Fuzzy Integral for Classification and Feature Extraction

Michel GRABISCH
Thomson-CSF, Corporate Research Laboratory
Domaine de Corbeville
91404 Orsay Cedex, France
email grabisch@thomson-lcr.fr

Abstract

We describe in this paper the use of fuzzy integral in problems of supervised classification. The approach, which can be viewed as a information fusion model, is embedded into the framework of fuzzy pattern matching. Results on various data set are given, with comparisons. Lastly, the problem of feature extraction is addressed.

1 Introduction

Many methods in pattern recognition have been proposed, based on various approaches, such as Bayesian inference, neural networks, linear methods, nearest prototypes, etc. (see [9] for a large survey of classical approaches). More recently, fuzzy set theory and related domains have brought new tools for pattern recognition, either based on the concept of fuzzy sets [37] as a fundamental concept for modelling classes, or based on non classical representations of uncertainty, such as possibility theory [6], and Dempster-Shafer theory [31] (see the collection of papers edited by Bezdek and Pal [3] to have a large survey of such approaches).

Essentially, these new approaches can often be viewed as a generalization of some classical method, these generalization introducing either some fuzziness in the modelling or some uncertainty measure more general than probability measures. In other cases, they follow the same kind of philosophy, but with different tools. Typical examples in the first category are the fuzzy k-nearest neighbours of Keller *et al.* [23], the fuzzy c-means of Bezdek [2], and all the literature about fuzzy grammars [29]. In the second one, the *fuzzy pattern matching* approach, as described by Dubois *et al.* [8] in the framework of possibility theory, can be viewed as the possibilistic counterpart of the Bayesian approach to classification (see in this respect the paper of Grabisch *et al.* [15] doing a close comparison of the two frameworks).

The fuzzy pattern matching approach, which will be described later on, essentially relies on a modelling of the classes with fuzzy sets, —expressing *typicality of values* and vagueness of the class—, and of the observations by a possibility distribution —expressing the imprecision. In a multiattribute classification problem, for a given class, a fuzzy set is defined for each attribute. The classification is done by finding the best *match* between the observation and the fuzzy classes. The matching is done for each attribute, then the matching degrees are aggregated, according to the way the class is built with respect to the attributes (i.e. we can imagine a disjunctive or a conjunctive construction).

The Bayesian approach proceeds similarly, since classes are described by probability densities for each attribute, expressing the *relative frequency* of the different values. Observations are also modelled by a probability distribution, and what corresponds to the matching operation is done by a convolution of the two probability densities, leading to what is called the *a posteriori probability of classification*. Attributes are often considered as statistically independent, so that the product over all attributes of the a posteriori probabilities are done. This last step corresponds to the aggregation step in fuzzy pattern matching.

The method of classification based on fuzzy integrals we introduce here can be embedded into the approach of fuzzy pattern matching, as it will be explained below. As one can expect, the fuzzy integral is used in the aggregation step, and provides some refinements over usual simpler methods. As the fuzzy measure underlying the fuzzy integral is defined in the set of attributes, it gives precious information on the importance and relevance of attributes to discriminate classes, via the Shapley and the interaction indices (see the companion papers in this book, on k-additive measures and multicriteria decision making). For this reason, this approach enables *feature selection*, a topic of interest in pattern recognition and classification.

Before entering the main subject, some historical considerations are in order. It seems that the first papers dealing with this approach are due to Keller and Qiu [24, 30], in the field of image processing. Later, Tahani and Keller [35] published a paper on classification by fuzzy integral, using the term *information fusion* —which is in fact a different view of this approach, the one taken in [16]. In both cases, the Sugeno integral was used with respect to a λ-measure [34], a particular case of fuzzy measures. Approximately at the same time, Grabisch and Sugeno proposed to use fuzzy t-conorm integrals for classification [20], and later published a complete method based on Choquet integral, including the learning of the fuzzy measure [21]. Since that time, many publications have been done, mostly by the above mentionned authors (see the paper of Keller and Gader in this book for a thorough survey of their works in this field), but also by some others. Let us cite Arbuckle *et al.* [1], Miyajima and Ralescu [27] for face recognition, Cho and Kim [4] for fusion of neural networks, and recently Mikenina and

Zimmermann [26].

We present in a first part basic concepts of fuzzy measures, and describe the method in subsequent parts. More details can be found in previous publications of the author, mainly [12, 16, 17]. See also the paper of Keller and Gader in this book, presenting other applications of this methodology.

In the sequel, $\wedge, \vee$ denote min and max respectively.

2 Background on fuzzy measures and integrals

Let X be a finite index set $X = \{1, \ldots, n\}$.

Definition 1 *A* fuzzy measure μ *defined on* X *is a set function* $\mu : \mathcal{P}(X) \longrightarrow [0, 1]$ *satisfying the following axioms:*

(i) $\mu(\emptyset) = 0, \mu(X) = 1$.

(ii) $A \subseteq B \Rightarrow \mu(A) \leq \mu(B)$

$\mathcal{P}(X)$ *indicates the power set of* X*, i.e. the set of all subsets of* X.

A fuzzy measure on X needs 2^n coefficients to be defined, which are the values of μ for all the different subsets of X.

Fuzzy integrals are integrals of a real function with respect to a fuzzy measure, by analogy with Lebesgue integral which is defined with respect to an ordinary (i.e. additive) measure. There are several definitions of fuzzy integrals, among which the most representative are those of Sugeno [33] and Choquet [5].

Definition 2 *Let* μ *be a fuzzy measure on* X*. The* discrete Choquet integral *of a function* $f : X \longrightarrow \mathbb{R}^+$ *with respect to* μ *is defined by*

$$\mathcal{C}_\mu(f(x_1), \ldots, f(x_n)) := \sum_{i=1}^{n} (f(x_{(i)}) - f(x_{(i-1)}))\mu(A_{(i)}) \tag{1}$$

where $\cdot_{(i)}$ *indicates that the indices have been permuted so that* $0 \leq f(x_{(1)}) \leq \cdots \leq f(x_{(n)}) \leq 1$*. Also* $A_{(i)} := \{x_{(i)}, \ldots, x_{(n)}\}$*, and* $f(x_{(0)}) = 0$.

