Publishers' page
Publishers’ page
to World Scientific Publishing Company
Preface
Contents

Preface vii

Chapter 1 Fuzzy Pattern Recognition by Fuzzy Integrals and Fuzzy Rules 1
1.1 Introduction ........................................... 1
1.2 Classification by fuzzy rules .......................... 2
  1.2.1 Fuzzy Rules ...................................... 2
  1.2.1.1 Fuzzy Logic Implications .................. 2
  1.2.2 The Generalised Modus Ponens ................. 4
  1.2.3 Classification by fuzzy rules .................. 4
  1.2.4 Other approaches based on fuzzy rules .......... 9
1.3 Classification by fuzzy integrals ..................... 10
  1.3.1 Background on fuzzy measures and integrals ...... 10
  1.3.2 Classification by fuzzy integral ................ 11
    1.3.2.1 General methodology ..................... 11
    1.3.2.2 The fuzzy pattern matching approach ......... 12
  1.3.3 Learning of fuzzy measures ..................... 13
    1.3.3.1 Performance .............................. 15
  1.3.4 Importance and interaction of attributes ........ 16
    1.3.4.1 Shapley value and interaction index of a fuzzy
            measure .................................. 18
  1.3.5 Related works .................................. 20

Bibliography .............................................. 23
Chapter 1

Fuzzy Pattern Recognition by Fuzzy Integrals and Fuzzy Rules

1.1 Introduction

Since the introduction by Zadeh of the concept of fuzzy set in 1965 [35], and later of the concept of possibility measure in 1978 [36], that is, a non-additive representation of uncertainty, many methods in pattern recognition have been proposed, based on these two seminal ideas. We should cite in this respect the pioneering work of Ruspini [30], who gave the basis of fuzzy clustering in 1969, an area which has known an important development later due to the work of Bezdek [3].

It would be a difficult task to cite here all important contributions to the field of pattern recognition done by researchers in fuzzy logic (but the reader can consult the edited volume of Bezdek and Pal [4] for a selected collection of papers on this topic), and we will limit ourself in this paper to the presentation of only two methodologies, which are representative of the two seminal ideas of Zadeh cited above. The first one is based on fuzzy sets, as a mean to model imprecise concepts, and uses fuzzy rules to represent the classes of interest. The second one is based on non-additive representations of uncertainty, namely fuzzy measures [32], which are yet more general than possibility (and also probability) measures.

The approaches described here are based on the work done by the author. Related works will be also indicated.

In the sequel, $\wedge, \vee$ denote min and max respectively,
1.2 Classification by fuzzy rules

1.2.1 Fuzzy Rules

The following exposition is based on possibility theory (see [8] for basic definitions). The typology of fuzzy rules we use below is based on [9].

1.2.1.1 Fuzzy Logic Implications

We call multivalued implication any operation \( I \) from \([0, 1]\) to \([0, 1]\) which extends the material implication of classical binary logic, i.e.

\[
\begin{align*}
I(0, 0) &= 1 \\
I(0, 1) &= 1 \\
I(1, 0) &= 0 \\
I(1, 1) &= 1.
\end{align*}
\] (1.1)

(1.2)

The material implication can be written as

\[
I(a, b) = \neg a \lor b. \tag{1.3}
\]

The first way to extend \( I \) on \([0, 1]\) is to replace usual binary operations by their corresponding fuzzy operations, i.e.

\[
I(a, b) = n(a) \bot b
\]

where \( \bot \) is a t-conorm, and \( n \) a strong negation (i.e. strictly decreasing, involutive, and such that \( n(0) = 1 \)). These implications are called S-implications since S often denotes t-conorms. Examples of S-implications are those of Kleene-Dienes, Reichenbach and Lukasiewicz, defined by:

\[
\begin{align*}
I^{KD}_{KD}(a, b) &= (1 - a) \lor b \\
I_{R}(a, b) &= 1 - a + ab \\
I^{L}_{L}(a, b) &= (1 - a + b) \land 1.
\end{align*}
\] (1.4) (1.5) (1.6)

The second way of considering an implication is related to sets. We say that \( p \rightarrow q \) if in all worlds where \( p \) is true, then \( q \) is true, i.e.

