ELSEVIER

Fuzzy Sets and Systems 107 (1999) 197–218

FUZZY sets and systems

www.elsevier.com/locate/fss

Improved feature selection and classification by the 2-additive fuzzy measure

L. Mikenina *, H.-J. Zimmermann

Lehrstuhl für Unternehmensforschung, RWTH Aachen, Templergraben 64, 52062 Aachen, Germany

Received May 1998; received in revised form October 1998

Abstract

This paper focusses on the investigation of a pattern recognition method based on the fuzzy integral. Until now this method has used a general fuzzy measure, which is characterized by exponential complexity. Naturally this led to some difficulties in practical applications of this pattern recognition method. In this paper, a heuristic algorithm for the identification of the 2-additive fuzzy measure, which is a particular type of k-additive fuzzy measures, is proposed. This algorithm can be used to reduce complexity of feature selection and classifier design. A further topic considered in this paper is the development of a feature selection algorithm for the fuzzy integral classifier. The proposed heuristic algorithm is based on two feature-evaluation criteria such as the importance and the interaction indexes. They were earlier defined in the literature using the semantic interpretation of the fuzzy measure. To validate the proposed algorithms, the feature selection algorithm and the pattern recognition method based on the fuzzy integral are applied to a problem of acoustic quality control.

Keywords: Pattern recognition; Feature selection; Fuzzy measure and integral theory

1. Introduction

Fuzzy pattern recognition presents one of the largest application areas of fuzzy set theory. The primary advantage of fuzzy methods compared to classical methods is the ability of a system to classify patterns in a non-dichotomous way as it is done by humans and to handle vague information [39,40].

Fuzzy pattern matching techniques represent a group of fuzzy methods for supervised pattern recognition. The most general framework among these techniques was introduced in [13] as a pattern recognition method based on the fuzzy integral. Its mathematical background is fuzzy measure and integral theory, which was proposed by Sugeno [31]. Fuzzy measures represent a generalization of classical measures and are obtained by replacing the additivity property with a weaker requirement of monotonicity. Thus they are often referred to as non-additive measures [37]. Fuzzy integrals are considered as averaging operators and, compared to probability theory, they correspond to non-additive expected values [19,37]. Although a lot of theoretical research has been done in this field [4,16,27,28,32–34,36,38], the number of practical applications is still moderate. The main application areas of fuzzy integral theory cover multiattribute decision theory and pattern recognition, where a fuzzy

* Corresponding author. Tel.: +49 241 806182; fax: +49 241 8888168.

0165-0114/99/$ – see front matter
PII: S0165-0114(98)00429-1

measure is used for modeling the importance of a group of elements (criteria or features) and the fuzzy integral for aggregating partial evaluations.

A large advantage of using the fuzzy integral within a pattern recognition method is due to the unique behavioral property of the fuzzy integral. This is the only weighted aggregation operator, which takes into account not only the importance of elements, but also the importance of all subsets of them. The weights are represented by the coefficients of the fuzzy measure. In the context of pattern recognition, these coefficients express the importance of each feature and the interaction between features. The importance of features is considered with respect to classes and is related to a discrimination ability of a feature for a class. The interaction between features expresses the contribution of a subset of features to the recognition process. Information about the importance and the interaction between features is used for feature selection as well as for classification.

Despite unique modeling properties of the fuzzy measure and integral, difficulties in practical application of this pattern recognition method arise due to the exponential complexity of the identification of fuzzy measures. This is the central problem in the design procedure of the fuzzy integral classifier. One possible solution is provided by the concept of k-additive fuzzy measures [12], which can range between additive and general fuzzy measures. In this paper, a particular case of a 2-additive fuzzy measure is considered, which is sufficient for the semantic interpretation of the fuzzy measure. The main advantage of using the 2-additive fuzzy measure instead of a general one is that the pattern recognition method based on the fuzzy integral can be applied to problems described by a large number of features.

Since the performance of a classifier depends to a large degree on the quality of features used, the development of an efficient feature selection algorithm is of primary importance. Its objective is to select the smallest set of features, which is sufficient for recognizing correctly classes of objects [5,20,24,35]. In this paper, two feature-evaluation criteria based on the semantic interpretation of the fuzzy measure [26,30] are discussed and a heuristic algorithm for feature selection using these two criteria is proposed. The feature selection procedure depends on the fuzzy integral classifier and thus the 2-additive fuzzy measure seems to be much more suitable for real applications than a general fuzzy measure.

This paper is structured as follows:

In Section 2, basic definitions concerning k-additive fuzzy measures are presented and a heuristic algorithm for the identification of the 2-additive fuzzy measure is proposed. In Section 3, two feature-evaluation criteria based on the fuzzy measure are described and a heuristic algorithm for feature selection is proposed. In Section 4, a general description of a pattern recognition method based on the fuzzy integral is given, and the use of the proposed identification algorithm within the classifier design is considered in more detail. Section 5 presents an application of the feature selection algorithm and a pattern recognition method based on the fuzzy integral to automatic bearing diagnosis.

2. k-order additive fuzzy measures and an algorithm for the identification of 2-additive fuzzy measures

Consider a finite set of elements $X = \{1, \ldots, n\}$.

Definition 2.1 (Grabisch [12]). A discrete fuzzy measure on X is a set function $\mu : \mathcal{P}(X) \to [0, 1]$ satisfying

1. $\mu(\emptyset) = 0$, $\mu(X) = 1$.
2. $A \subset B$ imply $\mu(A) \leqslant \mu(B)$ for $A \in \mathcal{P}(X)$, $B \in \mathcal{P}(X)$ (monotonicity).

In complexity theory, pseudo-Boolean functions are often used to represent a set function. This idea can be applied to represent the fuzzy measure, which is characterized by exponential complexity.

Definition 2.2 (Hammer and Holzman [15]). A pseudo-Boolean function is a real-valued function $f : \{0, 1\}^n \to \Re$.

It can be shown that any pseudo-Boolean function can be expressed in the form of a multilinear polynomial in n variables:

$$f(x) = \sum_{T \subset X} \left[a(T) \prod_{i \in T} x_i \right] \tag{2.1}$$

with $a(T) \in \Re$ and $\boldsymbol{x} = (x_1, \ldots, x_n) \in \{0, 1\}^n$.

It can be seen that a fuzzy measure is a particular case of the pseudo-Boolean function, defined for any $A \subset X$ such that A is equivalent to a point

$x=(x_1,\ldots,x_n)$ in $\{0,1\}^n$ where $x_i=1$ if and only if $i\in A$.

The coefficients $a(T)$, $T\subset X$ can be viewed as a set function, which in fact corresponds to the Möbius transform. Denote by μ any set function $\mu:\mathcal{P}(X)\to\Re$.

The Möbius transform of μ is a set function a on X defined by [29]

$$a(T)=\sum_{K\subset T}(-1)^{|T\setminus K|}\mu(K),\quad \forall T\subset X. \tag{2.2}$$

This transformation is inversible. When a is given, it is possible to recover the original μ by the so-called Zeta-transform:

$$\mu(T)=\sum_{S\subset T}a(S),\quad \forall T\subset X. \tag{2.3}$$

Consider the case of additive measures. According to (2.1), additive measures have a linear representation

$$f(x)=\sum_{i=1}^{n}a_ix_i,$$

where $\mu_i\equiv a_i$ and the notations $\mu_i=\mu(\{i\}), a_i=a(\{i\})$ are used. By extension, fuzzy measures having a polynomial representation of degree 2, or 3, or any fixed integer k can be defined. These fuzzy measures are called k-order additive or simply k-additive measures.

Definition 2.2 (Grabisch [12]). A fuzzy measure μ defined on X is said to be k-order additive if its corresponding pseudo-Boolean function is a multilinear polynomial of degree k, i.e. its Möbius transform $a(T)=0$ for all T such that $|T|>k$, and there exist at least one subset T of X of exactly k elements such that $a(T)\neq 0$.

This definition is illustrated by an example of the 2-additive fuzzy measure.

Example 2.1. The 2-additive fuzzy measure is defined by

$$\mu(K)=\sum_{i=1}^{n}a_ix_i+\sum_{\{i,j\}\subseteq X}a_{ij}x_ix_j \tag{2.4}$$

for any $K\subseteq X$, $|K|\geqslant 2$ with $x_i=1$ if $i\in K$, $x_i=0$ otherwise.

Using the fact that $\mu_i=a_i$ for all i, the following expression is obtained:

$$\mu_{ij}=a_i+a_j+a_{ij}=\mu_i+\mu_j+a_{ij}.$$

The general formula for the 2-additive fuzzy measure is [12]

$$\begin{aligned}\mu(K)&=\sum_{i\in K}a_i+\sum_{\{i,j\}\subseteq K}a_{ij}\\&=\sum_{\{i,j\}\subseteq K}\mu_{ij}-(|K|-2)\sum_{i\in K}\mu_i\end{aligned} \tag{2.5}$$

for any $K\subseteq X$ such that $|K|\geqslant 2$. It is clear that the 2-additive fuzzy measure is determined by the coefficients μ_i and μ_{ij}.

According to its definition, k-additive fuzzy measures for $k<n$ need less than 2^n coefficients to be defined. It can be shown that only n coefficients are required for $k=1$, $n(n+1)/2$ coefficients for $k=2$, and in general $\sum_{j=1}^{k}\binom{n}{j}$ for k-additive measures.

For a one-to-one correspondence between the Möbius representation and the fuzzy measure satisfying monotonicity, the values $a(T)$ in (2.1) must obey some constraints. They are formulated in the following theorem.

Theorem 2.1 (Chateauneuf and Jaffray [3]). *A set of 2^n coefficients $a(T), T\subset X$ corresponds to the Möbius representation of a fuzzy measure if and only if*

1. $a(\emptyset)=0$, $\sum_{T\subset X}a(T)=1$;
2. $\sum_{i\in B\subset T}a(B)\geqslant 0$ *for all* $T\subset X$, *for all* $i\in T$.

The concept of k-additive fuzzy measures provides a tradeoff between richness and complexity of fuzzy measures.