The discrete Sugeno integral *of a function* $f : X \longrightarrow [0, 1]$ *with respect to* μ *is defined by*

$$\mathcal{S}_\mu(f(x_1), \ldots, f(x_n)) := \bigvee_{i=1}^{n} (f(x_{(i)}) \wedge \mu(A_{(i)})), \tag{2}$$

with the same notations.

The Choquet integral coincides with the Lebesgue integral when the measure is additive, but this is not the case for the Sugeno integral.

3 Classification by fuzzy integral

3.1 General methodology

Let $C_1, \ldots, C_m$ be a set of given classes, and patterns be described by a n-dimensional vector $X^T = [x_1 \cdots x_n]$. We have n sensors[1] (or sources), one for each feature (attribute), which provide for an unknown sample X° a degree of confidence in the statement "X° belongs to class C_j", for all C_j. We denote by $\phi_i^j(X^\circ)$ the confidence degree delivered by source i (i.e. feature i) of X° belonging to C_j.

The second step is to combine all the partial confidence degrees in a consensus-like manner, by a fuzzy integral. It can be shown that fuzzy integrals constitute a vast family of aggregation operators including many widely used operators (minimum, maximum, order statistic, weighted sum, ordered weighted sum, etc.) suitable for this kind of aggregation [18]. In particular, fuzzy integrals are able to model some kind of interaction between features: this is the main motivation of the methodology (more on this in section 6.1). Thus the global confidence degree in the statement "X° belongs to C_j" is given by:

$$\Phi_{\mu^j}(C_j; X^\circ) := \mathcal{C}_{\mu^j}(\phi_1^j, \ldots, \phi_n^j) \tag{3}$$

(or similarly with the Sugeno integral). Finally, X° is put into the class of highest confidence degree. Here, the fuzzy measures μ^j (one per class) are defined on the set of attributes (or sensors), and express the importance of the sensors and groups of sensors for the classification. For example, $\mu^j(\{x_1\})$ expresses the relative importance of attribute 1 for distinguishing class j from the others, while $\mu^j(\{x_1, x_2\})$ expresses the relative importance of attributes 1 *and* 2 *taken together* for the same task. A precise interpretation of this will be given in section 6.1.

The above presentation is done in an information fusion fashion. It is very general and allows many methods to be used. However, it is interesting to embed this methodology in the *fuzzy pattern matching* methodology, a fundamental approach for classification in possibility theory, which is the counterpart of the Bayesian approach in probability theory, as explained in the introduction. Due to its importance, we devote the next paragraph to the presentation of this methodology, its connection with fuzzy integral, and the Bayesian approach.

[1] In what follows, we have taken the point of view of multi-attribute classification, i.e. one attribute per sensor. Nothing prevents us to restate the problem in a multi-sensor or multi-classifier framework, where each source (which could be a classifier) deals with several attributes, possibly overlapping. This is the position adopted in e.g. Cho and Kim [4].

3.2 The fuzzy pattern matching approach

We assume some familiarity of the reader with possibility theory (see [6] for this topic, and [8] for fuzzy pattern matching). Let us denote by U_i the universe of attribute x_i. Each class C_j is modelled by a collection of fuzzy sets $C_j^1, \ldots, C_j^n$ defined on $U_1, \ldots, U_n$ respectively, expressing the set of typical values taken by the attribute for the considered class. An observed datum x is modelled by a possibility distribution $\pi_x(u_1, \ldots, u_n)$, representing the distribution of possible locations of the (unknown) true value of x in $\times_{i=1}^n U_i$. If attributes are considered to be non-interactive, then $\pi_x(u_1, \ldots, u_n) = \wedge_{i=1}^n \pi_i(u_i)$. Now the possibility and necessity degrees that datum x matches class C_j w.r.t attribute i is given by

$$\begin{aligned} \Pi_{\pi_i}(C_j^i) &:= \sup_{u_i \in U_i} (C_j^i(u_i) \wedge \pi_i(u_i)) \\ \mathrm{N}_{\pi_i}(C_j^i) &:= \inf_{u_i \in U_i} (C_j^i(u_i) \vee (1 - \pi_i(u_i))). \end{aligned}$$

The first quantity represents the degree of overlapping between typical values of the class and possible value of the datum, while the second one is an inclusion degree of the set of possible values of x_i into C_j^i. If x is a precise datum, π_x reduces to a point, and the two above quantities collapse into $C_j^i(x_i)$, which corresponds to $\phi_i^j(X^\circ)$.

The next step is the aggregation of these matching degrees, according to the way the class C_j is built. If for example the class is built by the conjunction of the attributes, i.e.