\[
P \subseteq Q
\]

where \( P \) (resp. \( Q \)) denotes the set of worlds where \( p \) (resp. \( q \)) is true. But inclusion is a particular order relation, and in lattice theory, residuated operations are defined for lattices having a structure of semi-group. Using
a t-norm \( \mathcal{T} \) for the semi-group operation, we can define

\[
I(a, b) = \sup\{c \in [0, 1]|a \mathcal{T} c \leq b\}.
\]

These implications are called residiuated implications or R-implications. Examples of R-implications are those of Łukasiewicz and Gödel, the last one being defined by:

\[
I_G(a, b) = \begin{cases} 
1, & \text{if } a \leq b \\
 b, & \text{otherwise}
\end{cases}
\]

Using these implications, two types of rules can be defined [9]:

**Uncertainty rules** (or uncertainty-qualifying rules)

They have the following semantical interpretation:

**The more** \( X \) **is** \( A \), **the more certain** \( Y \) **is** \( B \)

i.e. \( \mu_A(x) \leq C(B) \), where \( C(B) \) is the certainty degree of having \( B \) when the input is \( x \), and \( \mu_A \) is the membership function of the fuzzy set \( A \). It can be shown that S-implications are suitable for this case.

**Gradual rules** (see [10]) (or truth-qualifying rules)

They have semantically the following meaning:

**The more** \( x \) **is** \( A \), **the more** \( y \) **is** \( B \)

where \( A, B \) are fuzzy sets with membership functions \( \mu_A, \mu_B \). This kind of rule expresses a relation between \( x \) and \( y \) in the sense that as \( X \) goes closer to \( A \), \( Y \) is constrained to go closer to \( B \). This constraint can be expressed by \( \mu_A(x) \leq \mu_B(y) \). The least specific solution (i.e. leading to the greatest possibility distribution) is the Gaines-Rescher implication defined by

\[
I_{GR}(a, b) = \begin{cases} 
1, & \text{if } a \leq b \\
0, & \text{otherwise}
\end{cases}
\]

since if \( (x, y) \) is such that \( A(x) \leq B(y) \), then \( I(A(x), B(y)) = 1 \) and 0 otherwise. In fact, all R-implications are suitable for modelling gradual rules.
1.2.2 The Generalised Modus Ponens

The classical modus ponens is the following reasoning process:

\[ p \rightarrow q \] (rule)
\[ p \] (fact)

\[ q \] (conclusion)

The extension to fuzzy sets becomes:

\[ X \in A \rightarrow Y \in B \] (rule)
\[ X \in A' \] (fact)

\[ Y \in B' \] (conclusion)

The conclusion \(B'\) is computed as follows. We consider the rule as a conditional possibility distribution \(\pi_{Y|X}(x, y) = I(A(x), B(y))\), and the possibility distribution of \(y\) after inference is obtained by the combination/projection principle [9]

\[ \pi_Y(y) = \mu_{B'}(y) = \sup_x (\mu_A(x) \land \pi_{Y|X}(x, y)). \]

In the case of a precise input \(x_0\) (i.e. a distribution reduced to a single value), the above expression reduces to:

\[ \pi_Y(y) = I(\mu_A(x_0), \mu_B(y)). \]

Let us now consider the case of several parallel rules \(R_1, \ldots, R_t\). Each rule \(R_j\) is expressed by a possibility distribution \(\pi^j_{X|Y}(x, y)\). It is known from [9] that inferred possibility distributions have to be aggregated by a minimum operator. Moreover, it is better (i.e. more informative) to aggregate all the rules in one single rule before performing the inference, than aggregating the inferred result of each rule, thus:

\[
\sup_x \left( \pi_X(x) \land \left[ \bigwedge_j \pi^j_{Y|X}(x, y) \right] \right) \leq \bigwedge_j \sup_x (\pi_X(x) \land \pi^j_{Y|X}(x, y)).
\]

1.2.3 Classification by fuzzy rules

Let \(C_1, \ldots, C_m\) be a set of classes, which can be described by a set of features or attributes \(X_i, i = 1, \ldots, n\), i.e. a given pattern to classify is
Classification by fuzzy rules

an element $x = (x_1, \ldots, x_n)$ of $X_1 \times \cdots \times X_n$, where $x_i$ is the value taken by attribute $i$ for this pattern. In the sequel, $X_i$ will indicate either the attribute (or variable) itself or its set of values, while $x_i$ indicate possible values of $X_i$. Fuzzy rules can be used in two different respects:

- **gradual rules** express relations between variables, and can be used to compute approximately the value of an attribute which could be difficult or costly to obtain or compute by an exact mathematical way.

