Non deterministic Fuzzy Classification Systems

Vincenzo CUTELLO*
Dept. of Computer and Inf. Sci.
Polytechnic University
Brooklyn, New York

Javier MONTERO
Dept. of Statistics and O.R.
Complutense University
Madrid, Spain

Abstract

We discuss upon the problem of learning fuzzy membership functions from structurally given, non deterministic fuzzy classification rules with non fuzzy conclusions.

Keywords: Fuzzy Sets, Aggregation Operators, Classification Systems, Inductive Learning.

1 Introduction

A wide research area in mathematics is devoted to the formalization what they call *Decision Making*, either from a *descriptive* or a *prescriptive* point of view. This is justified because individual decisions appears sometimes as the only *objective* information we get about people, and one can argue that individual decisions are the essential human fact. A human being is viewed just as a *decision maker.*

We should realize that a portion of the Decision Making researchers have a particular view about what a *decision* is. Decision Making in a Bayesian context, for example, is a Decision Making about acts, i.e., about alternatives which have been absolutely well defined.

Is that crisp Decision Making the essential human fact?

We really think that the essential human fact is not so related to decision problems about actions, but to decision problems about, lets say, *ethical principles* and more in general *arguments.* An act is indeed the final observable output of a previous thinking process, hierarchically structured (see [11]), in which at each step we select among ill-defined families of possible alternatives, until we get a particular quite well defined alternative. These are the key *Information Process* processing problems other researchers refer to. As pointed out by Shafer [12], what people really look for are arguments on which their choices should be based. Choices are most of the time a consequence of arguments mixed with basically *random* (uncontrolable, a priori not known) inputs.

Most human beings do not care too much about *optimized* solutions to their problems. They do care about fast and cheap enough solutions even if approximate. What we do always want is a better knowledge of each problem. The key issue for us is how information can be obtained and processed. The good solution, if any, should follow ¿from a good comprehensive analysis of the problem, ¿from a good information processing process.

Acts are crisp while arguments are quite often fuzzy. This is a key issue in Fuzzy Logic and it is the main topic of this paper. Very often in life we are faced with the problem of making a decision based on some imprecise knowledge. Typical cases of this kind occur in the medical field. In an emergency room of a hospital doctors have to decide whether or not to admit a patient in the hospital (see [9]). Their knowledge is what they have learned in Medical School (enriched by the experience gained at work) and the symptoms they observe in the patient, that is to say their knowledge is a set of fuzzy rules which on any given input must produce a Yes or No output.

By introducing ad hoc membership functions, one can use standard fuzzy inference mechanisms to come up with a value of admissibility for the patient. Then based on this value (typically introducing a security threshold parameter) one can decide whether to admit the patient or not. However, since in this case we are dealing with human lives, we need to be reasonably sure that the used membership functions are well tuned and in turn that the inference mechanisms and the system of fuzzy rules with a non-fuzzy conclusion we are working with has been validated (see [2]). How can we do that ?

*On leave from Dept. of Mathematics, University of Catania, Catania, Italy

We will deal with such information processing and decisional processes, by studying the underlying aggregation problems and by formalizing an inductive learning framework.

Many real-life problems are solved by means of some information aggregation procedures to be designed and used in an "intelligent" manner. Information is passed to an aggregation operator as an ordered sequence of real numbers, which without loss of generality can be supposed to belong to the unit interval (see [3, 4, 5, 6, 7] for a particular formalization of this problem and [13, 14, 15, 16] for related work)

Aggregation operators are quite simple maps. Formally, an aggregation operator of dimension n, is any mapping

$$\phi : [0,1]^n \rightarrow [0,1].$$

We can also have hierarchical aggregations, that is to say aggregations of chunks of information which in turn represent aggregated information. The practical consequences of such hierarchical aggregations are quite interesting. If we have aggregation maps of very big dimensions, hierarchical aggregations will allow us to deal with sub-aggregation operators of smaller dimensions, whose computational jobs can be parallelized. Thus, it will be possible to obtain a significant speed up of the whole aggregation process.

For instance, hierarchical aggregation procedures for individual preferences are defined by means of a basic classification of the individuals. The set of individual is divided into groups, in such a way that each individual is present in at least one of these groups. Then partial amalgamations of opinions within each group are to be amalgamated into one global amalgamation.

As pointed out in [8], when we think about an aggregation rule, we really mean a *family* of aggregation rules, the same way in which a *decision* of ours is in fact a *family* of possible actions with a clear conceptual analogy or being different due to some unavoidable measurement error or some very last guessing, for example. At the end, only one aggregation rule and only one act will come out ¿from the families we have in our mind.