$x \in C_j$ `if (`$x_1 \in C_j^1$`) and (`$x_2 \in C_j^2$`) and` $\cdots$ `and (`$x_n \in C_j^n$`)`

then it can be shown that, letting $C_j := C_j^1 \times \cdots \times C_j^n$,

$$\begin{aligned} \Pi_\pi(C_j) &= \bigwedge_{i=1}^n \Pi_{\pi_i}(C_j^i) \\ \mathrm{N}_\pi(C_j) &= \bigwedge_{i=1}^n \mathrm{N}_{\pi_i}(C_j^i). \end{aligned}$$

Similarly, if the class is built by a disjunction of the attributes, or a weighted conjunction, a weighted disjunction, the above result still holds, replacing the minimum by a maximum, a weighted minimum or a weighted maximum respectively. More generally, if we consider that C_j is built by a Sugeno integral w.r.t. a given fuzzy measure μ, a construction which encompasses all previous cases, $\Pi_\pi(C_j)$ and $\mathrm{N}_\pi(C_j)$ are also obtained by the (same) Sugeno integral. More specifically:

Proposition 1 *Let μ be a fuzzy measure, and consider that class C is expressed by a Sugeno integral, i.e. $C(u_1, \ldots, u_n) = \bigvee_{i=1}^n \left[C^{(i)}(u_{(i)}) \wedge \mu(A_{(i)})\right]$. Then, the*

possibility and necessity degrees that a datum x belongs to class C is given by

$$\begin{aligned}\Pi_\pi(C) &= \mathcal{S}_\mu(\Pi_{\pi_1}(C^1),\ldots,\Pi_{\pi_n}(C^n))\\ \mathrm{N}_\pi(C) &= \mathcal{S}_\mu(\mathrm{N}_{\pi_1}(C^1),\ldots,\mathrm{N}_{\pi_n}(C^n))\end{aligned}$$

Proof: (only for $\Pi_\pi(C)$, the case of $\mathrm{N}_\pi(C)$ is similar) Applying definitions and elementary calculus, we have:

$$\begin{aligned}\Pi_\pi(C) &= \sup_{u_1,\ldots,u_n}\left[C(u_1,\ldots,u_n)\wedge\bigwedge_{i=1}^n(\pi_i(u_i))\right]\\ &= \sup_{u_1,\ldots,u_n}\left[\bigvee_{i=1}^n[C^{(i)}(u_{(i)})\wedge\mu(A_{(i)})]\wedge\bigwedge_{i=1}^n(\pi_i(u_i))\right]\\ &= \sup_{u_1,\ldots,u_n}\left[\bigwedge_{i=1}^n(\pi_i(u_i))\wedge(C^{(1)}(u_{(1)})\wedge\mu(A_{(1)}))\right]\vee\\ &\quad\sup_{u_1,\ldots,u_n}\left[\bigwedge_{i=1}^n(\pi_i(u_i))\wedge(C^{(2)}(u_{(2)})\wedge\mu(A_{(2)}))\right]\vee\\ &\quad\ldots\\ &= \bigvee_{i=1}^n\sup_{u_1,\ldots,u_n}\left[\left(\bigwedge_{j\neq i}(\pi_j(u_j))\right)\wedge(\pi_i(u_i)\wedge C^{(j)}(u_{(j)})\wedge\mu(A_{(j)}))\right]\\ &= \bigvee_{i=1}^n\sup_{u_{(i)}}[\pi_{(i)}(u_{(i)})\wedge C^{(i)}(u_{(i)})\wedge\mu(A_{(i)})]\\ &= \bigvee_{i=1}^n\left[\sup_{u_{(i)}}[\pi_{(i)}(u_{(i)})\wedge C^{(i)}(u_{(i)})]\wedge\mu(A_{(i)})\right]\\ &= \bigvee_{i=1}^n(\Pi_{\pi_{(i)}}(C^{(i)})\wedge\mu(A_{(i)}))\\ &= \mathcal{S}_\mu(\Pi_{\pi_1}(C^1),\ldots,\Pi_{\pi_n}(C^n)).\end{aligned}$$

The fourth inequality comes from the fact that $\sup_{u_j}\pi_j(u_j)=1$ for every $j=1,\ldots,n$. □

This method can be viewed also under the Bayesian point of view. Let $p(x|C_j)$, $j=1,\ldots,m$ be the probability densities of classes, and $p(x_i|C_j)$, $i=1,\ldots,n$, $j=1,\ldots,m$, the marginal densities of each attribute. The Bayesian inference approach is to minimize the risk (or some error cost function), which amounts, in the case of standard costs, to assign x to the class maximizing the following discriminating function:

$$\Phi(C_j|x)=p(x|C_j)P(C_j)$$

where $P(C_j)$ is the a priori probability of class C_j. If the attributes are sta-

tistically independent, the above formula becomes :

$$\Phi(C_j|x) = \prod_{i=1}^{n} p(x_i|C_j)P(C_j) \tag{4}$$

If the classes have equal a priori probability, formulae (3) and (4) are similar: in probability theory and in the case of independence, the product operator takes place of the aggregation operator.

4 Learning of fuzzy measures

We give now some insights on the identification of the fusion operator, that is, the fuzzy integral, using training data. We suppose that the ϕ_i^j have already been obtained by some parametric or non parametric classical probability density estimation method, after suitable normalization (see Dubois *et al.* [7] for a study on the transformations between probability and possibility): possibilistic histograms, Parzen windows, Gaussian densities, etc.

The identification of the fusion operator reduces to the identification (or learning) of the fuzzy measures μ^j, that is, $m(2^n - 2)$ coefficients. We focus on the case of Choquet integral, since its derivability allows the application of standar optimization techniques. Several approaches have been tried here, corresponding to different criteria. We restrict to the most interesting, and state them in the two classes case ($m = 2$) for the sake of simplicity. We suppose to have $l = l_1 + l_2$ training samples labelled $X_1^j, X_2^j, \ldots, X_{l_j}^j$ for class C_j, $j = 1, 2$. The criteria are the following.

