muy cercanos a cero). Es muy probable que el filtro inicial de los datos no fuese lo suficientemente preciso y que aumentando el requisito impuesto sobre la venta media buena parte de estas series fuesen descartadas.

Sin embargo, pueden existir series en esta categoría con un volumen alto de ventas, pero que su comportamiento no se pueda explicar únicamente con los valores de las mismas.

El gráfico del vector del representante de la categoría es:



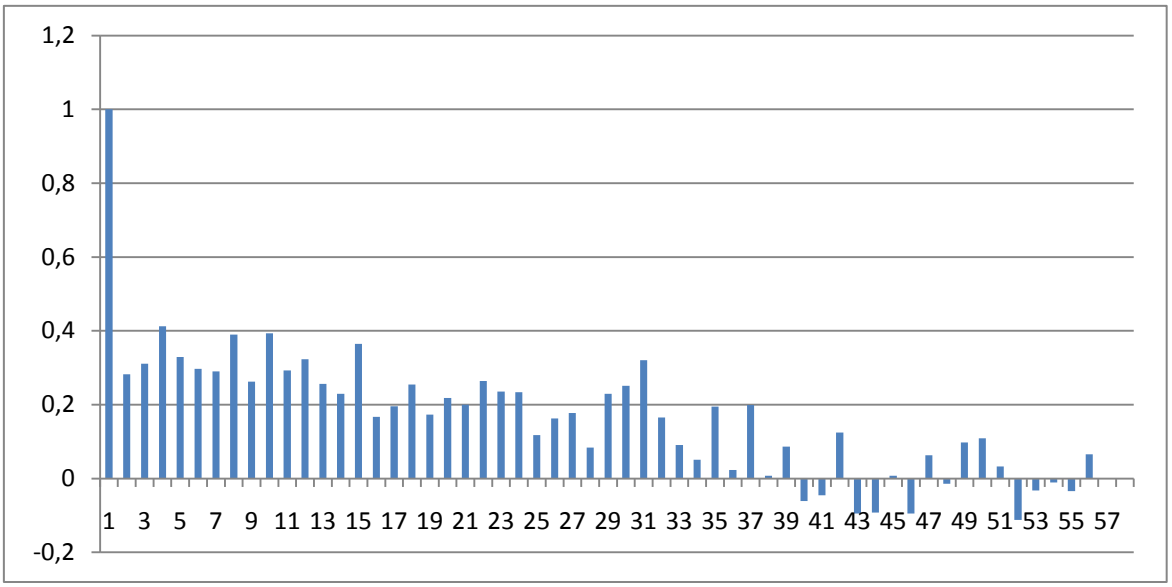
4-3 Gráfica del vector 7073790000

4.4.1.2. Categoría 7964670000

En el caso de la segunda categoría, la cual agrupa al 25,44% de las series se

observan unos valores bajos de la función ACF y prácticamente nulos de la función PACF. A priori no se le puede asociar algún modelo ARIMA sencillo.

El gráfico del vector del representante de la categoría es:



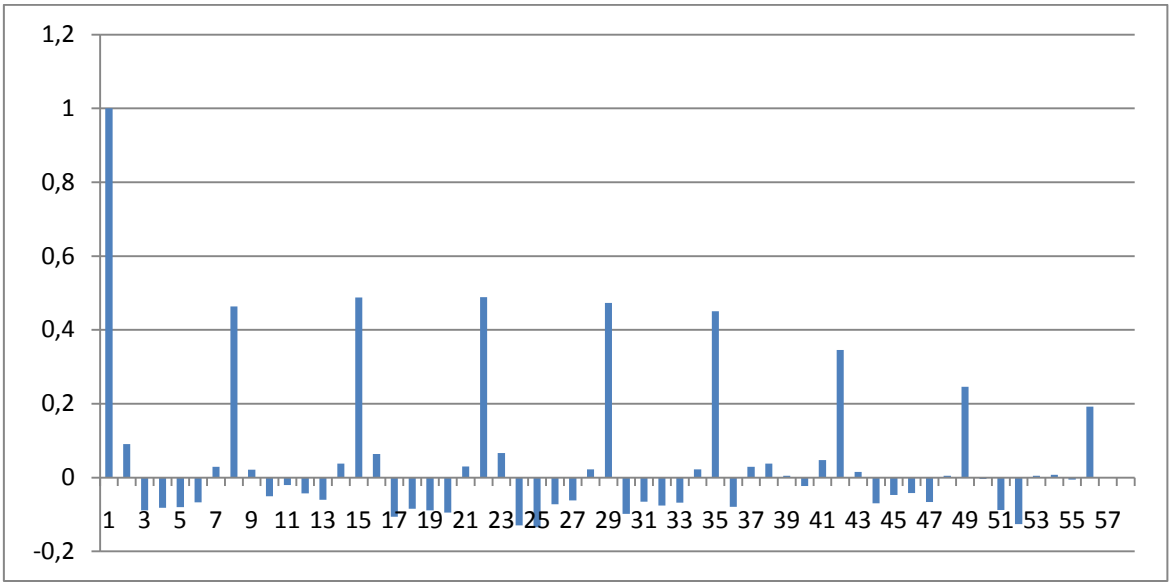
4-4 Gráfica del vector 7964670000

4.4.1.3. Categoría 922670000

El caso de tercera serie, la que agrupa a un 17,35% de las series, es quizás el más interesante desde un punto de vista estadístico. Claramente se puede apreciar una relación relativamente fuerte entre los valores con período 7. Es decir, estas series muestran una claro comportamiento cíclico de carácter semanal.

Para este tipo de series los estimadores más adecuados son los modelos ARIMA estacionales. Concretamente, se está probando con un modelo $\text{ARIMA}(0,0,2)(1,1,0)_7$ con buenos resultados.

El gráfico del vector del representante de la categoría es:

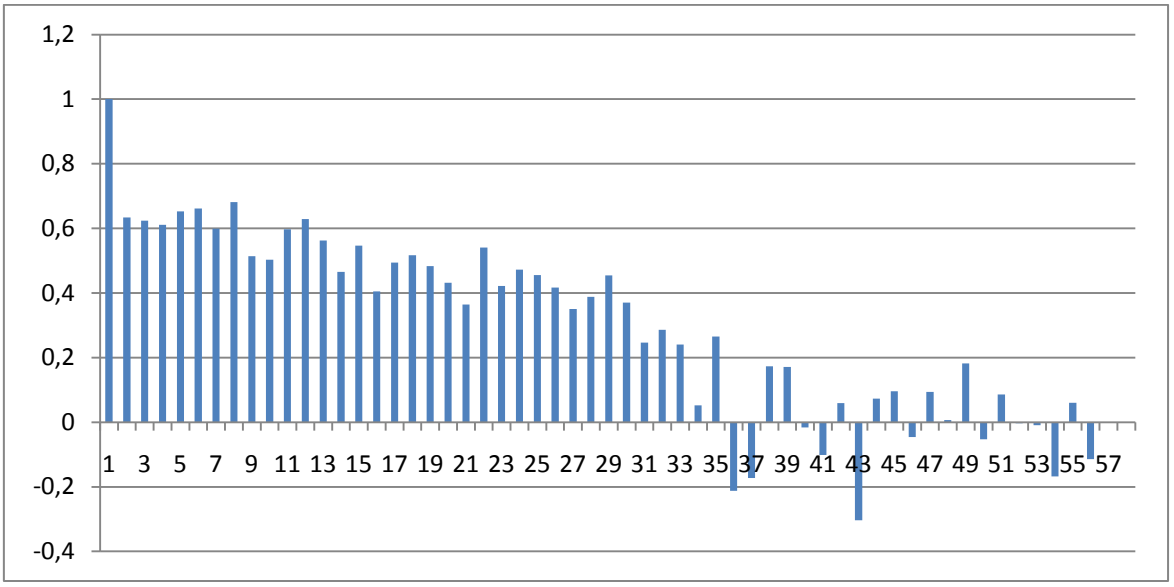


4-5 Gráfica del vector 922670000

4.4.1.4. Categoría 1235480102

El caso que nos ocupa, el de la cuarta categoría (representa al 3,62%) es parecido al de la segunda, ya que se obtienen valores altos de la función ACF. Pero en este caso, existen valores negativos mayores en los valores de la función PACF lo que indicaría un modelo ARIMA diferente.