Such situations often occur, when it is sufficient to have a rough approximation of the value of an attribute for classification. More specifically, let $X_j$ be such an attribute, whose value depends on variables $Z_1, \ldots, Z_p$, which could be attributes among $X_1, \ldots, X_n$, or additional variables. Instead of giving an explicit model of the form $x_j = f(z_1, \ldots, z_n)$, with $f$ being a deterministic function, the model is given under a set of gradual rules of the form:

* **The more** $Z_1$ **is** $A_1$ **and** ... **and the more** $Z_p$ **is** $A_p$.
  * **the more** $X_j$ **is** $B_j$.

where $A_1, \ldots, A_p, B_j$ are fuzzy sets on the respective universes. This kind of rule expresses a constraint between $X_j$ and $Z_1, \ldots, Z_p$: as $Z_1, \ldots, Z_p$ go nearer the core of the fuzzy sets $A_1, \ldots, A_p$, the value of $X_j$ is constrained to go nearer the core of $B_j$. A gradual equivalence (i.e. a gradual rule $p \rightarrow q$ together with its converse $\neg p \rightarrow \neg q$) can be used as well if necessary. In this case, there will be similar constraints on the negations of $B_j$.

Note that the result of a gradual rule is a fuzzy set, even if the input is crisp. Depending on the application, one should decide to keep the result under the form of a fuzzy set, and to forward it in subsequent layers of the reasoning, or to extract from it the most **representative** or **plausible** value. We avoid here the term “defuzzification”, which is so often associated with the center of gravity method. If such a method can be justified in the domain of fuzzy control (where fuzzy rules are Mamdani’s type rules, not belonging to the taxonomy given in section 1.2.1.1), it has no meaning in the context of possibility theory, where the representative value should be of maximal degree of possibility.

See an example of application of gradual rules in classification in [2].
• **uncertainty rule** express relations between variables and classes.
  More specifically, let us introduce a new variable $Y$ defined on the set of classes $\{C_1, \ldots, C_m\}$, which we can call the *classification variable*. A classification rule will have the form:

$$
\text{The more } X_1 \text{ is } A_1, \text{ and } \ldots \text{ and the more } X_n \text{ is } A_n,
\text{ the more certain } Y \text{ is } B
$$

where $A_1, \ldots, A_n$ are fuzzy sets expressing fuzzy domains of values for attributes, and $B$ is the possibility distribution on classes, i.e. $B(C_i)$ expresses to which degree it is possible that $C_i$ is the right class when the pattern matches perfectly the fuzzy sets $A_1, \ldots, A_n$.

In practice, several such rules will be used in parallel. It is to be noted that, since the aggregation of rules is made in a conjunctive way, if one of the rules concludes that the possibility degree of class $C_i$ is 0, then no other rule can “save” this conclusion, i.e. the class $C_i$ will be considered as impossible.

So basically, the classification is performed by a set of uncertainty rules, with possibly some additional gradual rules to compute some attributes. The result of the inference is thus a possibility distribution $\pi_Y$ indicating the possibility degree for each class. If $\pi_Y(C_i) = 0$, then the rule has inferred that the observed object cannot belong to class $C_i$. If $\pi_Y(C_i) = 1$, then it is totally possible (but not certain) that the object belongs to class $C_i$. The question is now how to exploit this possibility distribution in order to draw a conclusion on the class of the pattern in consideration.

Recall that a possibility distribution $\pi$ induces a possibility measure $\Pi$ defined by $\Pi(A) = \sup_{x \in A} \pi(x)$ and a necessity measure $N(A) = 1 - \Pi(A)$. Then:

- if there is a unique class $C_i$ such that $\pi_Y(C_i) = 1$, class $C_i$ is the most certain class since $N(C_i) > 0$ and $N(C_j) = 0$ for all $j \neq i$.
- if several $C_i$ are such that $\pi_Y(C_i) = 1$, we cannot decide between these classes, which are the most certain: the rule base has not enough knowledge to discriminate. However, if the classes with possibility degree equal to 1 corresponds to a meaningful subset of classes, then the certainty of this subset is strictly positive. If $\pi_Y(C_i) = 1$ everywhere, then there is total uncertainty about the class.
Classification by fuzzy rules

- if there is no class $C_i$ such that $\pi_Y(C_i) = 1$, then no class matches perfectly the pattern. It means that the pattern could belong to a class which is not included into the initial set of classes. It could mean also that there is some contradiction into the rule base.