- the squared error (or quadratic) criterion, i.e. minimize the quadratic error between expected output and actual output of the classifier. This takes the following form.

$$\begin{aligned} J &= \sum_{k=1}^{l_1} (\Phi_{\mu^1}(C_1; X_k^1) - \Phi_{\mu^2}(C_2; X_k^1) - 1)^2 \\ &\quad + \sum_{k=1}^{l_2} (\Phi_{\mu^2}(C_2; X_k^2) - \Phi_{\mu^1}(C_1; X_k^2) - 1)^2. \end{aligned}$$

  It can be shown that this reduces to a quadratic program with $2(2^n - 2)$ variables and $2n(2^{n-1} - 1)$ constraints (coming from the monotonicity of the fuzzy measure), which can be written:

  minimize $\frac{1}{2}\mathbf{u}^T\mathbf{D}\mathbf{u} + \mathbf{\Gamma}^T\mathbf{u}$
  under the constraint $\mathbf{A}\mathbf{u} + \mathbf{b} \geq \mathbf{0}$

  where $\mathbf{u}$ is a $2(2^n - 2)$ dimensional vector containing all the coefficients of the fuzzy measures μ^1, μ^2, i.e. $\mathbf{u} := [\mathbf{u}_1^T \ \mathbf{u}_2^T]^T$, with

  $\mathbf{u}_j := [\mu^j(\{x_1\})\mu^j(\{x_2\}) \cdots \mu^j(\{x_n\})\mu^j(\{x_1, x_2\}) \cdots \mu^j(\{x_2, x_3, \cdots, x_n\})]^T$

Note that $\mu^j(\emptyset) = 0,\ \mu^j(X) = 1$, so that there is no need to include them into the vector $\mathbf{u}$. (see full details in [16, 17]).

- the generalized quadratic criterion, which is obtained by replacing the term $\Phi_{\mu^1} - \Phi_{\mu^2}$ by $\Psi[\Phi_{\mu^1} - \Phi_{\mu^2}]$ in the above, with Ψ being any increasing function from $[-1, 1]$ to $[-1, 1]$. Ψ is typically a sigmoid type function:

$$\Psi(t) = (1 - e^{-Kt})/(1 + e^{-Kt}),\ K > 0.$$

With suitable values of K, differences between good and bad classifications are enhanced. Also, remark that the slope of $\Psi(t)$ at $t = 0$ is $K/2$. This means that with $K = 2$, we have more or less a criterion similar to the squared error criterion. On the other hand, when $K \to \infty$, we tend to the hard limiter, and then the criterion reduces to:

$$J_\infty = 4 l_{\text{miscl}} \tag{5}$$

where l_{miscl} is the number of misclassified samples. Thus, we tend to minimize the number of misclassified samples, as it is the case for the perceptron algorithm [9].

This is no longer a quadratic program, but a constrained least mean squares problem, which can also be solved with standard optimization algorithms when the Choquet integral is used. In fact, this optimization problem requires huge memory and CPU time to be solved, and happens to be rather ill-conditioned since the matrix of constraints is sparse. For these reasons, the author has proposed a heuristic algorithms better adapted to the peculiar structure of the problem and less greedy [10]. The algorithm, called hereafter heuristic least mean square (HLMS), although suboptimal, reduces greatly the computing time and the memory load without a sensible loss in performance.

5 Performance on real data

We give some experimental results of classification performed on real and simulated data. We have tested the Choquet integral with the quadratic criterion minimized with the Lemke method (QUAD), the generalized quadratic criterion minimized by a constrained least squared algorithm (CLMS), and by our algorithm (HLMS), and compared with classical methods. Table 1 (top) give the results obtained on the iris data of Fisher (3 classes, 4 attributes, 150 data), and on the cancer data (2 classes, 9 attributes, 286 data), which is a highly non-Gaussian data set. The results by classical methods come from a paper of Weiss and Kapouleas [36]. The good performance of HLMS on the difficult cancer data is to be noted. The bottom part of the table gives another series of results, obtained on simulated data (3 classes, 4 non-Gaussian attributes, 9000 data, one attribute is the sum of two others)

Method	iris (%)	cancer (%)
linear	98.0	70.6
quadratic	97.3	65.6
nearest neighbor	96.0	65.3
Bayes independent	93.3	71.8
Bayes quadratic	84.0	65.6
neural net	96.7	71.5
PVM rule	96.0	77.1
QUAD	96.7	68.5
CLMS	96.0	72.9
HLMS	95.3	77.4

Method	Classification rate (%)
Bayes linear	82.6
linear pseudo-inverse	84.4
cluster	86.9
adaptive nearest neighbour	87.8
Bayes quadratic	90.3
k nearest neighbour	90.4
tree	96.2
CLMS	90.7
HLMS	89.2

Table 1: Classification rate on various data set

used inside Thomson-CSF for testing purpose. These results show that if the Choquet integral-based classifier is not always the best one, it is nevertheless always among the best ones.

In [25], an experiment has been conducted on a problem of bank customer segmentation (classification). In a first step, we have performed a classification on a file of 3068 customers, described by 12 qualitative attributes, and shared among 7 classes. Classical methods in this context are linear regression, sequential scores, and polytomic scores. The problem happened to be very difficult, since no method (including fuzzy integrals), was able to go beyond 50% of correct classification (see table 2, top[2]). However, in many cases, the quantities $\Phi_{\mu^j}(C_j; x)$ were very near for two classes, showing that the decision of the classifier was not clear cut. In a second step, we have taken into account the "second choice", considering that the classification was also correct when the second choice gaves the correct class, provided the gap between the two greatest $\Phi_{\mu^j}(C_j; x)$ was below some threshold (here 0.05). Performing this way, the classification rate climbed to 65%.

[2]Table 2 gives the classification rate on the test population. 80% of the whole population has been taken for learning, and the remaining 20% for testing.

We have tried to apply the same approach to classical methods, but without good results, since there were very few cases where first and second choices were very near. Even taking systematically the two first choices, the rate obtained was at best 54%. A second experiment was performed on a second file of 3123 customers, described by 8 qualitative attributes, and shared among 7 classes. The results corroborate the fact that the classifier based on the fuzzy integral, when allowing the second choice in case of doubt, largely outperforms the other methods.