In the following, a particular case of a 2-additive fuzzy measure is considered. To identify this special type of fuzzy measures, only $n(n+1)/2$ coefficients μ_i and μ_{ij}, $i,j\in X$, have to be learned from training data. The coefficients for all other subsets $K\subseteq X$, $|K|>2$ are calculated from μ_i and μ_{ij}. In order to obtain the monotone fuzzy measure, the coefficients μ_i and μ_{ij} must satisfy particular conditions. Such conditions were formulated for the Möbius representation of a fuzzy measure in Theorem 2.1. They can also be expressed through the original representation of the fuzzy measure using (2.2).

For 2-additive fuzzy measures, coefficients of the Möbius representation are given by

$$a_i = \mu_i, \quad \forall i \in X,$$

$$a_{ij} = \mu_{ij} - \mu_i - \mu_j, \quad \forall \{i,j\} \subseteq X.$$

Then the monotonicity constraints on the coefficients of the 2-additive fuzzy measure can be derived from part 2 of Theorem 2.1 and are formulated as follows:

$$\sum_{j \in K} \mu_{ij} - \sum_{j \in K} \mu_j - (n-2)\mu_i \geqslant 0, \quad \forall i \in X,\ K \subseteq X \backslash i, \tag{2.6}$$

where $|X| = n$.

In order to obtain the fuzzy measure, which is normalized on the interval $[0,1]$, the coefficients μ_i and μ_{ij} must also satisfy the normalization condition. This can be formulated using (2.5) for $K = X$:

$$\sum_{\{i,j\} \subseteq X} \mu_{ij} - (n-2) \sum_{i \in X} \mu_i = 1. \tag{2.7}$$

For example, the monotonicity and normalization constraints for $n = 4$ are defined as

$$(i=1) \quad \mu_{12} + \mu_{13} + \mu_{14} - 2\mu_1 - \mu_2 - \mu_3 - \mu_4 \geqslant 0,$$
$$(i=2) \quad \mu_{12} + \mu_{23} + \mu_{24} - \mu_1 - 2\mu_2 - \mu_3 - \mu_4 \geqslant 0,$$
$$(i=3) \quad \mu_{13} + \mu_{23} + \mu_{34} - \mu_1 - \mu_2 - 2\mu_3 - \mu_4 \geqslant 0,$$
$$(i=4) \quad \mu_{14} + \mu_{24} + \mu_{34} - \mu_1 - \mu_2 - \mu_3 - 2\mu_4 \geqslant 0,$$
$$\mu_{12} + \mu_{13} + \mu_{14} + \mu_{23} + \mu_{24} + \mu_{34} - 2(\mu_1 + \mu_2 + \mu_3 + \mu_4) = 1.$$

Clearly, these relations are not as simple as the ones for a general fuzzy measure ($\mu_i \leqslant \mu_{ij}$), but using the lattice representation of the fuzzy measure [9] their formulation can be well summarized. The lattice of the 2^n coefficients of the fuzzy measure is equivalent to the lattice of elements of the power set with respect to set inclusion relations. An example of the lattice for the case $n = 4$ is shown in Fig. 1.

Nodes of the lattice represent subsets of the power set or fuzzy measure coefficients, and links of the lattice represent order relations such as inclusion for subsets or ordinary '$\leqslant$' for fuzzy measure coefficients. A set of chained links, starting from layer 0 (empty set) and arriving to layer n (whole set X) is called a path [9]. The path emphasized in the figure corresponds to a datum such that $x_3 < x_2 < x_4 < x_1$. The coefficients $\mu(\{1\})$, $\mu(\{1,4\})$ and $\mu(\{1,2,4\})$ are involved in the calculation of the fuzzy integral (the Sugeno or the Choquet integral [37]) of the given datum.

For a given node in layer k, the set of nodes in the layer $k-1$ (respectively $k+1$) linked to it is called its lower neighbors (respectively upper neighbors). A state, where each node of layer k is equidistant from any node of layer $k+1$ or $k-1$ is called the equilibrium state of the fuzzy measure.

Using the lattice representation of the fuzzy measure, relations (2.6) can easily be formulated. To verify the monotonicity relation for the node μ_i, its upper neighbors are determined and their values are added. Then the values of the nodes of the first layer are subtracted from the sum, where the value of μ_i is taken $(n-2)$ times. It can be noticed that each node of the first layer is involved in all monotonicity relations, thus n relations must be verified. For each node μ_{ij}, two relations must be checked. To formulate them, lower neighbors μ_i and μ_j of the node μ_{ij} are determined, and the monotonicity is verified for these two nodes as described above. If one of the monotonicity relations is violated, then the expression (2.6) is negative and its value is denoted by $-\Delta c_k$, $k = 1, \ldots, n$. The value of the node μ_i or μ_{ij} is corrected using the maximal degree of violation among all verified relations.

Finally, if coefficients μ_i and μ_{ij} are defined such that all constraints (2.6) and (2.7) are satisfied, the induced 2-additive fuzzy measure is monotone and normalized.

A heuristic algorithm for the identification of the 2-additive fuzzy measure (Algorithm 1) is based on this idea. Two main steps of the algorithm can be summarized as follows:

1. The training data are used to identify coefficients μ_i and μ_{ij} under the monotonicity and normalization constraints. All other coefficients $\mu(K)$, $|K| > 2$, are calculated from μ_i and μ_{ij}.
2. Only the nodes of the first and the second layers left unmodified in step 1 are shifted in the lattice as close as possible to the equilibrium state. The nodes' values are adjusted to keep the monotonicity and normalization constraints satisfied. For example, if a value of one node of the first layer is decreased (increased), then the values of all nodes of the second layer are equally decreased (increased) to some degree. After the adjustment of nodes μ_i and μ_{ij}, other coefficients $\mu(K)$, $|K| > 2$, must be calculated anew as in step 1.

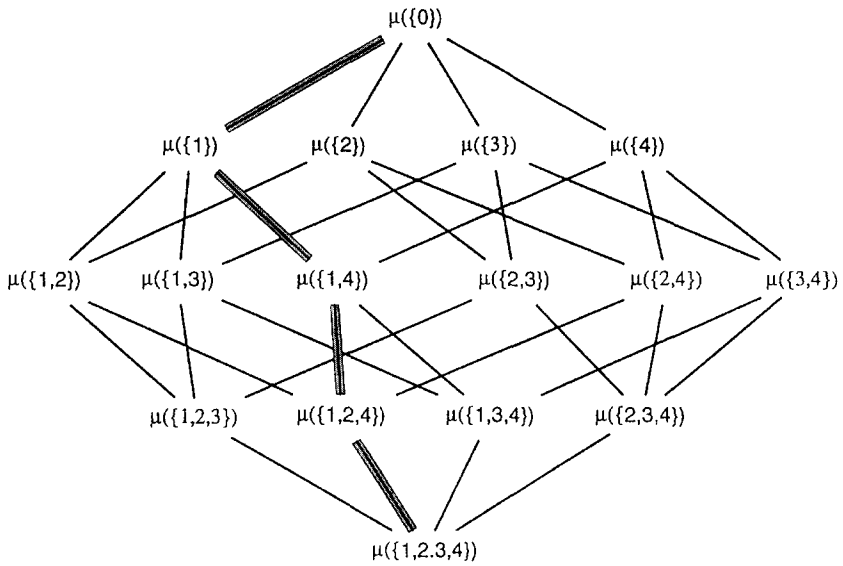

Fig. 1. The lattice of fuzzy measure coefficients for $n = 4$.

It should be noted that it is not necessary to verify the monotonicity for nodes left unmodified in step 1 (as it was done in the heuristic least mean-squares algorithm (HLMS) presented in [9]), since this is guaranteed, if at least one node of the first layer was modified.

Before presenting an identification algorithm, the concept of the fuzzy integral should be briefly introduced. Fuzzy integrals are non-linear functionals similar to Lebesgue integral and representing a particular case of averaging operators. Compared to probability theory, they correspond to non-additive expected values [19,37]. The most well-known fuzzy integrals are Sugeno [31] and Choquet [33] integrals. Since the proposed Algorithm 1 uses the Choquet integral, the definition of this type of the integral is considered.

Definition 2.4 (Sugeno and Murofushi [33]). Let $\mu: F \to [0,1]$ be a fuzzy measure on a measurable space (X,F) and $f: X \to [0,\infty)$ a measurable function. The Choquet integral of f with respect to μ is defined by

$$C_\mu(f) = (C)\int f \circ \mu = \int_0^\infty \mu(H_\alpha)\,\mathrm{d}\alpha,$$

where $H_\alpha = \{x \in X \mid f(x) \geqslant \alpha\}$ $\forall \alpha \in [0,1]$.

Suppose that the function $f(x)$ is discrete on $X = [x_1,\ldots,x_n]$ and denote the value of a function at a point $x_i \in X$ by f_i. Consider a permutation of the function values in increasing order denoted by $f_{(1)},\ldots,f_{(n)}$ and denote $A_{(i)} = \{x_{(i)}, x_{(i+1)},\ldots,x_{(n)}\}$. Then the Choquet integral can be written as follows:

$$C_\mu(f) = \sum_{i=1}^{n} (f_{(i)} - f_{(i-1)})\mu(A_{(i)}).$$

A new algorithm for the identification of the 2-additive fuzzy measure, which uses some basic steps of the HLMS algorithm, is called HLMS (2-add) and is presented below.

Algorithm 1.

Step 0: *Initialize the fuzzy measure in the equilibrium state.*

Step 1.1: *Consider a training datum* $(\boldsymbol{x}, y)$. *Compute the error between the actual and the expected output:* $E = \mathscr{F}_\mu(\boldsymbol{x}) - y$, *where* $\mathscr{F}_\mu(\boldsymbol{x}) = C_\mu(\boldsymbol{x})$ *is the Choquet integral. Denote the values of the nodes (fuzzy measure coefficients) on the path involved by* $\boldsymbol{x}$ *by* $u(0), u(1),\ldots,u(n)$, *where* $u(0) = 0$ *and* $u(n) = 1$.