El gráfico del vector del representante de la categoría es:

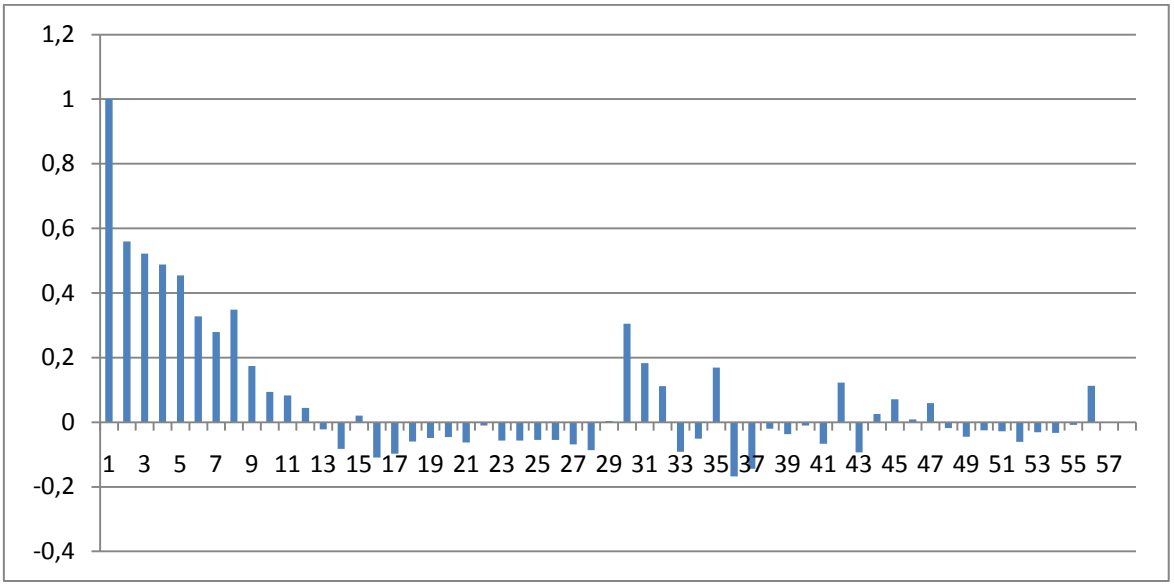


4-6 Gráfica del vector 1235480102

4.4.1.5. Categoría 864550000

La cuarta categoría (agrupa al 3,04% del total) muestra una gráfica conocida en los ejemplos académicos de los modelos ARIMA. Muy probablemente, estas series se pueden estimar con un modelo ARIMA(1,0,1) ó ARIMA(2,0,0).

El gráfico del vector del representante de la categoría es:

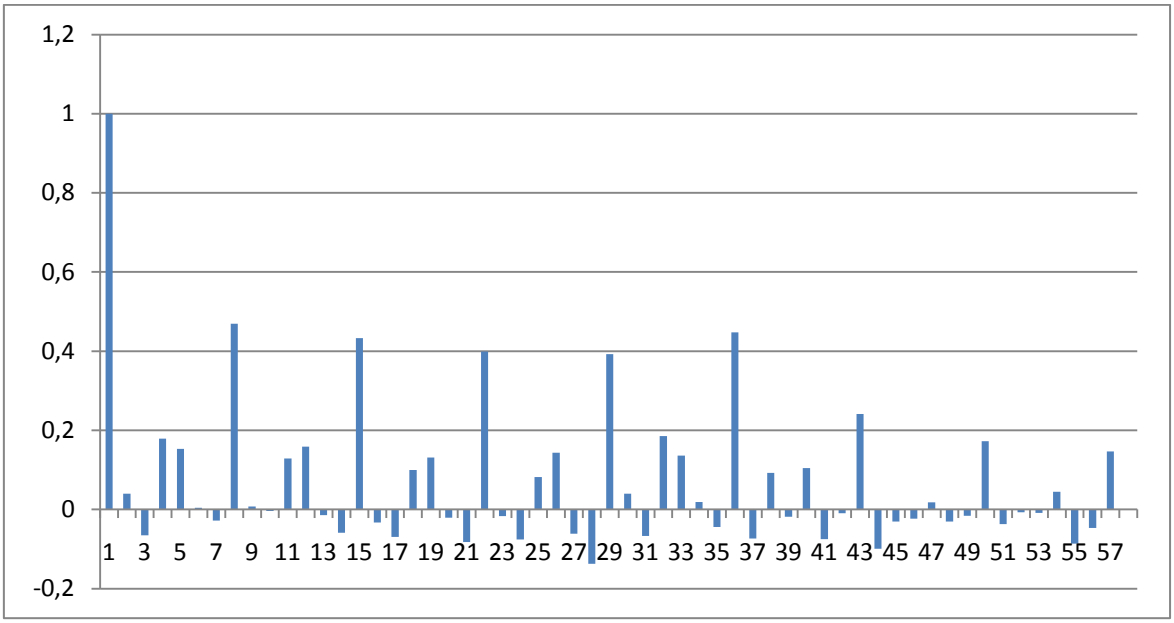


4-7 Gráfica del vector 864550000

4.4.1.6. Categoría 7346610000

El último caso, el de la categoría 7346610000 (representa al 2,22%) también muestra un comportamiento estacional pareció a la de la categoría 922670000. Sin embargo, también presenta otro patrón en las posiciones 4 y 5 de forma cíclica. En esta ocasión el modelo también debería ser un ARIMA estacional, pero es probable que los parámetros varíen.

El gráfico del vector del representante de la categoría es:



4-8 Gráfica del vector 7346610000

5. CONCLUSIONES Y TRABAJO FUTURO

El sector de la distribución se caracteriza por la rapidez con la que se producen cambios, es decir, es una cuestión conocida que las tendencias (que en muchos casos son modas) en el consumo cambian rápidamente. Y las grandes corporaciones deben adaptarse a estos cambios de la forma más rápida posible. Si a este hecho se le añade el gran volumen de datos que manejan, la ventaja de automatizar el descubrimiento y clasificación de los patrones de conducta (de los clientes) se vuelve casi una obligación.

En el caso que hemos abordado, al disponer de una muestra representativa de los datos reales (menos de 1% del total, compuesta por el conjunto de series temporales pertenecientes a un establecimiento) los resultados obtenidos son más que satisfactorios. Esto es así, ya que el clasificador de series temporales que hemos construido utilizando máquinas de vectores de soporte, genera un listado de clústeres muy reducido. Esto conlleva una serie de ventajas de cara a su aplicación a la totalidad de los datos.

Por un lado, simplifica la tarea de descubrimiento de modelos estadísticos. Y lo hace de tal forma que sólo con 12 modelos se puede representar al 98% de las series implica que la tarea de creación de modelos y su mantenimiento sea abordable con un grupo relativamente pequeño de personas. Aunque desde un punto de vista operativo, sea necesario calcular los coeficientes ARIMA para cada serie.

Por otro lado, como presuponemos que la muestra es suficientemente representativa (contiene todos los SKUs de un establecimiento) cabe suponer que al utilizar el método sobre el conjunto completo de series, el número de categorías no va a ser muy elevado.

De cara a la implementación del método, con miras a la creación de un

sistema autónomo y teniendo en cuenta que el entrenamiento / definición (de los juegos de prueba / validación) de las máquinas de vectores es un proceso automático y que el cálculo de los coeficientes ARIMA también lo es, nos encontramos con un sistema que sólo necesita de la intervención del usuario en tareas muy concretas: análisis de los nuevos clústeres y definición de los modelos ARIMA correspondientes.

Además, se ha conseguido identificar modelos ARIMA para cerca de un 20% de la muestra filtrada. Aunque será necesario realizar pruebas de ajuste, este hecho puede facilitar la implantación del sistema en una empresa real.

Cabe destacar también que al tener cálculos compartidos (la generación de los vectores de autocorrelación y autocorrelación parcial) en las fases de clasificación y construcción de los modelos, las necesidades de hardware también son menores.