File 1

Methods	Classification rate
Regression	45 %
Sequential scores	46.4 %
Fuzzy integral (HLMS)	47.1 %
Polytomic scores	50 %
Polytomic scores (2nd choice)	54%
Fuzzy integral (HLMS) (2nd choice)	65 %

File 2

Methods	Classification rate
Regression	29.9 %
Sequential scores	27.9 %
Fuzzy integral (HLMS)	31.1 %
Polytomic scores	32.2 %
Polytomic scores (2nd choice)	36%
Fuzzy integral (HLMS) (2nd choice)	50 %

Table 2: Segmentation of customers

6 Importance and interaction of attributes

In this section, we denote by $X = \{1, \ldots, n\}$ the set of attributes (or features), and by $x_1, \ldots, x_n$ the corresponding axes.

6.1 General discussion

We address in this section the problem of defining the importance of attributes, and their interaction, a problem closely related to the selection of the best attributes (features) for a given classification problem. The approach we employ here comes directly from our work in multicriteria decision making (see e.g. [11] or the paper by Grabisch and Roubens in this book), and is also closely related to cooperative game theory.

As we explained in section 3.1, a fuzzy measure μ^j is defined for each class C_j. The meaning of $\mu^j(A)$, for any group or *coalition* $A \subset X$, is the following:

> *$\mu^j(A)$ represents the discriminating power of coalition A for recognizing class C_j among the others.*

However, this information is too complex to be understood in the whole, — especially if we want to select the best features—, and we need a more comprehensive representation. Let us consider a particular class C, and drop superindex j. The first question we can ask is:

> *What is the contribution of a single attribute i in the recognition of class C ?*

Obviously, $\mu(\{i\})$ does not bring us the information, since $\mu(\{i\})$ may be very small, but nevertheless, all coalitions containing i may have a high value, which would mean that i is important for classification. Thus, naturally we are lead to:

- a feature i is important if whenever i is added to a coalition of attributes K, the importance of $K \cup \{i\}$, expressed by $\mu(K \cup \{i\})$, is much bigger than $\mu(K)$. The key quantities are

$$\Delta_i(K) = \mu(K \cup \{i\}) - \mu(K), \forall K \subset X \setminus \{i\}.$$

In the field of cooperative game theory, Shapley has provided a definition of importance of a single feature, using an axiomatic approach [32].

Definition 3 *Let μ be a fuzzy measure on X. The* importance index *or* Shapley index *of element i with respect of μ is defined by:*

$$v_i = \sum_{K \subset X \setminus i} \frac{(n - |K| - 1)!|K|!}{n!} \Delta_i(K), \quad \forall i \in X, \tag{6}$$

with $|K|$ indicating the cardinal of K, and 0!=1 as usual. The Shapley value *of μ is the vector $v = [v_1 \cdots v_n]$.*

The Shapley value has the property to be linear with respect to μ, and to satisfy $\sum_{i=1}^{n} v_i = 1$, showing that v_i represents a true sharing of the total importance of X. It is convenient to scale these indices by a factor n, so that an importance index greater than 1 indicates a feature more important than the average.

Although this index gives precious information on the real importance of each attribute, one may ask on what happens if we put together two attributes. We can consider the three following (qualitative) cases:

- *redundancy* or *negative synergy*: the discriminating power of the pair of attributes i, j is not greater than the sum of individual powers. In other words, we do not improve significantly the performance of recognition of a given class by combining attributes i and j, compared to i or j alone.
- *complementarity* or *positive synergy*: the discriminating power of the pair of attributes i, j is greater than the sum of individual powers. In other words, we do improve significantly the performance of recognition of a given class by combining attributes i and j, compared to the importance of i and j alone.
- *independency*: intermediate case, where each attribute brings its contribution to the recognition rate.

Reasoning similarly as for the case of importance of single attributes, we are lead to:

- two features i, j have a positive (resp. negative) synergy if when they are added both to a coalition of attributes K, there is (resp. there is no) significant difference with adding only one of them. Here the key quantities are

  $$\Delta_{ij}(K) = \mu(K \cup \{i,j\}) - \mu(K \cup \{i\}) - \mu(K \cup \{j\}) + \mu(K), \forall K \subset X \setminus \{i,j\},$$

  whose sign will be positive (resp. negative) in case of positive (resp. negative) synergy. This can be easily seen as follows [26].

  $$\Delta_{ij}(K) > 0 \iff$$
  $$\mu(K \cup \{i,j\}) - \mu(K) > \mu(K \cup \{i\}) - \mu(K) + \mu(K \cup \{j\}) - \mu(K).$$

A so-called interaction index can be derived, in a way similar to the Shapley value. Murofushi and Soneda [28] proposed the following index, based on considerations from multiattribute utility theory [22].

Definition 4 *The* interaction index *between two elements i and j with respect to a fuzzy measure μ is defined by:*

$$I_{ij} = \sum_{K \subset X \setminus \{i,j\}} \frac{(n - |K| - 2)!|K|!}{(n-1)!} \Delta_{ij}(K), \quad \forall i, j \in X. \tag{7}$$

It is easy to show that the maximum value of I_{ij} is 1, reached by the fuzzy measure μ defined by $\mu(K \cup \{i,j\}) = 1$ for every $K \subset X \setminus \{i,j\}$, and 0 otherwise. Similarly, the minimum value of I_{ij} is -1, reached by μ defined by $\mu(K \cup \{i\}) = \mu(K \cup \{j\}) = \mu(K \cup \{i,j\}) = 1$ for any $K \subset X \setminus \{i,j\}$ and 0 otherwise.

In fact, Grabisch has shown that the Shapley value and the interaction index can be both embedded into a general interaction index $I(A)$, defined

for any coalition $A \subset X$ [14], which can be axiomatized, as the Shapley value [19]. A positive (resp. negative) value of the index corresponds to a positive (resp. negative) synergy.

The Shapley and interaction indices bring precious help to select relevant features in a classification problem. For example, an attribute which has always negative interaction with other attributes can be removed, while an attribute with high Shapley index and which has a positive interaction with some attributes is essential in the classification process. Mikenina and Zimmermann have proposed an algorithm of selection based on these ideas [26].