Further processing can be done, but it depends strongly on the application: see the example below with satellite images.

It is important to stress here the advantage of possibility theory over probabilistic methods for the representation of uncertainty. The forced normalization of probability measures prevents them to distinguish between equi-certainty and low certainty due to the presence of an unknown class.

This methodology has been successfully applied by the author on target classification [2], face recognition [13], and in satellite image analysis (unpublished). It is to be noted that in every case, no automatic learning method was used to derive the rules, but only expert knowledge. The following example, borrowed from the satellite image analysis application, illustrates this. The problem is concerned with the recognition of various natural and artificial zones on an image (decametric image). The expert in the field is able to tell the following:

- water appears as regions of almost uniform texture. The grey level can be black (pure water) to middle grey (turbid water)
- forests form regions which are slightly more textured than water, and are generally dark (but less dark than pure water)
- fields on a small scale present almost no texture. The grey level varies from dark to light grey
- urban zones are highly textured and are bright.

From this expertise, the following attributes seem to be relevant:

- mean of grey levels on a $3 \times 3$ window, denoted $\text{Moy}$
- variance of the grey levels on a $3 \times 3$ window, denoted $\text{Var}$
- "busyness" type II on a $3 \times 3$ window, denoted $\text{Busy2}$.

The busyness index has been proposed by Dondoni and Rosenfeld [7; 29]. The type II busyness is defined as follows. Considering a $3 \times 3$ window:

\[
\begin{array}{ccc}
  a & b & c \\
  d & e & f \\
  g & h & i \\
\end{array}
\]
we compute the 12 absolute differences $\delta_{ab}, \delta_{bc}, \delta_{de}, \delta_{ef}, \delta_{gh}, \delta_{hi}, \delta_{ad}, \delta_{dg}, \delta_{be}, \delta_{eh}, \delta_{ef}, \delta_{fi}$, among $\delta_{ab} = |a - b|$, and similarly for the other quantities. The type II busyness index corresponds to the median of these 12 differences:

$$B_2 = \text{med}(\delta_{ab}, \delta_{bc}, \delta_{de}, \delta_{ef}, \delta_{gh}, \delta_{hi}, \delta_{ad}, \delta_{dg}, \delta_{be}, \delta_{eh}, \delta_{ef}, \delta_{fi}).$$

Using these three attributes, we have built a fuzzy rule base, containing 7 fuzzy rules, all of the uncertainty type, with Kleene-Dienes implication. We give some examples of this rule base.

- **Fuzzy rule 1**

  **IF** (Moy is VERY DARK) AND (Busy2 is LOW),
  **THEN** it is almost surely WATER

  The fuzzy sets concerning this rule are shown below. The possibility distribution over the classes is represented by vertical strokes.

- **Fuzzy rule 5**

  **IF** (Sig is HIGH),
  **THEN** it can be neither FIELD nor WATER

  The fuzzy sets and possibility distribution concerning this rule are shown below.
As explained above, the result of inference is a possibility distribution over the 4 classes. The final decision is taken as follows. For each pixel \( p \), we first assign the class of highest possibility degree. Then, we compute the difference \( \Delta_{12}(p) \) between the class of maximal possibility degree and the second highest possibility degree. This gives the quality of the decision (the highest \( \Delta_{12}(p) \), the best the quality), which is 0 if the 2 best classes have equal possibility degree. Considering a window around \( p \), we compute the average of decisions in the window, weighted by the quality of the decision. If the average is 0 for each class, then the pixel is classified into the \textsc{indeterminate} class.

Experimental results on various images show that this method, although without learning procedure, largely outperforms classical approaches such as neural nets, clustering and nearest prototype, and also fuzzy integrals, presented hereafter.