Step 1.2: *Compute a new value* $u^{\mathrm{new}}(i)$, $i = 1,2$, *of the considered node as in the gradient descent*

method [9]:

$$u^{\text{new}}(i) = u^{\text{old}}(i) - \alpha \frac{E}{E_{\max}}(x_{(n-i)} - x_{(n-i-1)}), \quad (2.8)$$

where $u(1) = \mu_j$ *and* $u(2) = \mu_{jh}$, $j, h = 1, \ldots, n$, $j \neq h$; $\alpha \in [0,1]$ *is a constant, or it can be decreasing at each iteration;* $E_{\max}$ *is the maximum value of the error. If* $y \in [0,1]$, *then* $E_{\max} = 1$. *The notation* $x_{(i)}$ *indicates the ith element of the vector* $\mathbf{x}$ *in ascending order.*

Step 1.3: *For every node* $u(i)$, $i = 1, 2$, *verify the monotonicity relations. If* $E < 0$, *the value of* $u(i)$ *is increased and n monotonicity relations of the form* (2.6) *are checked for* $u(1)$. *If the monotonicity of the kth relation is violated at the value* Δc_k, $k = 1, \ldots, n$, *then the maximum degree of violation among all relations is chosen to correct the value of* $u(1)$:

$$u(1) = u(1) - \max_{\substack{k=1,\ldots,n \\ k \neq j}} \left(\frac{\Delta c_j}{n-2}, \Delta c_k \right), \quad (2.9)$$

where j is the index of the considered fuzzy measure coefficient $u(1) = \mu_j$. *Since this coefficient is taken with factor* $(n-2)$ *in the jth relation, the degree of violation* Δc_k *have to be divided by this factor. Analogously, if* $E > 0$, *two monotonicity relations are checked for* $u(1)$ *and* $u(2)$. *For* $u(1)$, *they are just non-negativity conditions. For* $u(2)$, *these relations can be violated at values* Δc_k, $k = 1, 2$. *Then the value of* $u(2)$ *is corrected by using the maximum violation degree*:

$$u(2) = u(2) + \max(\Delta c_1, \Delta c_2). \quad (2.10)$$

Nodes $u(i)$, $i = 1, 2$, *are considered in steps* 1.2 *and* 1.3 *in the following order*:

- *if* $E < 0$, *begin by* $u(2), u(1)$;
- *if* $E > 0$, *begin by* $u(1), u(2)$.

Step 1.4: *Compute a new value* $u^{\text{new}}(i)$, $i = 3, \ldots, n-1$, *of the node* $\mu(K)$ *using*

$$u^{\text{new}}(i) = \mu(K) = \sum_{\{j,h\} \subset K} \mu_{jh} - (|K| - 2) \sum_{j \in K} \mu_j \quad (2.11)$$

for any $k \subset X$ *such that* $|K| > 2$.

During one iteration, Steps 1.1–1.4 *are repeated for all training data. For a convergence of the algorithm several iterations should be carried out.*

Step 2.1: *Adjust the value of each node of layers* 1 *and* 2 *left unmodified in Step* 1. *Begin from the lower level and denote the node considered by* $v(i)$. *To arrange node* $v(i)$ *into the lattice as close as possible to the equilibrium state, the following parameters are computed* [9]:

- *the mean value of lower neighbors*:

$$\underline{m}(i) = \frac{\sum_{A \subset LN} \mu(A)}{i},$$

where LN is a set of lower neighbors of node $v(i)$;

- *the mean value of upper neighbors*:

$$\overline{m}(i) = \frac{\sum_{A \subset UN} \mu(A)}{n - i},$$

where UN is a set of upper neighbors of node $v(i)$;

- *the minimum distance between node* $v(i)$ *and its lower neighbors*:

$$\underline{d}_{\min}(i) = \min_{A \subset LN} [v(i) - \mu(A)];$$

- *the minimum distance between node* $v(i)$ *and its upper neighbors*:

$$\bar{d}_{\min}(i) = \min_{A \subset UN} [\mu(A) - v(i)].$$

If node $v(i)$ *is closer to its lower than to its upper neighbors, that is,* $\overline{m}(i) + \underline{m}(i) - 2v(i) > 0$, *the value of* $v(i)$ *should be increased*:

$$v^{\text{new}}(i) = v^{\text{old}}(i) + \beta \frac{(\overline{m}(i) + \underline{m}(i) - 2v(i))\bar{d}_{\min}(i)}{2(\overline{m}(i) + \underline{m}(i))} \quad (2.12)$$

otherwise, if node $v(i)$ *is closer to its upper than to its lower neighbors, its value is decreased*:

$$v^{\text{new}}(i) = v^{\text{old}}(i) + \beta \frac{(\overline{m}(i) + \underline{m}(i) - 2v(i))\underline{d}_{\min}(i)}{2(\overline{m}(i) + \underline{m}(i))}, \quad (2.13)$$

where $\beta \in [0,1]$ *is a constant, or it can be decreasing at each iteration.*

To keep the monotonicity and normalization constraints satisfied, all other nodes of layers 1 *and* 2 *must be corrected. Denote by* Δb_1 *the degree, at which the value of node* $v(i)$ *should be changed. Four cases can be distinguished*:

- *If* $v(i)$ *is a node in layer* 1 *and increased by* Δb_1 *(the second term in* (2.12)*), then its upper*

neighbors must be increased by the same degree and all other nodes of layer 2 *must be decreased by*

$$\Delta b_2 = \frac{2}{(n-1)(n-2)}\Delta b_1;$$

- *If* $v(i)$ *is a node in layer* 1 *and decreased by* Δb_1, *then all nodes of layer* 2 *must be decreased by*

$$\Delta b_2 = \frac{2(n-2)}{n(n-1)}\Delta b_1;$$

- *If* $v(i)$ *is a node in layer* 2 *and increased by* Δb_1 (*the second term of* (2.13)), *then all nodes of layer* 1 *must be increased by*

$$\Delta b_2 = \frac{1}{n(n-2)}\Delta b_1;$$

- *If* $v(i)$ *is a node in layer* 2 *and decreased by* Δb_1, *then its lower neighbors must be decreased by*

$$\Delta b_2 = \frac{1}{(n-2)}\Delta b_1$$

and all other nodes of layer 1 must be increased by

$$\Delta b_2 = \frac{1}{(n-2)^2}\Delta b_1.$$

Step 2.2: *Calculate a new value of every node* $v(i)$, $i = 3,\ldots,n-1$, *using* (2.11). *Begin from lower levels.*

During one iteration, steps 2.1 *and* 2.2 *are repeated for all nodes left unmodified in the first step. Several iterations can be carried out.*

It can be observed that the adjustment of nodes in step 2 influences all nodes of the lattice. The lower and upper neighbors of node $v(i)$ are shifted together with node $v(i)$, but since the changes in their values are much smaller than the change of the value of $v(i)$ themselves, the procedure leads to a more homogeneous lattice, as supposed.

The HLMS (2-add) algorithm for the identification of the 2-additive fuzzy measure can be used within a pattern recognition method based on the fuzzy integral and its efficiency is crucial for the performance of the method. In the next sections, two main components of a pattern recognition system are considered in more detail: feature selection and classifier design. It should be noted that the proposed feature selection procedure is closely related to the fuzzy integral classifier.

3. Development of a feature-selection algorithm based on the fuzzy measure

In [12] it was stated that the fuzzy measure defined on a set X of elements can express the importance of any subset of elements. In the context of pattern recognition, where X is a set of features, the coefficients of the fuzzy measure can be used to evaluate the importance of each feature and the interaction between features. The importance of features is considered with respect to classes and is related to a discrimination ability of a feature for a class. The interaction between features expresses the contribution of a subset of features to the recognition process. These characteristics are modeled by the fuzzy measure as follows [11]:

- A feature i is important if the values of $\mu(K)$ are large for all subsets K containing i. Thus, not only the importance $\mu(\{i\})$ of the feature i is taken into account, but also the contribution of the feature in all subsets of features.

For pairs of features three cases can be distinguished:

- The importance of a pair of features i and j is almost the same as the individual importance of each feature, that is, $\mu(\{i,j\})$ is smaller than the sum of $\mu(\{i\})$ and $\mu(\{j\})$. This kind of interaction is called redundancy or negative synergy.
- The importance of a pair of features i and j is large, although these features are unimportant if they are considered separately, that is, the values of $\mu(\{i\})$ and $\mu(\{j\})$ are small whereas the value of $\mu(\{i,j\})$ is large. This kind of interaction is called complementarity or positive synergy.
- Features i and j have equal contributions to the recognition process without interfering, if the importance $\mu(\{i,j\})$ of a pair of features i and j is equal to the sum of the individual importances $\mu(\{i\})$ and $\mu(\{j\})$ of features. This is called independence.

Defining the interaction between features, all subsets K containing a pair i,j must be considered. Using this interpretation of the fuzzy measure, the global evaluation of the importance of features and the interaction between them is possible.

In [25] it was proposed to evaluate the importance of an element in X in analogy to n-person game theory. In game theory, a subset of players with the same goals is described as a coalition. A contribution of

each player of a coalition to the game is defined by the Shapley value [30]. It can be interpreted as a degree of importance of a player in the coalition. This idea can be used to determine the importance of an arbitrary element in the set X taking into account its importance in different subsets.

Definition 3.1 (Shapley [30]). The importance index or Shapley value v_i of element i with respect to a fuzzy measure μ is defined by

$$v_i = \sum_{k=0}^{n-1} \gamma_k \sum_{K \subset X \setminus i, |K|=k} [\mu(K \cup \{i\}) - \mu(K)], \quad (3.1)$$

where $\gamma_k = (n-k-1)!k!/n!$ and X is a set of n elements.

The Shapley value with respect to the fuzzy measure μ is a vector $v(\mu) = [v_1, \ldots, v_n]$. It expresses the global importance of each feature, taking into account the effect of adding element i to a subset K (which does not contain i) in terms of strengthening the subset K. The Shapley value has the property that the sum of all its components is equal to 1, thus it represents the sharing of the total importance of all features among them. It is convenient to scale these values by a factor n, in order to highlight the features which are more important than the average. Their importance indexes become then, greater than 1.

Analogously, the effect of adding two elements to a subset K can be calculated. The definition of the interaction of two elements was proposed in [26] using concepts from multiattribute utility theory [18] and is based on the following idea. Two elements i and j interact in a positive (cooperative) way, if adding i and j together to a subset K is more valuable than putting the individual values of i and j to the subset, that is,

$$\mu(K \cup \{i,j\}) - \mu(K) \geqslant \mu(K \cup \{i\}) - \mu(K) + \mu(K \cup \{j\}) - \mu(K).$$

In the case of $\leqslant$ relation, the elements interact in a negative (substitutive) way. Taking the average of all possible coalitions, the average value of interaction is obtained.