6.2 Importance and interaction of attributes for the iris data set

Let us apply these indexes to the iris data set. Figures 1 and 2 give the histograms of every feature for every class, as well as projections of the data set on some pairs of features.

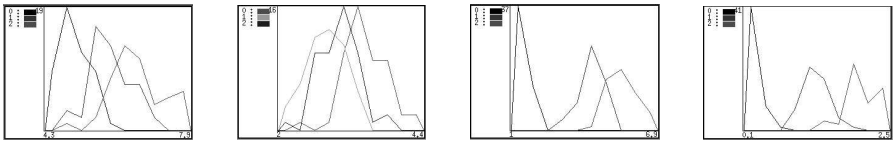

Figure 1: Histograms of the iris data (from left to right: features 1 to 4)

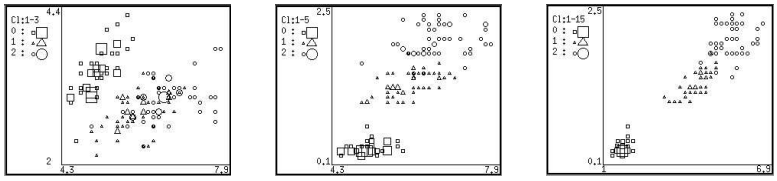

Figure 2: Projections of the iris data (from left to right: on features 1 and 2, 1 and 4, 3 and 4 resp.)

In these figures, samples of class 1 (resp 2, 3) are represented by squares (resp. triangles, circles).

Tables 3 give importance index and interaction indexes computed from the result of learning by HLMS (classification rate is 95.3%). We can see that the Shapley value reflects the importance of features which can be assessed by examining the histograms and projection figures. Clearly, x_1 and x_2 are not able to discriminate the classes, especially for classes 2 and 3. In contrast, x_3 and x_4 taken together are almost sufficient.

index of importance $v_i \times 4$			
feature	class 1	class 2	class 3
1	0.759	0.670	0.416
2	0.875	0.804	0.368
3	1.190	1.481	1.377
4	1.176	1.045	1.839

index of interaction I_{ij}			
features	class 1	class 2	class 3
1,2	0.128	-0.159	-0.065
1,3	0.051	0.281	0.052
1,4	0.054	-0.257	0.010
2,3	-0.009	0.114	0.002
2,4	-0.007	0.036	0.059
3,4	-0.051	0.132	-0.238

Table 3: Indexes of importance and interaction for the iris data set

The interaction indexes are not always so easy to interpret. However, remark that x_1 and x_2 are complementary for class 1: the projection figure on these two axes shows effectively that they are almost sufficient to distinguish class 1 from the others, although x_1 or x_2 alone were not. In contrast, these two features taken together are not more useful than x_1 or x_2 for classes 2 and 3 (redundancy). The fact that I_{14} for class 2 is strongly negative can be explained as follows. Looking at the projection figure on x_1, x_4, we can see that x_1 (horizontal axis) brings no better information than x_4 to discriminate class 2 from the others, so that the combination $\{x_1, x_4\}$ is redundant. Concerning x_3 and x_4, the examination of the projection figure shows that they are rather complementary for classes 2 and 3. Although I_{34} is positive for class 2 as expected, it is strongly negative for class 3.

Finally, we perform a Principal Component Analysis on the iris data set before training the classifier. As expected, the Shapley values for x_1 are very high, and the interaction index between x_1 and x_2 shows a strong complementarity (see figure 3 and corresponding table, where values concerning attributes 3 and 4 have been omitted).

6.3 Determination of the fuzzy measures by the interaction index

The previous results have shown the validity of the above defined concepts, and of the learning algorithm for fuzzy measures, despite some abnormalities in the values of I_{ij}. A new question arises now: *can we do the converse?*

	class 1	class 2	class 3
$4v_1$	2.129	1.873	2.198
$4v_2$	0.583	0.421	0.298
I_{12}	0.035	0.167	0.124

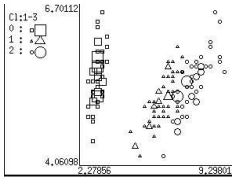

Figure 3: The iris data set after PCA, and the indexes of importance and interaction

Since by examining the data set by histograms, projections, or other means, we can have a relatively precise idea of the importance and interactions of features, are we able to find a fuzzy measure having precisely these values of importance and interaction indexes?

The question is of some importance in real pattern recognition problems, since we have not always a sufficient number of learning data to build the classifier, and in this case all information about features, even vague, is invaluable for improving the recognition. For the case of fuzzy integral classifier, a lower bound on the minimal number of training data has been established [17], which grows exponentially with the number of features. This fact, together with the difficulty of optimizing fuzzy measures, mean that when the number of features increases, the result of learning of fuzzy measures is more and more questionable.

The above question has been completely solved in the framework of k-*additive measures* and the *interaction representation* of a fuzzy measure (see full details in [13, 14], or in the the companion paper on k-additive measures in this book). We give here only the necessary details.

A *k-additive measure* μ is a fuzzy measure such that its interaction index $I(A)$, $A \subset X$, is zero for all subsets A of more than k elements. This means that, if an expert gives the Shapley index v_i of all attributes, and the interaction index I_{ij} for all pairs of attributes, it defines uniquely a 2-additive measure. More specifically, we have the following proposition.