5.1. Trabajo Futuro

Como posibles trabajos futuros pueden señalarse los siguientes:

- Ampliar el sistema para que las tareas de decisión (en las cuales se determina si una serie pertenece a una categoría) sean realizadas en paralelo. Debido al gran volumen de series a clasificar, sería una mejora necesaria de cara a su aplicación a un caso real.

- Crear un sistema de identificación automática de modelos ARIMA que reciba como entrada los resultados de las funciones ACF / PACF.

- Realizar una comparación con diferentes clasificadores (como los basados en lógica difusa), tanto en precisión en la clasificación como en rendimiento del sistema.

- Desarrollar y ajustar modelos de previsiones para las categorías que

no tienen uno asignado.

- Dado el porcentaje de series temporales que no admiten modelos ARIMA debido a la venta esporádica o aleatoria, investigar y desarrollar modelos estadísticos apropiados para este tipo de series.

REFERENCIAS

[1] Hau L Lee; V Padmanabhan; Seugjin Whang: "Information distortion in a supply chain: The bullwhip effect". *Management Science; Apr 1997; 43, 4;*

[2] Frank Chen, Zvi Drezner, Jennifer K. Ryan, and David Simchi-Levi: "Quantifying the Bullwhip Effect in a Simple Supply Chain: The Impact of Forecasting, Lead Times, and Information". Management Science 2000 46:436-443.

[3] Richard Metters: "Quantifying the bullwhip effect in Supply-Chains". Journal of Operations Management(15) 1997, 89-100.

[4] Forrester, Jay Wright (1961). "Industrial Dynamics". MIT Press.

[5] Manish Shukla; and Sanjay Jharkharia: "ARIMA models to forecast demand in fresh supply chains" ; International Journal of Operational Research; Volume 11, Number 1 / 2011.

[6] Volkan S. Ediger, Sertac Akar: "ARIMA forecasting of primary energy demand by fuel in Turkey"; Energy Policy, Volume 35, Issue 3, March 2007, Pages 1701-1708, ISSN 0301-4215, DOI: 10.1016/j.enpol.2006.05.009..

[7] Chi-Chen Wang: "A comparison study between fuzzy time series model and ARIMA model for forecasting Taiwan export", Expert Systems with Applications, Volume 38, Issue 8, August 2011, Pages 9296-9304, ISSN 0957-4174, DOI: 10.1016/j.eswa.2011.01.015.

[8] J. Shahrabi, S. S. Mousavi and M. Heydar, 2009. Supply Chain Demand Forecasting; A Comparison of Machine Learning Techniques and Traditional Methods. Journal of Applied Sciences, 9: 521-527.

[9] Orsenigo, C., Vercellis, C.: Time series classification by discrete support vector machines. In: Artifical Intelligence and Data Mining Workshop (2006).

[10] Li Wei , Eamonn Keogh: "Semi-supervised time series classification" Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, August 20-23, 2006, Philadelphia, PA, USA.

[11] Ratanamahatana, C.A. & Keogh, E. Making Time-series Classification More Accurate Using Learned Constraints. In proceedings of SIAM International Conference on Data Mining (SDM '04), Lake Buena Vista, Florida, April 22-24, 2004.

[12] Geurts, P. Pattern extraction for time series classification. In proceedings of Principles of Data Mining and Knowledge Discovery, 5th European Conference (2001). Freiburg, Germany, Sept 3-5. pp 115-127.

[13] Chatfield C., The Holt-Winters Forecasting Procedure, Journal of the Royal Statistical Society. Series C (Applied Statistics), Vol. 27, No. 3 (1978), pp. 264-279.

[14] Chatfield C., & Yar, M. (1988). Holt-Winters forecasting: Some practical issues, The Statistician, 37, 129-140.

[15] Box, George and Jenkins, Gwilym (1970) Time series analysis: Forecasting and control, San Francisco: Holden-Day.

[16] R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

[17] Chih-Chung Chang and Chih-Jen Lin, LIBSVM : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1--27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[18] Schölkopf, B., Sung, K., Burges, C., Girosi, F., Niyogi, P., Poggio, T. and Vapnik. V. Comparing support vector machines with gaussian kernels to radial basis function classifiers. IEEE Trans. Sign. Processing, 45:2758–2765, 1997.