2 Non deterministic aggregation operators

Similarly to Turing Machine, a non deterministic aggregation operator is such that the same input can produce different outputs. For Turing Machine non determinism has the intended meaning of computational power: in one step the machine can non deterministically choose any of the possible outcomes. Non deterministic algorithms "guess" the best outcome among all possible ones.

Analogously, a non deterministic aggregation operator can be defined in the following way:

- we have a set M of mappings

$$\phi : [0,1]^n \rightarrow [0,1]$$

- and a non deterministic choice function (Turing Machine) η such that

$$\Phi(x_1,\ldots,x_n) = \eta(\{\phi(x_1,\ldots,x_n) | \phi \in M\})$$

Such aggregation operators will be denoted by $\Phi(M)$. Thus, by definition

$$\Phi(M)(x_1,\ldots,x_n) = \eta(\{\phi(x_1,\ldots,x_n) | \phi \in M\}).$$

For aggregation operators non determinism will have the intended meaning of "intelligence" power. We can think of it as a process that in one step tests all the possible operators of the family M and chooses the best one.

It is now quite natural to introduce the concepts of non-deterministic T-norms and T-conorms and OWA operators.

DEFINITION 1 *A non deterministic T-norm is a non deterministic aggregation operator $\Phi(M)$ such that M is a set of T-norms.*

Analogously, a non deterministic T-conorm is a non deterministic aggregation operator $\Phi(M)$ such that M is a set of T-conorms.

Finally, a non deterministic OWA operator is a non deterministic aggregation operator $\Phi(M)$ such that M is a set of OWA operators.

The above definitions clearly extend the traditional definitions which are obtained for $|M| = 1$. Moreover, we can characterize non deterministic aggregation operators as finite, discrete and continuous according to the cardinality of M.

Example 1 Let M be the Dubois-Prade family of T-norms (see [10]). Therefore, the elements of M are of type

$$T_\alpha^{DB} = \frac{ab}{\max(a,b,\alpha)}$$

for $\alpha \in [0,1]$. In particular, for $\alpha = 0$ we obtain the minimum function and for $\alpha = 1$ the product function.

The associated family of T-conorms is obtained by computing the dual functions with respect to the standard negation operator $n(x) = 1 - x$. Thus

$$S_\alpha^{DB} = 1 - \frac{(1-a)(1-b)}{\max(1-a, 1-b, \alpha)}.$$

3 Non Deterministic Fuzzy Classification Systems

Let $\mathcal{U} \subseteq \Re^k$ for some integer k. Therefore, each element of $\mathcal{U}$ is a k-tuple of real numbers.

Let C be a concept over $\mathcal{U}$. A system S_m of classification rules for C (*Fuzzy Classification System*, FCS for short) is given in the following form:

$$\begin{array}{lll} Q_1^1(p_{Q_1^1}(x)) & \wedge \cdots \wedge Q_{n_1}^1(p_{Q_{n_1}^1}(x)) \rightarrow C(x) \\ \cdots & \cdots & \cdots \\ Q_1^m(p_{Q_1^m}(x)) & \wedge \cdots \wedge Q_{n_m}^m(p_{Q_{n_m}^m}(x)) \rightarrow C(x) \end{array}$$

where

- the predicates Q_i^j are taken from a set

  $$\mathcal{P} = \{Q_1, ..., Q_n\}$$

  of given unary predicates;

- for all Q_j, p_{Q_j} is a projection function that returns the parameter which is of significance for Q_j.

- for every component k, the predicates

  $$Q_1, ..., Q_{m_k}$$

  corresponding to k (i.e. such that p_{Q_i} returns the value of the k-th component) are *convex* and define a linguistic order, that is to say if $j > i$ then for every $0 \leq \alpha \leq 1$ we have

  $$\{x | Q_i(x) > \alpha\} \prec \{x | Q_j(x) > \alpha\},$$

  where $\prec$ is the standard order relation on real intervals:

  $$[a, b] \prec [c, d] \text{ iff } a \leq c \text{ and } b \leq d.$$

The intuitive meaning of the rule system is that a given element x is classified as a positive example for the concept C if one or more of the rule antecedents are *true* for x.

Therefore we can say that $x \in \mathcal{U}$ is a member of the concept C i.e. $C(x)$ is true if and only if

$$\tau\left(\bigvee_{j=1}^{m} \bigwedge_{i=1}^{n_j} Q_i^j(p_{Q_i^j}(x))\right) = 1$$

where τ is the the truth function. Since we are allowing some predicates in $\mathcal{P}$ to be fuzzy whereas C is not we make the following assumptions:

- A threshold parameter $0 < \theta < 1$ is given;
- $C(x)$ is θ-true if and only if

  $$\tau\left(\bigvee_{j=1}^{m} \bigwedge_{i=1}^{n_j} Q_i^j(p_{Q_i^j}(x))\right) > \theta$$

  and where the truth value above is computed as follows

  $$\tau\left(\bigvee_{j=1}^{m} \bigwedge_{i=1}^{n_j} Q_i^j(p_{Q_i^j}(x))\right) = S_{j=1}^{m}(T_{i=1}^{n_j}(Q_i^j(p_{Q_i^j}(x)))).$$

  where S and T are respectively a non deterministic T-conorm and a non deterministic T-norm.