Definition 3.2 (Murofushi and Soneda [26]). The interaction index between two elements i and j with respect to a fuzzy measure μ is defined by

$$I_{ij} = \sum_{k=0}^{n-2} \xi_k \sum_{K \subset X \setminus \{i,j\}, |K|=k} [\mu(K \cup \{i,j\}) - \mu(K \cup \{i\}) - \mu(K \cup \{j\}) + \mu(K)], \quad (3.2)$$

where $\xi_k = (n-k-2)!k!/(n-1)!$.

The generalization of the interaction index for any subset A of elements was proposed in [10]:

$$I(A) = \sum_{B \subset X \setminus A} \xi(B,A) \sum_{C \subset A} (-1)^{|A \setminus C|} \mu(C \cup B), \quad (3.3)$$

where $\xi(B,A) = (n - |B| - |A|)!|B|!/(n - |A| + 1)!$.

The value of $I(A)$ can be interpreted in the same way as I_{ij}, but it is more difficult to express its meaning directly. Thus, for a semantic analysis, the considerations are restricted to v_i and I_{ij}.

Using the importance and interaction indexes, a set of features can be described as follows:

- Feature i is more important than feature j for distinguishing class C_k from the others, if the Shapley values relate to each other as $v_i > v_j$.
- Features i and j are redundant for distinguishing class C_k from the others, if the value of I_{ij} is negative. It is sufficient to use one of the two features.
- Features i and j are complementary for distinguishing class C_k from the others, if the value of I_{ij} is positive. The combination of the two features must be used.
- Features i and j are independent for distinguishing class C_k from the others, if the value of I_{ij} is zero. Both features bring their contribution.

Hence, the importance and interaction indexes can be used as two evaluation measures for features. They provide information about the importance of single features for the classification as well as about the pairwise dependencies between features in the sense, whether two features taken together are complementary or redundant for the classification. Based on the interpretation of these two measures, the selection of a subset of the most appropriate features from a set of given features can be performed. Developing a method for feature selection, two points should be noticed:

- the evaluation measures are classifier dependent, i.e. for their computation the classification of training data must be performed to identify the fuzzy measure for each class;
- a search procedure for selecting the best subset of features should be defined based on the interpretation of both evaluation measures.

The first point can be considered as a restriction on the application area of the algorithm due to its dependence on a special classification method and its computational complexity. The second point emphasizes that existing search procedures cannot be applied, since they use other evaluation criteria.

The proposed method for feature selection consists of the following three steps:

1. identification of the fuzzy measure for each class within the classifier design;
2. computation of the importance index for all features and the interaction index for all pairs of features for each class;
3. selection of the best features based on the analysis of computed values (Algorithm 2).

The first step is in fact the procedure of classifier design, which will be described in the next section. Since it can be a very complicated task for a large number of features, it is reasonable to use 2-additive fuzzy measures, which are characterized by quadratic complexity.

Computation of the importance and the interaction indexes succeeds according to (3.1) and (3.2). The results of this step are two tables containing indexes for all classes.

The third step should be considered in more detail. There are in general two strategies for selecting a best set of features. The first one starts with one or two best features and increases its number by evaluating the criterion for different combinations of features. Another strategy starts with the whole set of features and eliminates the worst features based on the evaluation measure. The heuristic algorithm proposed in this section combines these two strategies. It should be noted that the evaluation of subsets is restricted to single features and pairs of features. Combinations of more than two features are not evaluated. Further, the selection of the best feature subset is performed for each class and the resulting features are put together. As a result the feature, which is important just for one class is chosen, although it is maybe not important for other classes. Alternative approach would be to calculate the importance and interaction indexes for each class and aggregate their values over classes by some rule. However, this approach cannot succeed, since it can happen that a large positive value of the interaction index for one class is compensated by negative values for other classes and, consequently, this pair, which is very important for one particular class, will not be selected. Therefore, features are selected with respect to each class.

In general, it is possible to select the desired subset of n features during one iteration. But it seems reasonable to do this in several iterations, in order to obtain reliable evaluations. Thus, for each iteration the number of features to be selected is predefined (empirically or according to some rules) such that the initial number n' is decreased after p iterations to the number n: $n' > n_1 > n_2 > \cdots > n_p = n$.

The following heuristic algorithm for selection of the best feature subset is proposed:

Algorithm 2.

Step 1: *Sorting of the values in descending order in the table of interaction indexes; setting of thresholds for positive and negative values for each class based on expert knowledge.*

Step 2: *Elimination of redundant features: find such a feature that the interaction index for all pairs containing this feature and for all classes is negative. This means that this feature is redundant for the classification and can be eliminated.*

Step 3: *Selection of the best n_i features: define the number n_i of features that should be selected in this iteration. For each class j, $j = 1, \ldots, m$, perform the following steps.*

Step 3.1: *Select the pair of features with the largest value of the iteraction index and add it to the 'Best subset'. These two features are complementary and should be considered together.*

Step 3.2: *Select the pair of features with the least value of the interaction index and check, whether at least one of two features is contained in the 'Best subset'. If yes, then continue with the next step. If not, look at the values of the importance index for each feature of the pair and add the feature with the largest value to the 'Best subset'. This feature combines the highest degree of redundancy with relative*

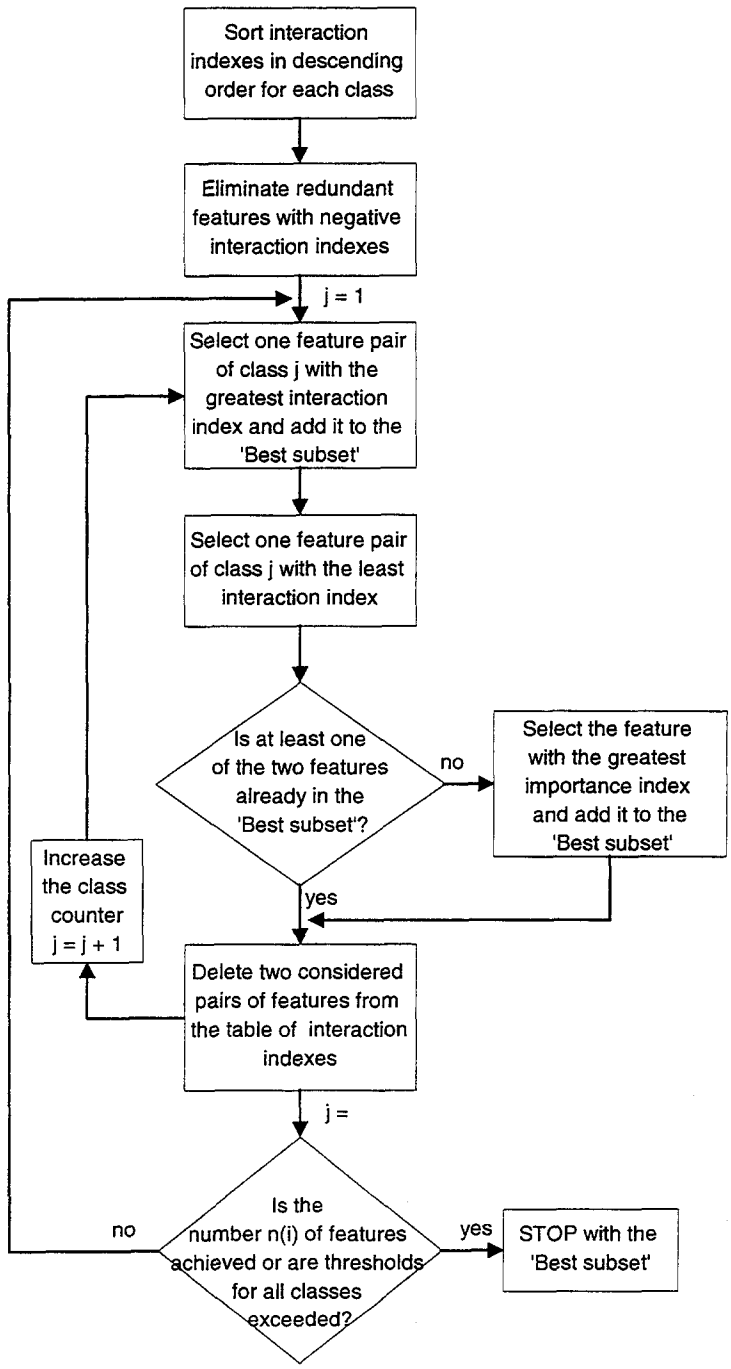

Fig. 2. A procedure for selection of the best subset of features.

high importance. Thus, it is sufficient to select only one feature of the pair.

Step 3.3: *Delete pairs with the largest and the least values of the interaction index from the table. Perform steps* 3.1 *and* 3.2 *for the next class.*

Steps 3.1 *and* 3.2 *are performed for each of m classes until thresholds for positive and negative values of the interaction index are exceeded. Step* 3 *is performed until either thresholds of the interaction index for all classes are exceeded or the maximum number of features to be selected in this iteration is achieved.*

Algorithm 2 constitutes one iteration of the feature selection method. The result of one iteration is the 'Best subset' of n_i features. It is possible that the algorithm terminates with the number of features smaller than n_i, due to the chosen thresholds for interaction indexes. After each iteration if the desired number n of features is not achieved yet, the classification of training data using the subset of selected features is performed and a classifier is designed. Based on new fuzzy measures, new values of importance and interaction indexes are calculated. Algorithm 2 is repeated to select the next best feature subset.

The algorithm is represented graphically in Fig. 2.

To define the thresholds for positive and negative values of the interaction index, an expert can apply graphical representation of index values. Since the range of values can be very different, the thresholds are defined separately for each class. By doing this it must be assured that the best pairs of features for each class will be selected. Moreover, it may be reasonable to set different positive and negative thresholds, because of different density of positive and negative values of the interaction index and different ranges of values. This approach can provide a balance in the selection procedure and an opportunity to consider to the same degree complimentary and redundant features.