[19] Schölkopf, B., A. Smola, R. Williamson, and P. L. Bartlett. New support vector algorithms. Neural Computation, 12, 2000, 1207-1245.

[20] A Practical Guide to Support Vector Classification, Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin, Department of Computer Science, National Taiwan University, Taipei 106, Taiwan. http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf

[21] B. E. Boser, I. Guyon, and V. Vapnik. A training algorithm for optimal margin classifiers. In Proceedings of the Fifth Annual Workshop on Computational Learning, Theory, pages 144-152. ACM Press, 1992

[22] C. Cortes and V. Vapnik. Support-vector network. Machine Learning, 20:273-297,1995.

[23] Support Vector Machines Definition, Wikipedia. http://en.wikipedia.org/wiki/Support_vector_machine

[24] Martin Stepnicka, Juan Peralta Donate, Paulo Cortez, Lenka Vavríková, German Gutierrez: Forecasting seasonal time series with computational intelligence: contribution of a combination of distinct methods. EUSFLAT, 464-471, volume 1. 2011

[25] CAO, Lijuan: Support vector machines experts for time series forecasting, Neurocomputing, Volume 51, April 2003, Pages 321-339

[26] KIM, Kyoung-jae: Financial time series forecasting using support vector machines, Neurocomputing, Volume 55, Issues 1-2 (September 2003), Pages 307-319

[27] CAO, Lijuan and Francis E. H. TAY: Financial Forecasting Using Support Vector Machines. Neural Computing & Applications, Volume 10, Number 2 (May 2001), Pages 184-192.

[28] Da-yong Zhang, Hong-wei Song, Pu Chen: Stock market forecasting model based on a hybrid ARMA and support vector machines. International Conference on Management Science and Engineering, 2008. ICMSE 2008. 15th Annual Conference Proceedings., pags. 1312 - 1317

ANEXO A: TABLAS DE DATOS

	Venta 2009											
	Enero	**Febrero**	**Marzo**	**Abril**	**Mayo**	**Junio**	**Julio**	**Agosto**	**Septiembre**	**Octubre**	**Noviembre**	**Diciembre**
1	8102	30934	42124	56705	17525	52081	49211	37143	45163	45564	35243	50012
2	47074	44314	51903	66120	26469	45093	49886	26292	41207	48872	47264	49106
3	37376	45543	51222	69342	49707	48391	52726	36652	45464	43267	43631	51485
4	32676	44359	52168	54686	59323	48893	39757	36799	47790	33077	45241	57470
5	41719	48760	55772	38164	52460	52236	30690	34280	34542	40665	50330	41154
6	16553	53892	59819	0	51884	44581	44934	41049	27338	41287	54635	26528
7	49649	49295	53607	58795	55387	33389	43142	43759	44096	44187	40390	40541
8	53643	24680	27198	63362	59619	44802	45134	28907	44195	50020	15438	27089
9	25083	53022	55463	16812	51110	47280	49778	11730	41283	56721	43989	47782
10	44904	47506	51421	12356	22987	52676	54632	37441	47428	40497	44485	47651
11	32728	49585	51761	38169	54314	30896	40815	32260	47766	14806	44971	48920
12	51644	62407	55531	19269	48414	50212	14195	32230	36540	28710	45405	44246
13	45811	0	63005	57054	53936	34276	44002	36733	16823	46426	49128	34423
14	47058	57990	52109	51181	65547	19932	46136	44871	46899	44462	43714	51957
15	47818	28812	24425	55019	39243	48340	48654	15236	40407	44723	21102	45893
16	61081	59301	57912	57665	39207	46413	49944	12501	42681	52382	46736	47348
17	57647	52418	57147	64186	20251	46071	51726	40443	44169	45036	43869	47100
18	26193	54151	62680	58084	50114	52202	38415	31764	49330	20293	45322	48575
19	56970	55043	37477	25518	46010	54588	16511	32139	38657	46084	44812	47931
20	52759	67000	54154	58763	37856	43941	44319	36575	18069	41980	47677	36795
21	54088	57459	40891	55341	30849	18983	42900	39962	45962	44044	46998	41127
22	52214	31017	22841	52447	66585	51972	46831	27631	42653	48994	20437	49887
23	68968	62134	57082	55243	49054	49195	49027	13046	41718	52581	47638	57247
24	58896	55941	51443	61608	20391	48757	53259	38804	46662	43459	43955	39385
25	29782	48847	54029	54900	49865	51588	35849	32854	50538	20363	46070	8811
26	59180	52374	54825	22278	48825	56236	14896	34379	41420	46018	47974	37233
27	57297	59638	62962	61314	43930	42115	48561	37228	17190	44429	53768	29823
28	55766	54726	49503	56978	51366	16592	42140	42548	47027	43279	47402	46674
29	57120		24866	63018	56882	50353	45239	28169	43291	48355	29461	48806
30	64330		57920	75822	47556	49979	53356	22385	46154	55386	52696	54167
31	54432		58925		20848		53790	46754		44458		38468