1.2.4 Other approaches based on fuzzy rules

Many authors have proposed fuzzy rule based approaches for classification, but to the knowledge of the author, all these approaches are more or less based on Mamdani's type rule or similar (e.g. Sugeno rules), i.e. rules where the implication is in fact a conjunction, namely the minimum operator for Mamdani's rules, and the product operator for Sugeno's rule. For this reason, let us call them "conjunctive rules". This type of rule, although widely used, cannot be called properly a rule in the logical sense of the term, but it performs a smooth interpolation between fuzzy domains. In this respect, conjunctive rules can be viewed as a mean to "fill in" the holes in a partial description of the relation between attributes and classes, given under the form of (fuzzy) examples. For this reason, these methods often lack the concise form of the rule base which can be obtained with the previous approach: for example, the rule base in [2] contains only 19 rules for a 4 classes and 16 attributes problem, and the above example uses 7 rules. On the other hand, some methods based on conjunctive rules use nearly as many rules as learning samples in the database!

In fact, most of the works devoted to the learning of fuzzy rules concern rather the modeling of a continuous function (e.g. for prediction) than classification (see the monograph of GIOrencec [14] for a thorough survey of this topic), and are not completely appropriate. Few methods are available, which construct fuzzy rules from learning data, including the structure of
the fuzzy rule base. Such a complete method, including a detailed analysis of performance, has been proposed by Mandal et al. \cite{24, 25}.

1.3 Classification by fuzzy integrals

1.3.1 Background on fuzzy measures and integrals

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

**Definition 1** A fuzzy measure $\mu$ defined on $X$ is a set function $\mu : \mathcal{P}(X) \rightarrow [0, 1]$ 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 $\mu$ 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 \cite{32} and Choquet \cite{6}.

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

$$C_\mu(f(x_1), \ldots, f(x_n)) := \sum_{i=1}^n (f(x_{(i)}) - f(x_{(i-1)})) \mu(A_{(i)}) \quad (1.7)$$

where $x_{(i)}$ 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 $f : X \rightarrow [0, 1]$ with respect to $\mu$ is defined by

$$S_\mu(f(x_1), \ldots, f(x_n)) := \bigvee_{i=1}^n (f(x_{(i)}) \wedge \mu(A_{(i)})), \quad (1.8)$$

with same notations.

Choquet integral coincides with Lebesgue integral when the measure is additive, but this is not the case for the Sugeno integral.
1.3.2 Classification by fuzzy integral

1.3.2.1 General methodology

As before, 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 (or sources), one for each feature (attribute), which provide for an unknown sample $x^o$ a degree of confidence in the statement “$x^o$ belongs to class $C_j$”, for all $C_j$. We denote by $\phi^i_j(x^o)$ the confidence degree delivered by source $i$ (i.e. feature $i$) of $x^o$ belonging to $C_j$.

The second step is then 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 [15]. 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 1.3.4). Thus the global confidence degree in the statement “$x^o$ belongs to $C_j$” is given by:

$$\Phi_{\mu^i}(C_j; x^o) := C_{\mu^i} (\phi^1_j(x^o), \ldots, \phi^n_j(x^o))$$  \hspace{1cm} (1.9)

(or similarly with the Sugeno integral). Finally, $x^o$ is put into the class of highest confidence degree. Here, the fuzzy measures $\mu^i$ (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, $\mu^i(\{X_1\})$ expresses the relative importance of attribute 1 for distinguishing class $j$ from the others, while $\mu^i(\{X_1, X_2\})$ expresses the relative importance of attributes 1 and 2 taken together for the same task. The precise way of how to interpret this will be given in section 1.3.4.

The above presentation 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. Due to its importance, we devote the next paragraph to the presentation of this methodology, its connection with fuzzy integral, and the Bayesian approach.
1.3.2.2 The fuzzy pattern matching approach

As for the section on fuzzy rules, we assume some familiarity of the reader with possibility theory (see [8] for this topic, and [12] for fuzzy pattern matching). Following previous notations, \( X_i \) denotes the universe of the \( i \)th attribute. Each class \( C_j \) is modelled by a collection of fuzzy sets \( C_1^i, \ldots, C_n^i \) defined on \( X_1, \ldots, X_n \) 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 \( \pi_x(x_1, \ldots, x_n) \), representing the distribution of possible locations of the (unknown) true value of \( x \) in \( \times_{i=1}^n X_i \). If attributes are considered to be non-interactive, then \( \pi_x(x_1, \ldots, x_n) = \land_{i=1}^n \pi_i(x_i) \). Now the possibility and necessity degrees that datum \( x \) matches class \( C_j \) w.r.t attribute \( i \) is given by

\[
\Pi_{\pi_i}(C_j^i) := \sup_{x_i \in X_i} (C_j^i(x_i) \land \pi_i(x_i)) \]

\[
N_{\pi_i}(C_j^i) := \inf_{x_i \in X_i} (C_j^i(x_i) \lor (1 - \pi_i(x_i))).
\]

