The representation of importance and interaction of features by fuzzy measures

Michel GRABISCH
Thomson-CSF, Central Research Laboratory
Domaine de Corbeville, 91404 Orsay cedex, France
Tel +33-1-69-33-92-95 Fax +33-1-69-33-08-65
e-mail grabisch@thomson-lcr.fr

Abstract
We present a new technique for pattern recognition by fuzzy integral, and show how to estimate importance of features, and their interaction. Reciprocally, it is shown how to use expert information about importance and interaction of features to improve recognition.

Keywords: fuzzy pattern matching, fuzzy integral, fuzzy measure, Choquet integral, Shapley value, interaction index, feature extraction

1 Introduction
One of the most important and difficult problem in pattern recognition remains feature extraction, that is, how in a practical problem to choose among a set of measurable features, those which will be most relevant for recognition. A related question concerns the dependency or interaction between features. Statistical dependency is well known, but it is not sure if this is the right concept in pattern recognition. We propose in this paper a quite different concept of interaction, and a tool to model it. The concept of interaction we deal with here is somehow related to information fusion.

We give here a rather informal presentation of our ideas. Suppose we have several sensors, each sensor measuring one (or several) features, which are supposed to be useful for recognition. Each sensor has a decision device, which tells to which class the observed pattern is likely to belong, or more precisely, which gives for each class a degree of confidence or a degree of membership of the pattern into the class. To take a final decision considering all the sensors together, we must take into account both importance of sensors and interaction between them. By importance we mean the following.

Sensor $A$ is more important than sensor $B$ for recognizing class $C_i$ against the others if sensor $A$ has a better discrimination ability than $B$ for $C_i$ (i.e. it can better distinguish $C_i$ from the others classes).

Next consider a pair of sensors $i, j$. We define three kinds of interaction.

- **redundancy or negative synergy**: the importance of the pair of sensors $i, j$ (in the above sense) is not greater than the sum of individual importances. In other words, we do not improve significantly the performance of recognition of a given class by combining sensors $i$ and $j$.

- **complementarity or positive synergy**: the importance of the pair of sensors $i, j$ (in the above sense) is greater than the sum of individual importance. In other words, we do improve significantly the performance of recognition of a given class by combining sensors $i$ and $j$, compared to the importance of $i$ and $j$.

- **independency**: intermediate case, where each sensor brings its contribution to the recognition rate.
It is important to stress that all these concepts are defined with respect to a given class.

The aim of the paper is to present a new method which can model the above defined concepts of importance and interaction. This method is based on the concept of fuzzy measure, also called non-additive measure, and has already lead to a new pattern recognition method with good performance [7]. In the following, we will do a brief presentation of fuzzy measures, fuzzy integrals, and the method of pattern recognition based upon them. Lastly, we will explain in detail the representation of importance and interaction by fuzzy measures.

2 Background on fuzzy measures and integrals

(for more details and definitions in the continuous case see [?]) Let $X$ be a finite index set $X = \{1, \ldots, n\}$.

2.1 Fuzzy measures, pseudo-Boolean functions and k-additive fuzzy measures

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

\begin{enumerate}[(i)]
    \item $\mu(\emptyset) = 0, \mu(X) = 1.$
    \item $A \subseteq B \Rightarrow \mu(A) \leq \mu(B)$
\end{enumerate}

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

When (ii) is not satisfied, $\mu$ is called a non monotonic fuzzy measure. A fuzzy measure on $X$ needs $2^n$ coefficients to be defined, which are the values of $\mu$ for all the different subsets of $X$. In order to avoid heavy notations, $\mu(\{i\}), \mu(\{i, j\})$ and $\mu(K \cup \{i\})$ will be denoted $\mu_i, \mu_{ij}$ and $\mu_{ik}$ respectively. These coefficients are not independent since they must satisfy monotonicity.

In a measure-theoretic point of view, fuzzy measures generalize the concept of classical measure by dropping additivity. In the finite case however, fuzzy measures can be viewed as particular cases of pseudo-Boolean functions, which are functions from $\{0, 1\}^n$ to $\mathbb{R}$. To see the correspondence, simply remark that for any $A \subseteq X$, $A$ is equivalent to a point $(x_1, \ldots, x_n)$ in $\{0, 1\}^n$ such that $x_i = 1$ iff $i \in A$. It can be shown that any pseudo-Boolean function can be put under a multilinear polynomial in $n$ variables:

$$f(x) = \sum_{T \subseteq X} a_T \prod_{i \in T} x_i$$

(1)

with $a_T \in \mathbb{R}$ and $x = (x_1, \ldots, x_n) \in \{0, 1\}^n$.