A-1 Datos ejemplo correspondientes a 2009

	Venta 2010											
	Enero	Febrero	Marzo	Abril	Mayo	Junio	Julio	Agosto	Septiembre	Octubre	Noviembre	Diciembre
1	8626	50006	53378	16217	17087	60583	56518	31754	51450	62620	23951	53289
2	41460	44766	47380	14070	45265	68430	55700	45882	51949	46967	58869	61059
3	30735	46171	48798	31307	51916	38161	39643	41507	57680	41079	56123	67118
4	43526	48312	47859	28312	50818	57431	35771	39244	38409	54915	54244	50959
5	41954	52051	47291	48586	54368	40046	47750	44477	34669	49743	61945	32129
6	15136	43576	45221	49373	55003	36423	53580	47184	49584	50825	48792	15444
7	50420	36608	37065	48839	57127	52724	46598	31789	48041	60036	34319	46411
8	47514	46088	46861	48108	47368	52287	56652	13777	45521	62117	50386	32336
9	44691	46171	46387	56002	19700	49868	60150	40569	50746	43648	46059	53111
10	29531	46358	45177	0	53301	51494	43383	36743	55525	16171	52607	53351
11	31351	49203	48372	0	51302	54880	15330	35687	40738	51366	51596	48625
12	43623	51470	54049	49803	48348	45266	50555	38641	20557	34504	55015	34893
13	42723	42679	46332	50525	53865	19638	54112	46382	55003	51964	49873	52258
14	47676	23014	20790	49073	59553	53484	51109	30253	47908	53142	20589	51991
15	47863	48090	50673	49544	34543	49748	55393	17092	50834	58389	55448	55710
16	44067	41114	48521	58207	19919	48064	57891	42955	53255	50200	50324	54255
17	20336	44154	48520	48463	53498	51180	41951	35856	56386	22941	51204	57804
18	50600	44771	54900	20290	50056	56810	17444	35620	44792	55646	53484	51683
19	44439	50949	32960	50843	49089	46000	49596	38960	19785	51074	57247	43352
20	46793	45057	37323	47349	54629	22127	46131	41230	52656	49981	49522	55081
21	50246	20896	19580	52049	61366	53496	49739	30223	53150	54393	25423	57124
22	47477	47092	51995	55685	45538	54430	56398	8483	50892	59935	56574	58909
23	49140	47109	45623	54937	21978	54311	58342	44269	52484	49576	52677	69266
24	17962	48895	47618	47115	52810	58095	37235	37257	58919	23667	52460	46142
25	52717	46940	52759	21025	48541	55244	15099	40445	43943	54785	52962	10715
26	45870	56186	59513	55996	48628	47125	49650	41217	25368	50703	59177	38647
27	45483	46243	51130	49137	55968	20544	47265	45015	56344	51670	48982	50916
28	49730	21535	19618	51789	63433	57761	50033	35002	60136	60062	39595	0
29	55294		50989	59553	53047	53820	55031	25601	34016	67233	56707	50743
30	45409		52012	64432	24434	61953	60106	46397	54193	53095	54034	58587
31	22573		59135		60139		43442	46127		23173		42318

A-2 Datos ejemplo correspondientes a 2010