The heuristic algorithm for selection of the best feature subset is illustrated by the following example.

Example 3.1. Consider a set of training data, described by 5 features and representing two classes. The goal is to select 3 features. First, the classifier based on 5 features is designed and the fuzzy measures for two classes are identified. Thereafter, importance and interaction indexes are computed for two classes. Suppose that the values of indexes given in Tables 1 and 2 are obtained.

Table 1
An example of importance indexes for a set of 5 features

	Class 1	Class 2
$5v_1$	0.75	0.8
$5v_2$	0.95	1.2
$5v_3$	1.2	1.1
$5v_4$	0.8	0.5
$5v_5$	0.6	0.45

Table 2
An example of interaction indexes for a set of 5 features

	Class 1	Class 2
I_{12}	0.35	0.03
I_{13}	0.15	0.17
I_{14}	−0.1	0.12
I_{15}	−0.15	−0.07
I_{23}	0.24	0.28
I_{24}	0.01	−0.21
I_{25}	−0.3	−0.18
I_{34}	−0.35	−0.05
I_{35}	−0.28	−0.15
I_{45}	−0.05	−0.03

In the first step, the values of the interaction index are sorted in descending order from top to bottom to expedite the further analysis. Before the selection procedure starts, the thresholds for positive and negative values of the interaction index must be set. It can be done using, e.g. the graphical representation (Fig. 3). The interaction indexes are represented in the plane according to Table 2 and can all be marked. The x-axis corresponds to an index of a pair (or a pair themselves) and y-axis corresponds to the values of the interaction index. For class 1, positive threshold can be set, e.g. at 0.2 and the negative threshold at -0.2. For class 2, these thresholds can be chosen as 0.1 and -0.1, respectively.

In the second step, one tries to eliminate the least important features. Checking for all features the values of the interaction index, one finds that all pairs containing feature 5 have a negative value

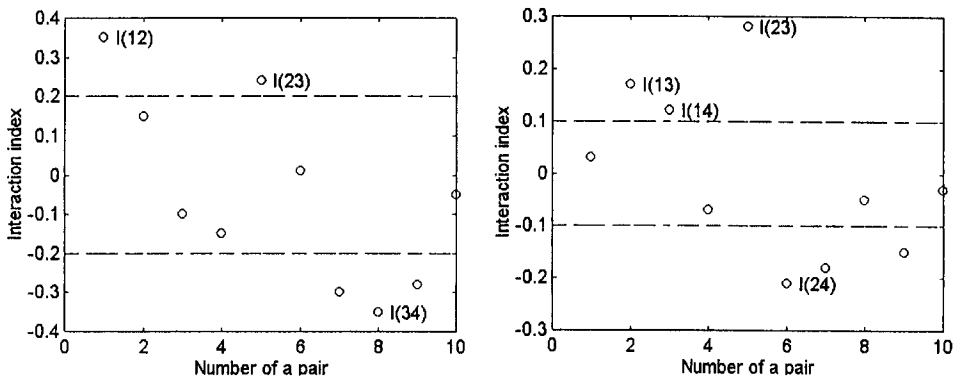

Fig. 3. Interaction indexes for class 1 (left) and class 2 (right).

Table 3
Sorted interaction indexes for features 1–4

Class 1	Class 2
$I_{12} = 0.35$	$I_{23} = 0.28$
$I_{23} = 0.24$	$I_{13} = 0.17$
$I_{13} = 0.15$	$I_{14} = 0.12$
$I_{24} = 0.01$	$I_{12} = 0.03$
$I_{14} = -0.1$	$I_{34} = -0.05$
$I_{34} = -0.35$	$I_{24} = -0.21$

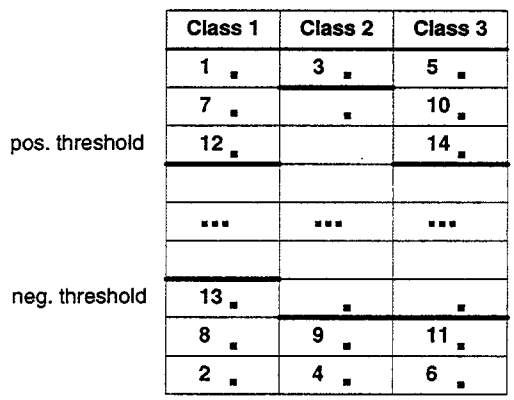

Fig. 4. The selection order of feature pairs in step 3.

of the index. This indicates that feature 5 is redundant for all pairs of features and can be eliminated from the initial feature set. The corresponding values of the interaction index can also be deleted from Table 2.

In the third step, only the interaction indexes for 4 features shown in Table 3 are considered. The number of features to be selected is 3. The selection procedure proceeds as follows:

Starting with class 1, the pair 1,2 with the largest interaction index is selected. Features 1 and 2 are added to the 'Best subset'. Then the pair 3,4 with the least interaction index is chosen. Since none of the two features is in the 'Best subset', one of them can be selected. The importance indexes for both features are compared (they are $v_3 = 1.2$ and $v_4 = 0.8$) and feature 3 with the largest value is added to the 'Best subset'. Since the desired number of features is achieved, the selection procedure terminates with the subset $\{1,2,3\}$.

Another termination criterion is based on positive and negative thresholds of interaction indexes. The selection procedure is repeated until all thresholds are violated, i.e. there is no more pair with valuable contribution. The order, in which pairs are considered in step 3, is generalized in Fig. 4 for the case of three classes. Points denote the sorted values of the interaction index and numbers correspond to the order of consideration of pairs.

In general, it should be noted that the feature selection procedure, which requires a classifier design for each selected subset of features, is not as efficient as some methods, operating independent of the classifier. But it is still better than computing a classifier for each possible combination of features. The number of iterations, and classifiers respectively, is predefined and could be $n' - n$ at the maximum, where n' is the ini-

tial number and n is the desired number of features, if one feature is eliminated in each iteration. For enumeration this number is equal to $\binom{n'}{n}$. The computational complexity of the algorithm by using the 2-additive fuzzy measure is mainly related to the computation of fuzzy measure coefficients in each iteration. The total number of coefficients can be evaluated by

$$\sum_{k=n}^{n'} \frac{k(k+1)}{2} \leqslant (n'-n)^3,$$

where n' is the initial number of features and n is the number to be selected. Thus, the complexity of the algorithm grows cubicly with respect to the difference between the initial number and the desired number of features.

In Section 5, the heuristic algorithm for feature selection based on the fuzzy measure will exemplarily be applied to the problem of frequency spectra analysis.

4. A pattern recognition method based on the fuzzy integral

In the following, the problem of supervised classification of objects is discussed. Consider a set of m classes $C_1,\ldots,C_m$, each class being described by a set of n features. Each object $\boldsymbol{x}$ is represented as an n-dimensional vector $\boldsymbol{x} = [x_1,\ldots,x_n]$. Suppose that a training set of objects for each class is given, which is used during the learning process to derive a classifier. When a new object $\boldsymbol{x}^o$ is observed, we want to find the class, to which $\boldsymbol{x}^o$ most likely belongs.

The idea of a pattern recognition method based on the fuzzy integral [14] is to build fuzzy prototypes of classes in the form of fuzzy sets and during the classification to match a new object with all class prototypes and to choose the class with the highest matching degree (this is also a general idea of the fuzzy pattern matching approach introduced in [2]). Specifically, consider for each class C_j, $j=1,\ldots,m$, a collection of fuzzy sets $v_1^j,\ldots,v_n^j$, defined on each feature and modeling the fuzzy sets of typical values of the features for class j. When a new object $\boldsymbol{x}^o = [x_1^o,\ldots,x_n^o]$ is observed, the matching process is carried out in two steps:

- partial matching with respect to feature i: determination of a partial matching degree ϕ_i^j between a feature value x_i^o of an object and a class prototype v_i^j. Partial matching is done for all features and all classes;
- global matching between the object $\boldsymbol{x}^o$ and the class prototype C_j: all partial matching degrees concerning class j are aggregated to a global degree by the fuzzy integral with respect to a fuzzy measure

$$\Phi(C_j \mid \boldsymbol{x}^o) = \mathscr{F}_\mu(\phi_1^j,\ldots,\phi_n^j). \tag{4.1}$$

A fuzzy measure μ is defined on a set of features for each class and expresses the importance of features and groups of features for the classification, i.e. the contribution of single features and groups of features into the recognition process. Object $\boldsymbol{x}^o$ is classified by assigning it to the class with the highest global matching degree.

A pattern recognition method based on the fuzzy integral represents the most general framework among fuzzy pattern matching techniques. This is due to the fact that the fuzzy integral used for aggregation contains most of the known averaging operators including minimum and maximum as the limit cases [14]. This method was proposed in [13] as a generalization of a fuzzy pattern matching approach introduced in [2] and then extended in [7]. In the approach of Grabisch and Sugeno [13], the assignment of objects to classes is crisp, although the global degrees of confidence, or matching degrees, take their values in interval $[0,1]$.

In the following, the pattern recognition method based on the fuzzy integral is slightly modified: the value of the fuzzy integral (4.1) expressing the matching degree between an object and a class is interpreted as the degree of membership of an object to a class. As a result, an object belongs to each class to some degree. The sum of the degrees of membership of an object to all classes is not necessarily 1, so they can represent the typicality of objects to classes. The classification procedure is illustrated in Fig. 5.

The classification procedure used in a pattern recognition method based on the fuzzy integral requires the knowledge of class prototypes and fuzzy measures for each class. Therefore, the design of the fuzzy integral classifier includes the following two steps [13]:

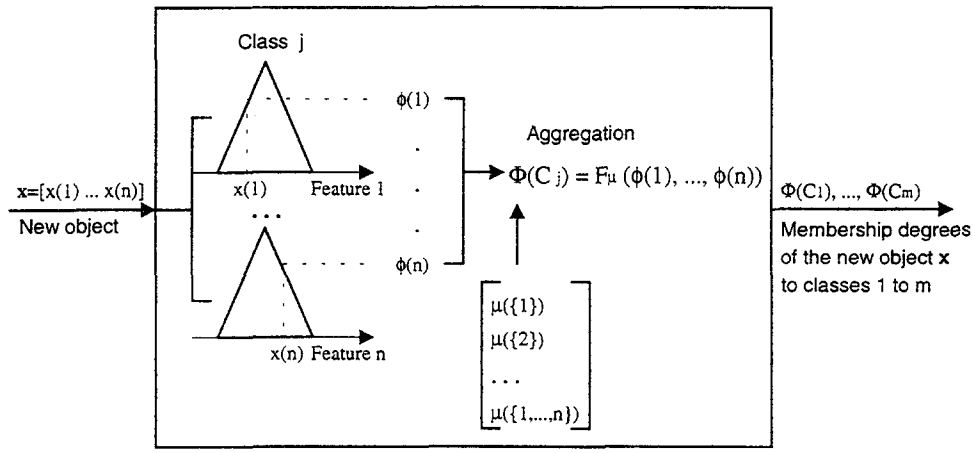

Fig. 5. Classification procedure used in a pattern recognition method based on the fuzzy integral.