Some particular cases of interest here. We say that a measure is additive if $\mu(A \cup B) = \mu(A) + \mu(B)$ whenever $A \cap B = \emptyset$. It is easy to see that additive measures can be expressed by the coefficients $\mu_i, i = 1, \ldots, n$ alone. Also, the corresponding pseudo-Boolean function will be linear, i.e. $f(x) = \sum_{i=1}^n a_i x_i$, and remark that $\mu_i \equiv a_i$. By extension, we define $k$-order additive (or simply $k$-additive) fuzzy measures.

Definition 2 A fuzzy measure $\mu$ defined on $X$ is said to be $k$-order additive if its corresponding pseudo-Boolean function is a multilinear polynomial of degree $k$.

Let us detail a little more the 2-additive case, since it will be used in the sequel. For 2-additive fuzzy measures, we have for any $K \subseteq X$:

$$\mu(K) = \sum_{i=1}^n a_i x_i + \sum_{\{i,j\} \subseteq X} a_{ij} x_i x_j$$

(2)

with $x_i = 1$ if $i \in K, x_i = 0$ otherwise. We deduce that $\mu_i = a_i$ for all $i$, and $\mu_{ij} = a_i + a_j + a_{ij} = \mu_i + \mu_j + a_{ij}$. The general formula is:

$$\mu(K) = \sum_{i \in K} a_i + \sum_{\{i,j\} \subseteq K} a_{ij} = \sum_{\{i,j\} \subseteq K} \mu_{ij} - (|K| - 2) \sum_{i \in K} \mu_i,$$

(3)

for any $K \subseteq X$ such that $|K| \geq 2$, $|K|$ being the cardinal of $K$. 

2
2.2 Fuzzy integrals

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 integral, among which the most representative are those of Sugeno and Choquet. We restrict here to the latter.

Definition 3 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

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

where \( A_{i} \) 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 definition can be extended to negative functions too (see e.g. [?]). Choquet integral coincides with Lebesgue integral when the measure is additive.

3 Classification by fuzzy integral

We present now a new method of pattern recognition based on fuzzy measures and integrals. This method will serve as a basis for the representation of importance of features as well as their interaction. For the sake of concision, we will omit details, and refer the reader to previous publications [?, ?] (see also chapter 9 in [?]).

3.1 General presentation

There are several ways to present this method: either as a pseudo-Bayesian approach based on possibility theory [?], as a fuzzy pattern matching based approach [?], or as an information fusion approach, as it was done originally by Tahani and Keller [?]. We will adopt here the latter, as we did in [?].

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, 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 [?]. In particular, fuzzy integrals are able to model some kind of interaction between features: this is the main motivation of the paper. Thus the global confidence degree in the statement “\( X^o \) belongs to \( C_j \)” is given by:

\[
\Phi_{\mu^j}(C_j; X^o) = C_{\mu^j}(\phi^1_j, \ldots, \phi^n_j)
\]

Finally, \( X^o \) is put into the class of highest confidence degree. Here, the fuzzy measures \( \mu^j \) (one per class) are defined on the set of features (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 feature 1 for distinguishing class \( j \) from the others, while \( \mu^j(\{x_1, x_2\}) \) expresses the relative importance of features 1 and 2 taken together for the same task. The precise way of how to interpret this will be given in section 4.

In the sequel, we will call this classifier the fuzzy integral classifier.

3.2 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 \( \phi^i_j \) have already been obtained by some parametric or non parametric classical probability density estimation method, after suitable normalization: possibilistic histograms, Parzen windows, Gaussian densities, etc. In this paper, possibilistic histograms have been used [?].

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 \(X^j_1, X^j_2, \ldots, X^j_{l_1}\) for class \(C_j\), \(j = 1, 2\). The criteria are the following.

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

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

(6)

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 \([2, 3]\)).

- the generalised 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 optimisation algorithms when the Choquet integral is used. In fact, this optimisation 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 recently looked towards heuristic algorithms better adapted to the peculiar structure of the problem and less greedy \([2]\). A satisfying algorithm has been found, which although suboptimal, reduces the computing time by a factor 200.

### 3.3 Performance

It has been proven that the quadratic and generalised quadratic criteria lead to the minimum mean-squared error approximation to the Bayes optimal classifier, within the class of all possible fuzzy integral classifiers \([2]\). This result shows that these criteria are suitable.