The basic idea is clear: during the inference process, the *best* logical, <u>dual</u> aggregation operators ¿from the classes S and T are non-deterministically chosen and applied.

Notice that for internal consistency, we are imposing the condition on the non deterministic choice of dual T-norms and T-conorms.

Given a set F=$\{\mu_1, ..., \mu_n\}$ of membership functions associated to the predicates in $\mathcal{P}$, the above fuzzy classification system (FCS) is denoted by S_m^F. The notation S_m then denotes the collection of all possible FCS's S_m^F, which in turn can be characterized as the collection of all sets F of membership functions. Such FCS's are called *convex*, non deterministic, fuzzy classification systems.

In case the system is deterministic and the truth values are computed according to the min-max semantic, i.e.

$$\tau\left(\bigvee_{j=1}^{m} \bigwedge_{i=1}^{n_j} Q_i^j(p_{Q_i^j}(x))\right) = \max_{j=1,\dots,m}(\min_{i=1,\dots,n_j}(Q_i^j(p_{Q_i^j}(x))).$$

Then, as proven in [1], the membership functions in the system can be quickly tuned from examples.

We introduce the following definitions

DEFINITION 2 *An element x is* justifiably *classifiable as a θ-positive example for the non deterministic concept $C \equiv S_m^F$ if*

$$\tau\left(\bigvee_{j=1}^{m}\bigwedge_{i=1}^{n_j} Q_i^j(p_{Q_i^j}(x))\right) > \theta$$

for some possible choices of the aggregation operators.

Analogously, x is justifiably *classifiable as a θ-negative example for the non deterministic concept $C \equiv S_m^F$ if*

$$\tau\left(\bigvee_{j=1}^{m}\bigwedge_{i=1}^{n_j} Q_i^j(p_{Q_i^j}(x))\right) \leq \theta$$

for some possible choices of the aggregation operators.

Given a non deterministic concept S_m^F, which is an approximation of the concept C, and given an element x we define $error(S_m^F, x)$ to be true if the obtained classification is wrong, i.e. if x is *justifiably* classifiable θ-positive [resp. θ-negative] for C and is not *justifiably* classifiable θ-positive [resp. θ-negative] for S_m^F.

Given a set E of justifiably θ-positive and justifiably θ-negative examples of C, drawn ¿from an unknown, but fixed distribution D over the universe $\mathcal{U}$, we define

$$\begin{aligned}
\underline{\text{global errors}}\quad Errors &= \{x \in X | error(S_m^F, x)\}\\
\underline{\text{observed error}}\quad Err_0 &= \frac{|Errors \cap E|}{|E|}\\
\underline{\text{expected error}}\quad Err &= \sum_{x \in Errors} D(x)
\end{aligned}$$

Our main goal is to produce a learning algorithm, that is to say an algorithm which by observing examples, justifiably negative or positive, is able to produce membership functions which will minimize Err_0 and Err.

The question to be answered is now the following: can learning take place and if, yes, under what conditions ?

4 A learnability result

Some non deterministic fuzzy classification systems allow the design of inductive learning procedures, by observing examples. The general problem of characterizing such systems remains open. We will show now that if the underlying sets of T-norms and T-conorms are finite and linearly parameterized then we have a positive result.

Suppose then that the evaluation

$$S_{j=1,\ldots,m}(T_{i=1,\ldots,n_j}(Q_i^j(p_{Q_i^j}(x)))$$

is done by choosing S and T to be two finite subsets of respectively the Dubois-Prade classes of T-conorms and T-norms. Thus, S and T can be characterized as follows:

- there exists n values $0 = \alpha_1, \cdots, \alpha_n = 1$ such that (see example 1)

$$\begin{aligned}
S &= \{S_{\alpha_1}^{DB}, S_{\alpha_2}^{DB}, \cdots, S_{\alpha_n}^{DB}\}\\
T &= \{T_{\alpha_1}^{DB}, T_{\alpha_2}^{DB}, \cdots, T_{\alpha_n}^{DB}\}
\end{aligned}$$

- by definition $S_{\alpha_i}^{DB}$ is the dual of $T_{\alpha_i}^{DB}$ for every $1 \leq i \leq n$
- a non deterministic choice corresponds to a choice of a constant α_i.