- learning of class prototypes in the form of fuzzy sets based on training data;
- identification of the fuzzy measure for each class. That is, for m classes and n features, $m2^n$ coefficients of the fuzzy measure for all subsets of a set of n elements and for each class must be identified.

In this paper, class prototypes are learned using a method described in [6], which provides possibilistic histograms of data. The ideal of this method is to build a classical probability density function from the data and to map it into a possibility distribution, which has the same shape.

Identification of the fuzzy measure for each class using learning data is based on the idea of modeling a system with n input variables $x_1, \ldots, x_n$ and one output in the form of the fuzzy integral $\mathscr{F}_\mu(\boldsymbol{x})$, where $\boldsymbol{x}$ is an n-dimensional input vector and μ is a fuzzy measure on the set of n inputs [14]. Suppose that a set of l learning data in the form of couples $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_l, y_l)$ is given, where y_k, $k = 1, \ldots, l$, is the output of the system, if the input is $\boldsymbol{x}_k$. The goal of the identification is to find the best fuzzy measure μ for a given fuzzy integral (Choquet or Sugeno) such that the error between the actual output and the expected output of the system is minimized (in general the fuzzy measure is dependent on the class):

$$E^2(\mu) = \sum_{k=1}^{l} (\mathscr{F}_\mu(\boldsymbol{x}_k) - y_k)^2. \tag{4.2}$$

The variable in this expression is the fuzzy measure μ, which can be expressed as a $(2^n - 2)$-dimensional vector $\boldsymbol{u}$ of coefficients of the fuzzy measure (coefficients $\mu_\emptyset$ and μ_X are not included, since their values are 0 and 1, respectively):

$$\boldsymbol{u} = [\mu_1 \mu_2 \ldots \mu_n \ \mu_{12} \mu_{13} \ldots \mu_{1n} \ldots \mu_{n-1,n} \mu_{123} \ldots \mu_{23\ldots n}], \tag{4.3}$$

where $\mu_i = \mu(\{x_i\})$, $\mu_{ij} = \mu(\{x_i, x_j\})$, etc.

According to the definition of the fuzzy measure [37], the components of this vector must satisfy a set of monotonicity constraints, which take the form $u_i \leqslant u_j$ or $u_i \leqslant 1$, $i, j = 1, \ldots, (2^n - 2)$.

Thus, to solve the identification problem, a constraint optimization technique must be applied. The choice of the particular optimization method depends on the fuzzy integral chosen and the error criterion used.

Using the Choquet integral, optimization methods based on gradient techniques can be used to minimize criterion (4.2), but in the case of the Sugeno integral methods such as simulated annealing [1] or genetic algorithms [22] can be applied, since functions induced by the operators minimum and maximum are not always differentiable. All criteria discussed in this section consider the Choquet integral, which makes the solution of the optimization problem easier.

In [14] three types of criteria for the correct identification of the fuzzy measure were proposed and it was

shown that the best one is the generalized quadratic criterion. For the sake of clarity, the criterion is given only for the case of two classes:

$$E^2 = \sum_j \sum_k |\Psi(\Delta\Phi_{12}(\boldsymbol{x}_k^j)) - 1|^2, \tag{4.4}$$

where Ψ is any increasing function from $[-1,1]$ to $[-1,1]$, preferably a sigmoid-type function, and $\Delta\Phi_{12}(\boldsymbol{x}_k^j)$ denotes the degree of class discrimination for each datum $\boldsymbol{x}_k^j$ of class j, $j=1,2$. This value is defined by [14]

$$\Delta\Phi_{12}(\boldsymbol{x}_k^1) = \Phi_{\mu 1}(C_1|\boldsymbol{x}_k^1) - \Phi_{\mu 2}(C_2|\boldsymbol{x}_k^1), \tag{4.5}$$

for each datum $\boldsymbol{x}_k^1$ of class 1, where $\Phi_{\mu j}(\cdot)$ is a global matching degree between an object and the class prototype defined by (4.1). The subscript j is added to the fuzzy measure μ^j to express its dependency on the class. The degree of class discrimination for each datum $\boldsymbol{x}_k^2$ of class 2 is defined analogously with indexes being inverted. A correct classification of training samples produces positive values of $\Delta\Phi_{12}$, a misclassification leads to negative values.

Criterion (4.4) can be minimized using constrained least mean-squares techniques or some heuristic algorithms. Optimization algorithms exhibit a number of problems due to the sparse constraint matrix and for a low number of training data. Also they are very time consuming and not incremental, i.e. the obtained solution cannot be updated using new data.

Therefore, several heuristic algorithms for the identification of the fuzzy measure were proposed in the literature based on intuitive considerations rather than explicit criteria. It is known that they are usually much less time consuming and can be better adjusted to specific characteristics of the problem, but they do not guarantee the optimal solution. Nevertheless, such algorithms can be very useful in practical situations. A heuristic algorithm for the identification of the fuzzy measure in the case of the Sugeno integral was first proposed in [17]. This algorithm was later modified for the case of the Choquet integral in [23]. A considerable improvement of the latter algorithm was presented in [9], where a heuristic least mean-squares (HLMS) algorithm was proposed.

According to results given in [9], the HLMS algorithm requires much less computing time than constraint minimization methods with a slight loss of optimality. The HLMS algorithm does not have restrictions on the number of training data and can be used for a larger number of features than optimization techniques. However, the number of features is still restricted, due to the necessity to compute $m2^n$ coefficients of the fuzzy measure. In general, the exponential complexity of the fuzzy measure is a serious obstacle for an application of the classification method based on the fuzzy integral. Therefore, the next obvious step in the development of this classification approach is to use k-additive fuzzy measures introduced in [10].

In Section 2, the HLMS (2-add) algorithm for the identification of 2-additive fuzzy measures was presented. Applied to the procedure of classifier design, it requires some specifications.

To step 1.1: As was stated above, the best error criterion for the correct identification of the fuzzy measure is the generalized quadratic criterion in the form (4.4). Thus, to minimize this criterion using the HLMS (2-add) algorithm, the expression

$$E = \Psi(\Delta\Phi_{12}(\boldsymbol{x}_k^j)) - 1$$

must be used for the computations. Due to the form of a sigmoid function $\Psi(t)$, the value of E is always negative or equal zero. Considering the case of two classes for simplicity, note that each training datum involves coefficients of two fuzzy measures μ^1 and μ^2 describing two classes. Thus, the modifications in step 1.2 are made for two vectors $\boldsymbol{u}^1$ and $\boldsymbol{u}^2$.

To step 1.2: For a training datum belonging to class 1, minimizing E means increasing the value of $\Delta\Phi_{12}$. For this purpose, the value $\boldsymbol{u}^1(i)$ must be increased according to (2.8) and the values $\boldsymbol{u}^2(i)$ must be decreased using $-E$ instead of E in (2.8). These considerations are based on the fact that the Choquet integral is a linear function of the fuzzy measure.

The main advantage of the HLMS (2-add) algorithm compared to the HLMS algorithm is its lower computational complexity, which is quadratic with respect to the number of features. Therefore, the classification method using the 2-additive fuzzy measure can be applied to problems described by much larger number of features than in the case of a general fuzzy measure.

In the next section, both algorithms HLMS and HLMS (2-add) are compared by being applied to the problem of automatic bearing diagnosis.

5. Application of the fuzzy integral classifier to automatic bearing diagnosis

The goal of this section is to design a pattern recognition system based on the fuzzy integral for machine diagnosis, in particular for automatic diagnosis of bearings. The main objective of machine diagnosis is a precocious recognition of mechanical defects in a machine, which results in a reduction of machine failures and high machine availability. Since a permanent machine monitoring by a trained operator is very expensive and in some cases not sufficient, automatic diagnostic systems are developed.

It was shown in [8] that bearing conditions can be evaluated using the envelope frequency spectrum of the vibration signal generated by an operating bearing. In the envelope spectrum special defects of bearings can be easily recognized by using particular frequencies as features. In this application, only two types of bearings, intact and with a damaged outer ring, are considered.

5.1. Training data used for classifier design

For the purpose of classifier design, 30 intact bearings and 30 bearings with a damaged outer ring were observed and the corresponding vibration signals were registered by a sensor. All measurements were taken on the test device and preprocessed using a special hardware. With the data acquisition board, 8192 values were selected from each time signal, transmitted by a special software into the computer and processed using Fourier transformation. The resulting data cover the interval from 0 to 4000 Hz and correspond to a frequency resolution of 0.488 Hz. These frequency spectra can be used as observation vectors describing bearings. They contain all acoustic information about a possible damage of a bearing.

However, for correct recognition of a damage it is not necessary to consider all frequencies of the envelope spectrum as features. For the purpose of detailed analysis, the initial set of features was limited according to suggestions of an experienced expert to the set of the most promising features. It contains frequencies with large amplitudes, in particular characteristic frequency and its first and second harmonics. They can be calculated if the bearing geometry and the revolution speed are known, depending on a bearing defect [8].

Table 4
Features used for classification of bearings

Features	Frequency interval
1	118–126
2	239–249
3	359–371
4	85–95
5	140–150
6	260–270
7	390–400

For roller bearings with outer ring damage considered here, the kinematic frequency with 1 RPM is equal to 3.565 Hz. In the conducted experiments with a revolution speed of 1000 RPM the characteristic frequencies correspond to 59.42, 118.84 and 178.26 Hz, respectively. Because of a high revolution speed, they show some variations and thus, certain intervals containing characteristic frequencies should be considered. Taking into account a frequency resolution in measured data, three intervals were chosen: 118–126, 239–249 and 359–371. Moreover, four other frequency intervals were suggested by an expert, which are all given in Table 4.