Numerous experiments have been done in order to test the fuzzy integral classifier on various data sets. We give here some of these results, in order to show its efficiency. Fuzzy integral classifiers trained with the quadratic criterion (QUAD), the generalized quadratic criterion minimized by the constrained least mean squares algorithm (CLMS) or the heuristic least mean square algorithm (HLMS)\([2]\) have been tested, together with more classical methods. Table 1 (left) gives results obtained on the famous iris data set (3 classes, 4 features, 50 data per class), and the highly non Gaussian cancer data set (2 classes, 9 features, 286 data). Results on classical methods have been borrowed from

<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 1: Classification rate on various data set

[2]. The estimation of the classification rate has been done by a 10-fold cross validation test. The results of the fuzzy integral classifier in the case of cancer data are remarkable.

We give another series of results on a simulated data set of 9000 patterns, 3 classes and 4 features, used inside Thomson-CSF for comparison tests. The data are non gaussian, with one attribute being the sum of two others (table 1 right). All these results show that a fuzzy integral classifier, if not always the best, gives in each case among the best results. It is a widely known fact that there exists no universal classification algorithm outperforming all the others on every data set.
4 Importance and interaction of features

As said before, the fuzzy measures \( \mu^j \) contain all the information about the importance of all individual features (or sources) and all groups of features 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 features, 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 \( i \in A \). Clearly, it is not enough to look solely at the value of \( \mu_i \), but also at \( \mu_{ij}, \mu_{ijk}, \) 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 \( \mu_1 \) 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 \( \mu_{12} \) 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 \( \{1, 2\} \subset A \).

Based on these ideas, it is possible to compute a global importance index and a interaction index. We define them in the next section.

4.1 Shapley value and interaction index of a fuzzy measure

**Definition 4** Let \( \mu \) be a fuzzy measure on \( X \). The importance index or Shapley value of element \( i \) with respect of \( \mu \) is defined by:

\[
v_i = \sum_{k=0}^{n-1} \gamma_k \sum_{K \subseteq X \setminus \{i\}, |K| = k} (\mu_{iK} - \mu_K)
\]

with \( \gamma_k = \frac{(n-k-1)!k!}{n!} = \frac{1}{C_k^{n-1}} |K| \) indicating the cardinal of \( K \), and \( 0! = 1 \) as usual.

This definition has been proposed by Shapley on an axiomatic basis in cooperative game theory [?], and possesses all suitable properties for representing importance of indexes. These values have the property that \( \sum_{i=1}^{n} v_i = 1 \). It is convenient to scale these values by a factor \( n \), so that an importance index greater than 1 indicates a feature more important than the average.

**Definition 5** The interaction index between two elements \( i \) and \( j \) with respect to a fuzzy measure \( \mu \) is defined by:

\[
I_{ij} = \sum_{k=0}^{n-2} \xi_k \sum_{K \subseteq X \setminus \{i,j\}, |K| = k} (\mu_{ijK} - \mu_{iK} - \mu_{jK} + \mu_K)
\]

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

It is easy to show that the maximum value of \( I_{ij} \) is 1, reached the fuzzy measure \( \mu \) defined by \( \mu_{ijK} = 1 \) for every \( K \subset X \), and 0 otherwise. Similarly, the minimum value of \( I_{ij} \) is -1, reached by \( \mu \) defined by \( \mu_{iK} = \mu_{jK} = 1 \) for any \( K \subset X \) and 0 otherwise. This definition has been proposed by Murofushi and Soneda [?], by using concepts of multiattribute utility theory. We will see later why this definition is perfectly coherent with the Shapley value. A positive (resp. negative) value of the index corresponds to a positive (resp. negative) synergy.

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.

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

Tables 2 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 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).