Our inductive learning procedure is (of course) non deterministic

(Step 1) Draw a sufficient (see [1]) number of examples;

(Step 2) If there is no example classified as positive (i.e. not justifiably negative) or negative (i.e. not justifiably positive) then output **NOTHING**; otherwise apply for these examples and for $\alpha = 0$ the algorithm in [1] and output a first approximation of the membership functions and the threshold value θ;

(Step 3) For each of the remaining examples "guess" the value α_i used to classify it;

(Step 4) For α_i and the corresponding set of examples try to modify the membership functions and the threshold value so to obtain the same classification for the examples; if modifications are possible output the new values, otherwise output **NOTHING**;

(Step 5) If there are n "equal" outputs at Step 4 then output such common values otherwise output **NOTHING**;

A few comments to intuitively explain the learning algorithm the algorithm looks first for certain values, that is to say positive and negative examples. If all examples are both justifiably negative and positive then no learning can take place. This is not surprising. The examples give no information at all.

Using the set of certain examples, the algorithm makes a guess on the values of the membershio functions used in the classification system and on the threshold. It does so using the inductive procedure described in [1] which uses max-min evaluation (i.e. $\alpha = 0$).

For the non certain examples, we guess the values of α that were used and we try the built membership functions on them.

5 Final Comments

In this paper we formalize the notion of non deterministic fuzzy classification system, following the work done in [8] and [1]. Our final goal was to show that inductive learning procedures can be built for the membership functions within a non-deterministic fuzzy classification system with non fuzzy conclusions. This first positive result opens a large set of problems of high interest in the general field of learning ¿from examples membership functions. Such problems are both of a theoretical (formalization and proof of learnability for specific fuzzy classification systems) and experimental (we need a system that does so and it uses the right heuristics to bypass computational hardness) nature.

Acknowledgment: This research has been partially supported by Dirección General de Investigación Científica y Técnica of Spain (PB95-0407).

References

[1] F. Bergadano, V. Cutello. Probably Approximate Correct (PAC) learning in fuzzy classification systems. *IEEE Transactions on Fuzzy Systems*, Vol.3 N. 4: 473-478, 1995.

[2] A.M. Chang and L.O. Hall. The validation of fuzzy knowledge-based systems. In *Fuzzy logic for the management of uncertainty*, 589-604. L.A. Zadeh and J. Kacprzyk, editors. John Wiley and Sons, Inc., 1992

[3] V. Cutello and J. Montero. A characterization of rational amalgamation operations. *International Journal of Approximate Reasoning* 8:325–344 (1993).

[4] V. Cutello and J. Montero. Hierarchies of intensity preference aggregations. *International Journal of Approximate Reasoning*, 10:123-133, 1994.

[5] V. Cutello and J. Montero. Hierarchies of aggregation operators. *International Journal of Intelligent Systems* , 9:1025-1045, 1994.

[6] V. Cutello and J. Montero. Hierarchical aggregation of OWA operators: basic measures and related computational problems. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 3:17–26 (1995).

[7] V. Cutello and J. Montero. Information and aggregation: ethical and computational issues. In: D. Ruan, Ed., *Fuzzy Sets Theory and Advanced Mathematical Applications*. Kluwer, pg. 175-200, Boston (1995)

[8] V. Cutello and J. Montero. The problem of defining non deterministic aggregation operators. In: D. Ruan, Ed., *Intelligent Systems and Soft Computing for Nuclear Science and Industry*. World Scientific, 1996, pg.105-112.

[9] D.L. Hudson and M.E. Cohen. The role of approximate reasoning in a medical expert system. In *Fuzzy Expert Systems*, 165-179. A. Kandel, editor. CRC Press, 1991

[10] G.J. Klir and B. Yuan. *Fuzzy sets and Fuzzy Logic. Theory and Applications*. Prentice Hall, Englewood Cliffs, NJ (1995).

[11] T.L. Saaty. Exploring the interface between hierarchies, multiple objectives and fuzzy sets. *Fuzzy Sets and Systems*, 1:57–68 (1978)

[12] G. Shafer. Savage revisited (with discussion). *Statistical Science*, 1:463–501 (1986)

[13] R.R. Yager. On ordered weighted averaging aggregation operators in multi-criteria decision making. *IEEE Transactions on Systems, Man and Cybernetics*, 18:183–190 (1988)

[14] R.R. Yager. Families of OWA operators. *Fuzzy Sets and Systems*, 59:125–148 (1993)

[15] R.R. Yager. MAM and MOM operators for aggregation. *Information Sciences*, 69:259–273 (1993)

[16] R.R. Yager. Aggregation operators and fuzzy systems modeling. *Fuzzy Sets and Systems*, 67:129–145 (1994)