These intervals are used as relevant features for the classification and their values are determined as a maximum amplitude over a corresponding interval. Hence, each training datum is described by a seven-dimensional feature vector containing maximum amplitudes of vibration signals in chosen frequency intervals. The goal is to select the best set of three features and to design a classifier for an automatic diagnosis of bearings, based on the given training data. The classification objects are bearing conditions and two classes, 'intact bearings' and 'damaged bearings', are considered.

5.2. Feature selection for bearing diagnosis

According to the feature selection algorithm described in Section 5, the classifier based on 7 features should be at first designed to be able to evaluate the performance of features and pairs of them. The design procedure proceeds in two steps. In the first step, class prototypes in the form of fuzzy sets representing typical values of each feature for each class are learned from training data. They are calculated as possibilistic histograms of data [6]. Fuzzy sets for the first class

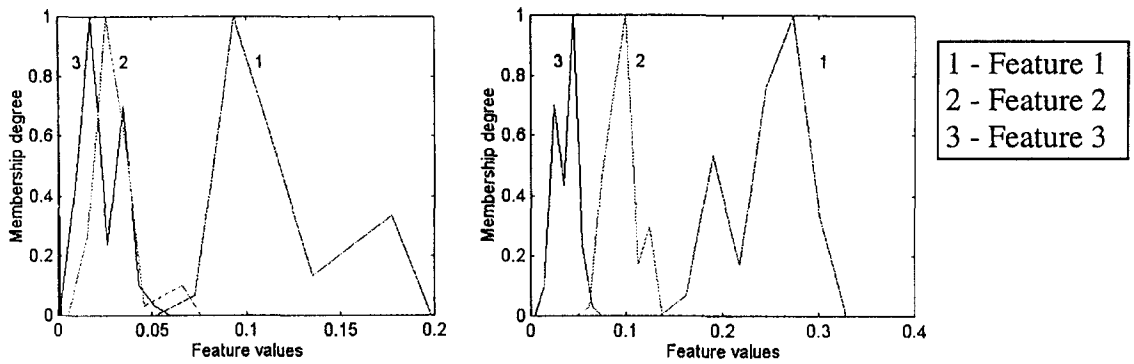

Fig. 6. Fuzzy sets representing class 1 (left) and class 2 (right).

of intact bearings and for the second class of damaged bearings are shown in Fig. 6. Since the values of features 4–7 are very small compared to the ones of features 1–3, one cannot see the corresponding fuzzy sets on the figures.

Considering the class prototypes, it can be noticed that fuzzy sets corresponding to the same features have different supports for two classes, i.e. typical values of the same features cover different intervals of the domain for two classes, which is a necessary requirement for a good classification. However, it is not clear from this representation, how good the single features are in distinguishing between classes. To get a better understanding, another illustration of data can be used. Fig. 7 shows for each feature two fuzzy sets representing two classes. It can be observed that features 1–3 can recognize rather well two classes, while features 4–7 are very poor in distinguishing between classes.

After the class prototypes have been learned, the next step of classifier design is the identification of fuzzy measures for each class. This can be done using either the HLMS algorithm for a general fuzzy measure or the HLMS (2-add) algorithm for the 2-additive fuzzy measure. As was stated earlier, the 2-additive fuzzy measure requires only $n(n+1)/2$ coefficients to be defined in contrast to 2^n coefficients for a general fuzzy measure, thus it is better suited for a large number of features. If the number of features is small, it is better to use a general fuzzy measure, because of its richness and flexibility. In this application for the purpose of analysis, both algorithms are applied and their results are compared.

The coefficients of two fuzzy measures obtained with the help of the HLMS algorithm are shown in Fig. 8. The coefficients are numbered from 1 for μ_1 to 128 for $\mu_{1234567}$ and are given on the x-axis. The y-axis shows the values of the corresponding coefficients. One can easily recognize equilibrium states of the measure and small deviations of the coefficients from these states. The algorithm preserves a homogeneous structure of the fuzzy measure.

The corresponding coefficients of the fuzzy measures obtained with the HLMS (2-add) algorithm are illustrated in Fig. 9. The fuzzy measures are monotone, but there are no equilibrium states in their structure (except first two layers initialized as in HLMS). The range of variations of the coefficients is much larger than in the previous case, which can be explained by the fact that most coefficients are calculated from coefficients of the first two layers, but not modified with respect to the initialized state. This does not mean, however, a decline of the classifier performance. The rate of reclassification of the training data is 100% in both cases.

Thus, it can be deduced that the 2-additive fuzzy measure is equally good for classification. Numerous tests have also shown that for the correct identification of the 2-additive fuzzy measure less iterations are needed in general, since only first two layers of coefficients must be learned.

After the classifier is designed and the fuzzy measures for both classes are identified, the feature selection algorithm described in Section 5 can be applied. For the purpose of comparison, feature selection is performed based on general and 2-additive fuzzy

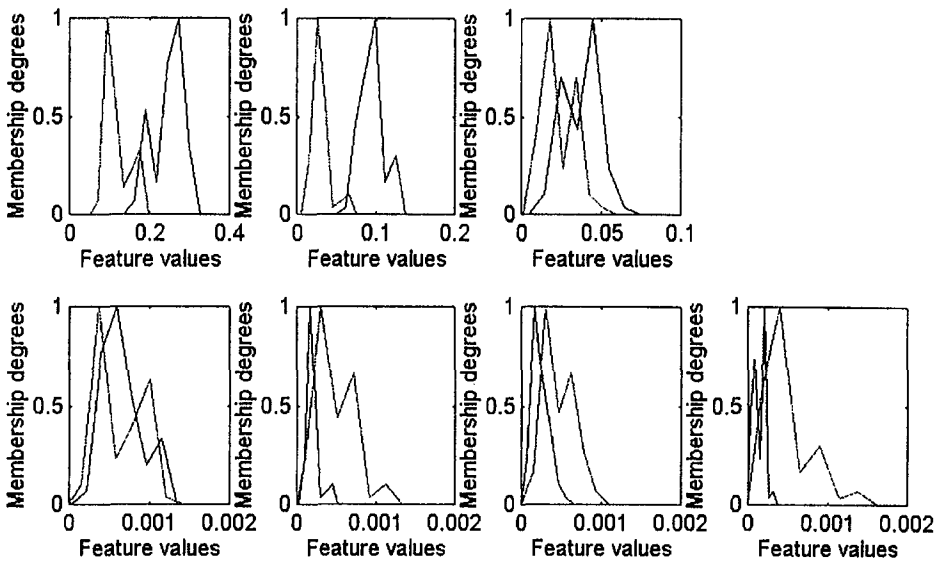

Fig. 7. Fuzzy sets of features from 1 to 7 for two classes (from left to right and from top to bottom).

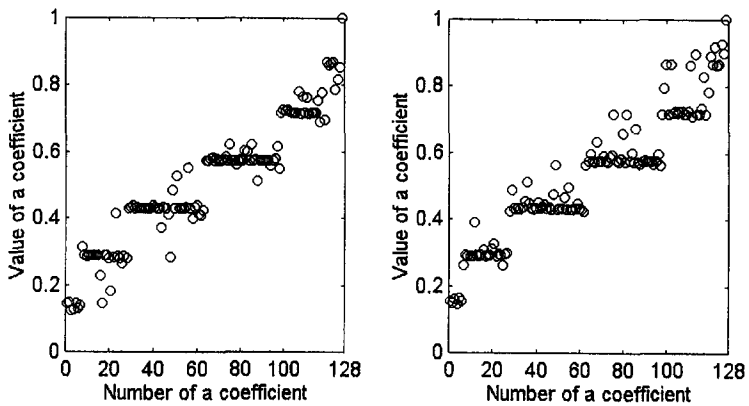

Fig. 8. Coefficients of the fuzzy measure for class 1 (left) and class 2 (right) identified with HLMS.

measures. Suppose that the selection procedure consisting of two iterations is desirable: 5 features should be selected in the first iteration and 3 features in the second iteration. Consider the importance and interaction indexes obtained for the 2-additive fuzzy measure (the interaction indexes are already sorted) (Table 5).

Suppose that positive and negative thresholds for interaction indexes are chosen as marked in Table 6. It can be observed that there is no feature that can be eliminated as being redundant (all interaction indexes containing a feature must be negative). Based on the selection procedure, the 5 following features can be selected before all thresholds are exceeded:

Table 5
Importance indexes for a set of 7 features obtained with HLMS (2-add)

	Class 1	Class 2
1	0.9023	1.9379
2	1.4516	1.7105
3	0.8553	0.5838
4	0.6141	0.7285
5	1.414	0.6859
6	0.773	0.7136
7	0.9897	0.6398

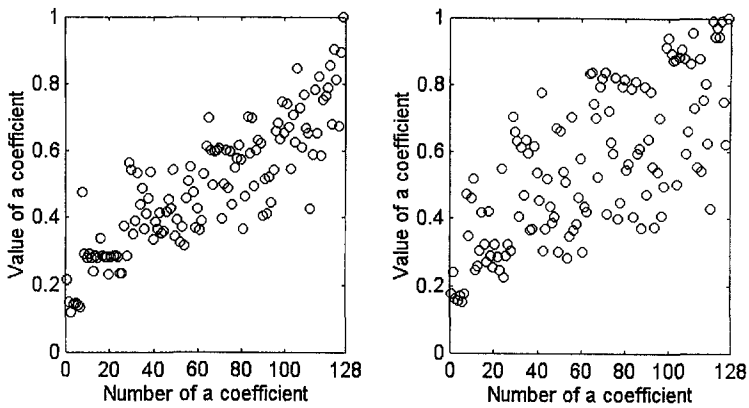

Fig. 9. Coefficients of the fuzzy measure for class 1 (left) and class 2 (right) identified with HLMS (2-add).

$\{1,2,3,4,5\}$. The feature selection algorithm is repeated from the beginning using the new set of features to design a classifier.