### 4.2 The inverse problem

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? 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\[\text{[?]}\], 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.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Class 1</th>
<th>Class 2</th>
<th>Class 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.759</td>
<td>0.670</td>
<td>0.416</td>
</tr>
<tr>
<td>2</td>
<td>0.875</td>
<td>0.804</td>
<td>0.368</td>
</tr>
<tr>
<td>3</td>
<td>1.190</td>
<td>1.481</td>
<td>1.377</td>
</tr>
<tr>
<td>4</td>
<td>1.176</td>
<td>1.045</td>
<td>1.839</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Features</th>
<th>Class 1</th>
<th>Class 2</th>
<th>Class 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,2</td>
<td>0.128</td>
<td>-0.159</td>
<td>-0.065</td>
</tr>
<tr>
<td>1,3</td>
<td>0.051</td>
<td>0.281</td>
<td>0.052</td>
</tr>
<tr>
<td>1,4</td>
<td>0.054</td>
<td>-0.257</td>
<td>0.010</td>
</tr>
<tr>
<td>2,3</td>
<td>-0.009</td>
<td>0.114</td>
<td>0.002</td>
</tr>
<tr>
<td>2,4</td>
<td>-0.007</td>
<td>0.036</td>
<td>0.059</td>
</tr>
<tr>
<td>3,4</td>
<td>-0.051</td>
<td>0.132</td>
<td>-0.238</td>
</tr>
</tbody>
</table>

Table 2: Indexes of importance and interaction for the iris data set
We will solve completely the above inverse problem in the case of 2-additive fuzzy measures. For the sake of brevity, we will omit proofs. The following result can be shown [?].

**Theorem 1** Let $v_1, \ldots, v_n$ be a set of Shapley values, satisfying $\sum_{i=1}^{n} v_i = 1$, and $I_{ij}, \{i, j\} \subseteq X$ a set of interaction indexes. There exists a unique 2-additive fuzzy measure (eventually non monotonic) whose equivalent pseudo-Boolean function $f(x) = \sum_{i=1}^{n} a_i + \sum_{\{i,j\} \subset X} a_{ij}$ is defined by:

$$a_i = v_i - \frac{1}{2} \sum_{j \in X \setminus i} I_{ij}, \quad i = 1, \ldots, n$$

$$a_{ij} = I_{ij}, \quad \{i, j\} \subseteq X$$

Remark that $a_{ij}$ coincides with $I_{ij}$, similarly to $a_i$ which coincides with $v_i$ in the first order case. Clearly, there exists an infinity of $k$-additive fuzzy measures, $k > 2$, whose Shapley values and interaction indexes are the same. The unicity is ensured only for $k = 2$. It can be said that this 2-additive fuzzy measure is the least specific regarded 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 $v_i$ and $I_{ij}$ must verify a set of constraints. We can show the following [?].

**Theorem 2** A set of Shapley values $v_1, \ldots, v_n$ and interaction indexes $I_{ij}, \{i, j\} \subseteq X$ lead to a monotonic 2-additive fuzzy measure if and only if they satisfy the following set of constraints:

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

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 3 the following set of importance and interaction index (this set satisfies the above constraints). It can be seen that very simple values have been given, setting $v_i$ 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 4. We have used these identified 2-additive fuzzy measures for recognition. The results are shown on table 5. 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.

---

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

<table>
<thead>
<tr>
<th></th>
<th>class 1</th>
<th>class 2</th>
<th>class 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$v_{11}$</td>
<td>2.129</td>
<td>1.873</td>
<td>2.198</td>
</tr>
<tr>
<td>$v_{21}$</td>
<td>0.583</td>
<td>0.421</td>
<td>0.298</td>
</tr>
<tr>
<td>$v_{12}$</td>
<td>0.035</td>
<td>0.167</td>
<td>0.124</td>
</tr>
</tbody>
</table>

**Figure 3:** The iris data set after PCA, and the indexes of importance and interaction.
Table 3: Set of scaled importance and interaction indexes for the iris data set

<table>
<thead>
<tr>
<th></th>
<th>class 1</th>
<th>class 2</th>
<th>class 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n \cdot v_1$</td>
<td>0.4</td>
<td>0.4</td>
<td>0.4</td>
</tr>
<tr>
<td>$n \cdot v_2$</td>
<td>0.4</td>
<td>0.4</td>
<td>0.4</td>
</tr>
<tr>
<td>$n \cdot v_3$</td>
<td>1.6</td>
<td>1.4</td>
<td>1.4</td>
</tr>
<tr>
<td>$n \cdot v_4$</td>
<td>1.6</td>
<td>1.8</td>
<td>1.8</td>
</tr>
<tr>
<td>$I_{12}$</td>
<td>0.1</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>$I_{13}$</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>$I_{14}$</td>
<td>0.0</td>
<td>-0.2</td>
<td>-0.2</td>
</tr>
<tr>
<td>$I_{23}$</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>$I_{24}$</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>$I_{34}$</td>
<td>0.0</td>
<td>0.45</td>
<td>0.45</td>
</tr>
</tbody>
</table>

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 lead us to the conclusion that the concepts of importance and interaction indexes presented here are meaningful and useful in applications.