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, \( \pi_x \) reduces to a point, and the two above quantities collapse into \( C_j^i(x_i) \), which corresponds to \( \phi_i(x^i) \).

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 \( C_j := C_j^1 \times \cdots \times C_j^n \),

\[
\Pi_{ \pi_j}(C_j) = \land_{i=1}^n \Pi_{\pi_i}(C_j^i)
\]

\[
N_{ \pi_j}(C_j) = \land_{i=1}^n N_{\pi_i}(C_j^i).
\]

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 \( \mu \), a construction which encompasses
all previous cases, \( \Pi_x(C_j) \) and \( N_x(C_j) \) are also obtained by the (same) Sugeno integral. More specifically:

**Proposition 1** Let \( \mu \) be a fuzzy measure, and consider that class \( C \) is expressed by a Sugeno integral, i.e. \( C(x_1, \ldots, x_n) = \bigvee_{i=1}^n [C^{(i)}(x_{\{i\}}) \land \mu(A_{\{i\}})] \). Then, the possibility and necessity degrees that a datum \( x \) belongs to class \( C_j \) is given by

\[
\Pi_x(C) = S_\mu(\Pi_{x_1}(C^1), \ldots, \Pi_{x_n}(C^n)) \\
N_x(C) = S_\mu(N_{x_1}(C^1), \ldots, N_{x_n}(C^n))
\]

For a proof, see [18].

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 statistically independent, the above formula becomes:

\[
\Phi(C_j|x) = \prod_{i=1}^n p(x_i|C_j)P(C_j) \tag{1.10}
\]

If the classes have equal a priori probability, formulae (1.8) and (1.10) are similar: in probability theory and in the case of independence, the product operator takes place of the aggregation operator.

1.3.3 **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 \( \phi^j_i \) have already been obtained by some parametric or non-parametric classical probability density estimation method, after suitable normalization: possibilistic histograms (that is, transformations of probability density functions

*Sugeno integrals, as shown by Marichal [26], represent a wider class of Boolean polynomials, i.e. polynomials formed uniquely by a combination of \( \lor \) and \( \land \).*
into possibility distributions, see [11]), Parzen windows, Gaussian densities, etc.

The identification of the fusion operator reduces to the identification (or learning) of the fuzzy measures \( \mu^j \), that is, \( m(2^n - 2) \) coefficients. 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 \( x^1_j, x^2_j, \ldots, x^j_{l_j} \) 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.

\[
J = \sum_{k=1}^{l_1} (\Phi_{\mu^1}(C_1; x^1_k) - \Phi_{\mu^2}(C_2; x^1_k) - 1)^2 + \sum_{k=1}^{l_2} (\Phi_{\mu^2}(C_2; x^2_k) - \Phi_{\mu^1}(C_1; x^2_k) - 1)^2.
\]

It can be shown that this reduces to a quadratic program with \( 2(2^n - 2) \) variables and \( 2n(2^{n-1} - 1) \) constraints in the case of Choquet integral (see full details in [20; 19]).

- 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 \( \Psi \) being any increasing function from \([-1, 1]\) to \([-1, 1]\). \( \Psi \) 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. This is no more 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 looked towards heuristic algorithms better adapted to the peculiar structure of the problem and less greedy [16]. A satisfying algorithm has been found (called hereafter HLMS), which although suboptimal, reduces the computing time by a factor 200.
1.3.3.1 Performance

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 1.1 (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 are highly non-Gaussian. The results by classical methods come from a paper of Weiss and Kapouleas [34]. The good performance of HLMS on the difficult cancer data is to be noted. The bottom part