Proposition 2 *Let $[v_1 \cdots v_n]$ be a Shapley value, satisfying $\sum_{i=1}^{n} v_i = 1$, and I_{ij}, $\{i,j\} \subset X$ a set of interaction indices. There exists a unique 2-additive measure (possibly non monotonic) defined by*

$$\mu(\{i\}) = v_i - \frac{1}{2} \sum_{j \in X \setminus \{i\}} I_{ij}, \quad i = 1, \ldots, n \tag{8}$$

$$\mu(\{i,j\}) = v_i + v_j - \frac{1}{2} \sum_{k \neq i,j} (I_{ik} + I_{jk}), \quad \{i,j\} \subset X. \tag{9}$$

It can be said that this 2-additive fuzzy measure is the *least specific* regarded to the information given. Any different fuzzy measure implicitly adds some

information at a higher level (which could be defined as interaction indexes of more than 2 features).

A problem arises with the monotonicity of fuzzy measures. The above theorem does not ensure that the resulting fuzzy measure is monotonic as requested in the definition. Although non monotonic fuzzy measures exist and can have some applications, they are not suitable here since non monotonicity of the fuzzy measure implies non monotonicity of the integral. But a fusion operator which would be non monotonic will inevitably lead to inconsistent results. In order to ensure monotonicity of the 2-additive fuzzy measure, the v_i's and I_{ij}'s must verify a set of constraints. Adapting the general result of [14] to our case, we get the following.

Proposition 3 *A Shapley value $[v_1 \cdots v_n]$ and a set of interaction indices I_{ij}, $\{i,j\} \subset X$ lead to a monotonic 2-additive fuzzy measure if and only if they satisfy the following set of constraints:*

$$-v_i \leq \frac{1}{2}\left(\sum_{j \in X \setminus K \cup \{i\}} I_{ij} - \sum_{k \in K} I_{ik}\right) \leq v_i, \quad K \subset X \setminus \{i\}, \quad i = 1, \ldots, n.$$

We apply these results on the iris data set. Following the same observations we have made on the histograms and projections, we propose in table 4 the following set of importance and interaction index (this set satisfies the above constraints). It can be seen that very simple values have

	class 1	class 2	class 3
$n \cdot v_1$	0.4	0.4	0.4
$n \cdot v_2$	0.4	0.4	0.4
$n \cdot v_3$	1.6	1.4	1.4
$n \cdot v_4$	1.6	1.8	1.8
I_{12}	0.1	0.	0.
I_{13}	0.	0.	0.
I_{14}	0.	-0.2	-0.2
I_{23}	0.	0.	0.
I_{24}	0.	0.	0.
I_{34}	0.	0.45	0.45

Table 4: Set of scaled importance and interaction indexes for the iris data set

been given, setting I_{ij} to 0 when there is no clear evidence of redundancy or complementarity. The satisfaction of the constraints is not difficult to obtain by a trial and error method, since few values of I_{ij} are non zero, and the constraint $\sum_{i=1}^{n} v_i = 1$ is easy to satisfy, and entails $\mu(X) = 1$. We give as illustration the values of the obtained fuzzy measure for class 2 in table 5.

We have used these identified 2-additive fuzzy measures for recognition.

subset A	$\mu(A)$	subset A	$\mu(A)$	subset A	$\mu(A)$
$\{1\}$	0.2	$\{1,2\}$	0.3	$\{1,2,3\}$	0.425
$\{2\}$	0.1	$\{1,3\}$	0.325	$\{1,2,4\}$	0.425
$\{3\}$	0.125	$\{1,4\}$	0.325	$\{1,3,4\}$	0.9
$\{4\}$	0.325	$\{2,3\}$	0.225	$\{2,3,4\}$	1.
		$\{2,4\}$	0.425		
		$\{3,4\}$	0.9		

Table 5: The 2-additive fuzzy measure for class 2 obtained from coefficients of table 4

Method	recognition rate with 90% data for learning (%)	recognition rate with 20% data for learning (%)
HLMS	95.3	93.8
MIN	94.	92.6
MEAN	95.3	92.5
2-additive	96.	95.5

Table 6: Results of recognition with identified 2-additive measures

The results are shown on table 6. Two experiments were done. The first is similar to the previous ones, taking a 10-fold cross validation (90% of data for learning and the remaining 10% for testing). The second one is a random subsampling with 20 runs and 20% of data for learning, in order to illustrate what was said above about the effect of the lack of a sufficient number of training data. We have compared in these two experiments one of the usual methods of learning of fuzzy measures (HLMS), the minimum (MIN) and the arithmetic mean (MEAN) operator (which are particular cases of fuzzy integrals), and the above explained method identifying a 2-additive fuzzy measure. MIN and MEAN can be considered as special cases when no information is available on the importance and interaction of features. Although there is no learning of the fuzzy measures for MIN and MEAN, it remains nevertheless a learning procedure for the ϕ_i^J's.

The results show that surprisingly enough, this last method is even better in the usual case of sufficient amount of learning data. The improvement obtained in the case of few training data is significant, which leads us to the conclusion that the concepts of importance and interaction indexes presented here are meaningful and useful in applications.

References

[1] T. Arbuckle, E. Lange, T. Iwamoto, N. Otsu, and K. Kyuma. Fuzzy information fusion in a face recognition system. *Int. J. of Uncertainty, Fuzziness and Knowledge-Based Systems*, 3(3):217–246, 1995.

[2] J.C. Bezdek. *Pattern recognition with fuzzy objective function algorithms*. Plenum Press, New York, 1981.

[3] J.C. Bezdek and S.K. Pal (eds). *Fuzzy Models for Pattern Recognition*. IEEE Press, 1992.

[4] S.B. Cho and J.H. Kim. Combining multiple neural networks by fuzzy integral for robust classification. *IEEE Tr. on Systems, Man, and Cybernetics*, 25(2):380–384, 1995.

[5] G. Choquet. Theory of capacities. *Annales de l'Institut Fourier*, 5:131–295, 1953.

[6] D. Dubois and H. Prade. *Possibility Theory*. Plenum Press, 1985.

[7] D. Dubois, H. Prade, and S. Sandri. On possibility/probability transformations. In R. Lowen and M. Roubens, editors, *Fuzzy Logic, State of the Art*. Kluwer Academic, 1993.