<table>
<thead>
<tr>
<th>Method</th>
<th>iris (%)</th>
<th>cancer (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>linear</td>
<td>98.0</td>
<td>70.6</td>
</tr>
<tr>
<td>quadratic</td>
<td>97.3</td>
<td>65.6</td>
</tr>
<tr>
<td>nearest neighbor</td>
<td>96.0</td>
<td>65.3</td>
</tr>
<tr>
<td>Bayes independent</td>
<td>93.3</td>
<td>71.8</td>
</tr>
<tr>
<td>Bayes quadratic</td>
<td>84.0</td>
<td>65.6</td>
</tr>
<tr>
<td>neural net</td>
<td>96.7</td>
<td>71.5</td>
</tr>
<tr>
<td>PVM rule</td>
<td>96.0</td>
<td>77.1</td>
</tr>
<tr>
<td>QUAD</td>
<td>96.7</td>
<td>68.5</td>
</tr>
<tr>
<td>CLMS</td>
<td>96.0</td>
<td>72.9</td>
</tr>
<tr>
<td>HLMS</td>
<td>95.3</td>
<td>77.4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Classification rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bayes linear</td>
<td>82.6</td>
</tr>
<tr>
<td>linear pseudo-inverse cluster</td>
<td>84.4</td>
</tr>
<tr>
<td>adaptive nearest neighbour</td>
<td>87.8</td>
</tr>
<tr>
<td>Bayes quadratic</td>
<td>90.3</td>
</tr>
<tr>
<td>k nearest neighbour tree</td>
<td>90.4</td>
</tr>
<tr>
<td>CLMS</td>
<td>90.7</td>
</tr>
<tr>
<td>HLMS</td>
<td>89.2</td>
</tr>
</tbody>
</table>

Table 1.1 Classification rate on various data set

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). 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 [27], 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 polytomous 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 1.2, top). 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 gives 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%. 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.

This fact is again an evidence of the advantage of dealing with possibility distributions\(^1\) rather than probability distributions as explained in section 1.2.3.

1.3.4 Importance and interaction of attributes

As said before, the fuzzy measures \( \mu \) contain all the information about the importance of all individual attributes (or features) and all groups of attributes for distinguishing class \( C_j \) from the others. Let us drop index \( j \) for simplicity, and denote by \( X = \{1, \ldots, n\} \) the set of attributes, and \( X_1, \ldots, X_n \) the corresponding axes. We can reasonably state the following.

- a feature \( i \) is important if the values of \( \mu(A) \) are high whenever

\(^1\)Strictly speaking, \( \Phi_{\mu_j}(C_j; x) \) is a possibility distribution only if a Sugeno integral is used. Nevertheless, there is no normalization of \( \Phi \) such as \( \sum_j \Phi_{\mu_j}(C_j; x) = 1 \).
File 1

<table>
<thead>
<tr>
<th>Methods</th>
<th>Classification rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>45 %</td>
</tr>
<tr>
<td>Sequential scores</td>
<td>46.4 %</td>
</tr>
<tr>
<td>Fuzzy integral (HLMS)</td>
<td>47.1 %</td>
</tr>
<tr>
<td>Polytomic scores</td>
<td>50 %</td>
</tr>
<tr>
<td>Polytomic scores (2nd choice)</td>
<td>54 %</td>
</tr>
<tr>
<td>Fuzzy integral (HLMS) (2nd choice)</td>
<td>65 %</td>
</tr>
</tbody>
</table>

File 2

<table>
<thead>
<tr>
<th>Methods</th>
<th>Classification rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>29.9 %</td>
</tr>
<tr>
<td>Sequential scores</td>
<td>27.9 %</td>
</tr>
<tr>
<td>Fuzzy integral (HLMS)</td>
<td>31.1 %</td>
</tr>
<tr>
<td>Polytomic scores</td>
<td>32.2 %</td>
</tr>
<tr>
<td>Polytomic scores (2nd choice)</td>
<td>36 %</td>
</tr>
<tr>
<td>Fuzzy integral (HLMS) (2nd choice)</td>
<td>50 %</td>
</tr>
</tbody>
</table>

Table 1.2 Segmentation of customers

$i \in A$. Clearly, it is not enough to look solely at the value of $\mu\{i\}$, but also at $\mu\{i, j\}$, $\mu\{i, j, k\}$, etc. But it seems very difficult to extract from these coefficients the **contribution of $i$ alone**.