[8] D. Dubois, H. Prade, and C. Testemale. Weighted fuzzy pattern matching. *Fuzzy Sets & Systems*, 28:313–331, 1988.

[9] R.O. Duda and P.E. Hart. *Pattern classification and scene analysis*. Wiley-Interscience, 1973.

[10] M. Grabisch. A new algorithm for identifying fuzzy measures and its application to pattern recognition. In *Int. Joint Conf. of the 4th IEEE Int. Conf. on Fuzzy Systems and the 2nd Int. Fuzzy Engineering Symposium*, pages 145–150, Yokohama, Japan, March 1995.

[11] M. Grabisch. The application of fuzzy integrals in multicriteria decision making. *European J. of Operational Research*, 89:445–456, 1996.

[12] M. Grabisch. The representation of importance and interaction of features by fuzzy measures. *Pattern Recognition Letters*, 17:567–575, 1996.

[13] M. Grabisch. Alternative representations of discrete fuzzy measures for decision making. *Int. J. of Uncertainty, Fuzziness, and Knowledge Based Systems*, 5:587–607, 1997.

[14] M. Grabisch. k-order additive discrete fuzzy measures and their representation. *Fuzzy Sets and Systems*, 92:167–189, 1997.

[15] M. Grabisch, G. Bienvenu, A. Ayoun, J.F. Grandin, A. Lemer, and M. Moruzzi. A formal comparison of probabilistic and possibilistic frameworks for classification. In *7th IFSA World Congress*, Prague, Czech Republic, June 1997.

[16] M. Grabisch, H.T. Nguyen, and E.A. Walker. *Fundamentals of Uncertainty Calculi, with Applications to Fuzzy Inference*. Kluwer Academic, 1995.

[17] M. Grabisch and J.M. Nicolas. Classification by fuzzy integral — performance and tests. *Fuzzy Sets & Systems, Special Issue on Pattern Recognition*, 65:255–271, 1994.

[18] M. Grabisch, S.A. Orlovski, and R.R. Yager. Fuzzy aggregation of numerical preferences. In R. Słowiński, editor, *Fuzzy Sets in Decision Analysis, Operations Research and Statistics*. Kluwer Academic, 1998.

[19] M. Grabisch and M. Roubens. An axiomatic approach to the concept of interaction among players in cooperative games. *Int. Journal of Game Theory*, submitted.

[20] M. Grabisch and M. Sugeno. Fuzzy integrals and dual measures : application to pattern classification. In *Sino-Japan Joint Meeting on Fuzzy Sets & Systems*, Beijing, China, October 1990.

[21] M. Grabisch and M. Sugeno. Multi-attribute classification using fuzzy integral. In *1st IEEE Int. Conf. on Fuzzy Systems*, pages 47–54, San Diego, CA, March 1992.

[22] R.L. Keeney and H. Raiffa. *Decision with Multiple Objectives*. Wiley, New York, 1976.

[23] J.M. Keller, M.R. Gray, and jr. J.A. Givens. A fuzzy k-nearest neighbor algorithm. *IEEE Trans. on Syst., Man & Cyb.*, 15(4):580–585, 1985.

[24] J.M. Keller, H. Qiu, and H. Tahani. The fuzzy integral and image segmentation. In *North Amer. Fuzzy Information Proc. Soc.*, pages 324–339, New Orleans, June 1986.

[25] O. Metellus and M. Grabisch. Une approche de la classification par filtrage flou — méthodologie et performances sur un problème de segmentation clientèle. In *Proc. Rencontres Francophones sur la Logique Floue et ses Applications (LFA)*, pages 215–220, Paris, France, November 1995.

[26] L. Mikenina and H.-J. Zimmermann. Improved feature selection and classification by the 2-additive fuzzy measure. *Fuzzy Sets and Systems*, to appear.

[27] K. Miyajima and A. Ralescu. Modeling of natural objects including fuzziness and application to image understanding. In *2nd IEEE Congr. on Fuzzy Systems*, pages 1049–1054, San Francisco, CA, March 1993.

[28] T. Murofushi and S. Soneda. Techniques for reading fuzzy measures (III): interaction index. In *9th Fuzzy System Symposium*, pages 693–696, Sapporo, Japan, May 1993. In Japanese.

[29] S.K. Pal and D.K. Dutta Majumder. *Fuzzy mathematical approach to pattern recognition*. Wiley Eastern Ltd., 1986.

[30] H. Qiu and J.M. Keller. Multiple spectral image segmentation using fuzzy techniques. In *Proc. North amer. Fuzzy Information Proc. Soc.*, pages 374–387, Purdue Univ., May 1987.

[31] G. Shafer. *A Mathematical Theory of Evidence*. Princeton Univ. Press, 1976.

[32] L.S. Shapley. A value for n-person games. In H.W. Kuhn and A.W. Tucker, editors, *Contributions to the Theory of Games, Vol. II*, number 28 in Annals of Mathematics Studies, pages 307–317. Princeton University Press, 1953.

[33] M. Sugeno. *Theory of fuzzy integrals and its applications*. PhD thesis, Tokyo Institute of Technology, 1974.

[34] M. Sugeno. Fuzzy measures and fuzzy integrals — a survey. In Gupta, Saridis, and Gaines, editors, *Fuzzy Automata and Decision Processes*, pages 89–102. 1977.

[35] H. Tahani and J.M. Keller. Information fusion in computer vision using the fuzzy integral. *IEEE Tr. on Systems, Man, and Cybernetics*, 20(3):733–741, 1990.

[36] S.M. Weiss and I. Kapouleas. An empirical comparison of pattern recognition, neural nets, and machine learning classification methods. In *11th IJCAI*, pages 781–787, 1989.

[37] L.A. Zadeh. Fuzzy sets. *Information and Control*, 8:338–353, 1965.