- if $\mu\{1\}$ and $\mu\{2\}$ are high (say 0.7), and $\mu\{1, 2\}$ is not very different (say 0.75), then clearly the importance of features 1, 2 taken together is much the same as 1 or 2 taken separately, and we have no interest in considering them both. We will speak here of **negative synergy** or **redundancy**. On the contrary, if $\mu\{1\}$ and $\mu\{2\}$ have low values (say 0.1) and $\mu\{1, 2\}$ is large (say 0.6), then although features 1 and 2 are unimportant when considered separately, they become very important when taken together. We speak then of **positive synergy**, or **complementarity**. Of course, as above we must consider the value of all coefficients $\mu(A)$ with $\{1, 2\} \subset A$, to see the effect of adding either $i, j$ or $\{i, j\}$ to a coalition.

Based on these ideas, it is possible to compute a global **importance index** and an **interaction index**. We define them in the next section.
1.3.4.1 Shapley value and interaction index of a fuzzy measure

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

\[
v_i = \sum_{K \subseteq X \setminus \{i\}} \gamma_K(\mu(K \cup \{i\}) - \mu(K)) \tag{1.11}
\]

with \( \gamma_k = \frac{(n-k-1)!k!}{n!} = \frac{1}{(\frac{n}{k})n!} \), \(|K|\) indicating the cardinal of \( K \), and \( 0! = 1 \) as usual. The Shapley value of \( \mu \) is the vector \( v = [v_1 \ldots v_n] \).

This definition has been proposed by Shapley on an axiomatic basis in cooperative game theory [31], and possesses all suitable properties for representing importance of indexes. In particular, the Shapley value has the property that \( \sum_{i=1}^{n} v_i = 1 \). 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.

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

\[
I_{ij} = \sum_{K \subseteq X \setminus \{i,j\}} \xi_K(\mu(K \cup \{i,j\}) - \mu(K \cup \{i\}) - \mu(K \cup \{j\}) + \mu(K)) \tag{1.12}
\]

with \( \xi_k = \frac{(n-k-2)!k!}{|n-2|!} = \frac{1}{(\frac{n-2}{k})|n-2|} \).

It is easy to show that the maximum value of \( I_{ij} \) is 1, reached by the fuzzy measure \( \mu \) defined by \( \mu(K \cup \{i,j\}) = 1 \) for every \( K \subseteq X \), and 0 otherwise. Similarly, the minimum value of \( I_{ij} \) is -1, reached by \( \mu \) defined by \( \mu(K \cup \{i\}) = \mu(K \cup \{j\}) = 1 \) for any \( K \subseteq X \) and 0 otherwise. This definition has been proposed by Murofushi and Soneda [28], by using concepts of multiattribute utility theory, and is very similar to the Shapley value. In fact, Grabisch has shown that both can be embedded into a general interaction index, defined for any coalition [17]. A positive (resp. negative) value of the index corresponds to a positive (resp. negative) synergy.

Let us apply these indices to the iris data set, Figures 1.1 and 1.2 give the histograms of every feature for every class, as well as projections of the data set on some pairs of features.

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

<table>
<thead>
<tr>
<th>index of importance $v_i$ (scaled)</th>
</tr>
</thead>
<tbody>
<tr>
<td>feature</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>index of interaction $I_{ij}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>features</td>
</tr>
<tr>
<td>1,2</td>
</tr>
<tr>
<td>1,3</td>
</tr>
<tr>
<td>1,4</td>
</tr>
<tr>
<td>2,3</td>
</tr>
<tr>
<td>2,4</td>
</tr>
<tr>
<td>3,4</td>
</tr>
</tbody>
</table>

Table 1.3  Indexes of importance and interaction for the iris data set

Tables 1.3 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.

The interaction index values 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 $X_1, X_1$, 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.
1.3.5 Related works

To our knowledge, the fuzzy integral has been first applied to classification in the beginning of the nineties, independently by Tahani and Keller [33], and by Grabisch and Sugeno [21, 22]. Later, many works have been done in image processing and character recognition around these people (see [23].
Classification by fuzzy integrals

for a thorough survey on this topic, with a detailed bibliography, see also the work of Arbuckle et al. [1]). Also, we mention the use of fuzzy integrals as a mean to combine the output of several classifiers (see e.g. S.B. Cho [5]).
Bibliography


Bibliography


Bibliography