In the second iteration of the selection procedure, three features are chosen already after three steps and the best subset of features contains $\{1,2,3\}$.

If a general fuzzy measure is used for the classifier design, then the result of the first iteration is different and the best subset of features is given by $\{1,2,3,4,6\}$. However, in the second iteration the same subset of features $\{1,2,3\}$ as in the previous case is chosen. The set of selected features corresponds in fact to cinematic frequency and its first and second harmonics. Thus, these features are indeed sufficient for detection of an outer ring damage in bearings as it was stated in [8].

The achieved results corroborate the fact that the 2-additive fuzzy measure shows the equivalent performance compared to a general fuzzy measure and can be used for modeling the importance of features and the pairwise interaction between features without loss of information. It is suited for classification in the case of a high number of features and provides considerable complexity reduction against a general fuzzy measure in the classifier design. Moreover, the application results validate the effectiveness of the feature selection algorithm proposed in this paper.

5.3. *Design of the fuzzy integral classifier for bearing diagnosis*

In the previous section, it was shown that the set of 3 features representing the characteristic frequencies in the envelope spectrum of the vibration signal is the most appropriate set for bearing diagnosis. To demonstrate this, consider a set of training data in a three-dimensional feature space (Fig. 10). One can easily recognize two distinct classes.

The classifier, designed using these training data, is represented by class prototypes shown in Fig. 6 and two fuzzy measures consisting of eight coefficients for two classes. For the identification of the fuzzy measure, the HLMS algorithm was used, since the

Table 6
Interaction indexes for a set of 7 features obtained with HLMS (2-add) with chosen positive and negative thresholds

Class 1		Class 2	
2,3	0.0242	4,5	0.0283
2,4	0.0174	2,3	0.0186
3,4	0.0134	**1,3**	**0.0100**
6,7	0.0114	1,7	0.0029
2,7	**0.0108**	4,6	0.0024
3,6	0.0075	6,7	0.0012
2,6	0.0055	4,7	−0.0031
1,4	0.0051	1,6	−0.0040
3,7	0.0042	1,4	−0.0057
1,3	0.0040	1,5	−0.0076
1,6	0.0020	3,4	−0.0113
5,7	0.0004	3,5	−0.0126
4,6	−0.0019	3,7	−0.0133
1,7	−0.0071	2,5	−0.0139
3,5	−0.0076	1,2	−0.0156
1,2	**−0.0183**	5,7	−0.0160
4,7	−0.0231	2,7	−0.0181
2,5	−0.0301	**5,6**	**−0.0186**
5,6	−0.0311	2,4	−0.0208
1,5	−0.0328	3,6	−0.0242
4,5	−0.0409	2,6	−0.0265

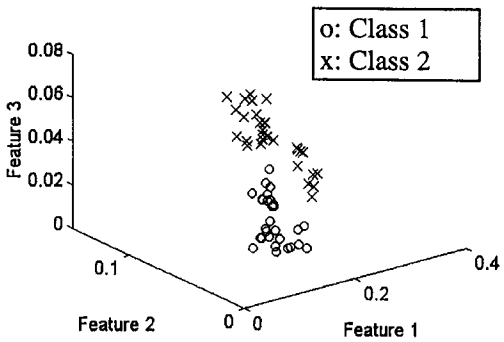

Fig. 10. A set of 60 training data in a three-dimensional feature space.

number of features is small and the 2-additive fuzzy measure has no advantage in this case.

The classifier was tested with the same set of training data (resubstitution test). The rate of reclassification is 96.7% (two objects are classified wrong).

To estimate the classification rate of the fuzzy integral classifier, the 10-fold cross validation test was performed. This test consists of the following steps: the set of training data is split into 10 pieces and the classifier is trained on nine pieces and then tested on the last. This procedure is repeated for all 10 permutations. The fuzzy integral classifier applied for the classification of bearings provides an estimated classification rate of 99.2% with a standard deviation of 0.8.

The results presented show that the fuzzy integral classifier can be successfully applied to automatic diagnosis of bearings. As to its general applicability the following considerations can be useful.

Each classification method can be characterized by its suitability to certain data structures and there is probably no classifier that is able to recognize all possible data structures. The purpose of this paragraph is to determine in general abilities of the fuzzy integral classifier based on the analysis of its internal structure and to detect its possible application areas.

The performance of different classification methods is influenced to a high degree by the criterion underlying the classification process, which can be based either on a similarity or a distance measure. In general, these two measures can be viewed to be inverse to each other, but algorithmically they lead to completely different techniques. The fuzzy integral classifier uses as a basic criterion for a design of classes a similarity between objects and class prototypes. In contrast to many clustering methods, a class prototype is defined not by a point representing the most typical object of a class, but by a collection of fuzzy sets representing typical values of each feature for this class. The similarity (or compatibility) of an object with a class prototype is determined by aggregating degrees of compatibility of each feature value of an object with typical values of corresponding features for a given class under consideration of importance of single features and their groups. This similarity value can be interpreted as a degree of membership of an object to a class. It is calculated for each class independent from other class prototypes and, thus, can represent typicality of an object to a class. This is an important property of the fuzzy integral classifier inherent also in the possibilistic c-means algorithm [21], because noise data are often present among objects in real applications and they can distort the classification process.

The fuzzy integral classifier can distinguish different shape, size and density of classes provided that at least some of the features describing objects have good discriminating ability, i.e. domains occupied by features for different classes are separated. The information about the form of classes is reflected in the form of fuzzy sets representing classes, which are determined in this paper as possibilistic histograms.

Based on the knowledge of the internal structure of the fuzzy integral classifier, its behavior in different situations can be predicted and limits of its abilities can be evaluated.

6. Conclusions

This paper is concerned with feature selection and classification based on the fuzzy integral as the most general framework among pattern matching techniques. Both procedures require the identification of the fuzzy measure, which presents the main difficulties for applications due to exponential complexity of the fuzzy measure. Therefore, a new heuristic algorithm for the identification of the 2-additive fuzzy measure was proposed. This type of the fuzzy measure represents an intermediate solution in the sense of richness and complexity and is sufficient for the semantic interpretation of the fuzzy measure. The computational complexity of the algorithm is quadratic with respect

to the number of features. Thus, using the 2-additive fuzzy measure it is possible to handle classification problems described by a large number of features.

Another research point of the paper was focused on the development of the feature selection procedure for the fuzzy integral classifier. The heuristic algorithm proposed is based on two feature-evaluation criteria, defined using the semantic interpretation of the fuzzy measure. The feature-selection algorithm depends on the fuzzy integral classifier, thus its complexity in the case of the 2-additive fuzzy measure is cubic with respect to a difference between the initial number and the desired number of features.

To investigate properties of the fuzzy integral classifier, it was applied to automatic bearing diagnosis. For the purpose of analysis, feature selection was based on a general fuzzy measure and on the 2-additive fuzzy measure. In both cases the same set of three features was selected as the best. These features correspond to characteristic frequencies in the envelope spectrum of the vibration signal. The classifier designed using these three features provides good discrimination ability and shows its applicability for automatic bearing diagnosis.

Based on the analysis of the fuzzy integral classifier, it was shown that the classifier can recognize the typicality of an object for a class. This ability of the classifier is very important for dealing with 'unclear' data and noise. Moreover, the classifier can recognize different shape, size and density of classes provided that features have good discriminating ability. If different classes have similar prototypes, then the classifier cannot provide desired results. This property of the classifier presents certain limits of its abilities.

Further research with respect to the fuzzy integral classifier should aim at the development of a method for the classification of fuzzy objects or, in general, objects described by features in the form of arbitrary functions. This problem can arise in dynamical pattern recognition or in the classification of systems described by some functional relationships.

Acknowledgements

This research was partially supported by a DAAD-Grant and by Project DFG-Zi 104/27-1

References

[1] E. Aarts, J. Korst, Simulated Annealing and Bolzmann Machines, Wiley, Chichester, 1980.
[2] M. Cayrol, H. Farreny, H. Prade, Fuzzy Pattern Matching, Kybernetes 11 (1982) 103–116.
[3] A. Chateauneuf, J.Y. Jaffray, Some characterizations of lower probabilities and other monotone capacities through the use of Möbius inversion. Math. Social Sci. 17 (1989) 263–283.
[4] D. Denneberg, Non-additive measure and integral, Theory and Decision Library Series B, vol. 27, Kluwer Academic, Dordrecht, Boston, 1994.
[5] P. Devijver, Statistical pattern recognition, in: K.S. Fu, Applications of Pattern Recognition, CRC Press, Boca Raton, 1982, pp. 15–36.
[6] D. Dubois, H. Prade, Unfair coins and necessity measures: toward a possibilistic interpretation of histograms, Fuzzy Sets and Systems 10 (1983) 15–20.
[7] D. Dubois, H. Prade, Possibility Theory – An Approach to Computerized Processing of Uncertainty, Plenum Press, New York, 1988.
[8] B. Geropp, Schwingungsdiagnose an Wälzlagern mit Hilfe der Hüllkurvenanalyse, Dissertation RWTH Aachen, 1995.
[9] M. Grabisch, A new algorithm for identifying fuzzy measures and its application to pattern recognition, Internat. Joint Conf. of the 4th IEEE Internat. Conf. on Fuzzy Systems and the 2nd Internat. Fuzzy Engineering Symp., Yokohama, Japan, March 1995, pp. 145–150.
[10] M. Grabisch, K-order additive fuzzy measures. Proc. 6th Internat. Conf. on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU), Granada, Spain, July 1996a, pp. 1345–1350.
[11] M. Grabisch, The representation of importance and interaction of features by fuzzy measures, Pattern Recognition Lett. 17 (6) (1996b) 567–575.
[12] M. Grabisch, K-order additive discrete fuzzy measures and their representation, Fuzzy Sets and Systems 92 (1997) 167–189.
[13] M. Grabisch, M. Sugeno, Multi-attribute classification using fuzzy integral, 1st IEEE Internat. Conf. on Fuzzy Systems, San Diego, March 1992, pp. 47–54.
[14] M. Grabisch, H.T. Nguyen, E.A. Walker, Fundamentals of Uncertainty Calculi with Applications to Fuzzy Inference, Kluwer Academic, Dordrecht, 1995.
[15] P.L. Hammer, R. Holzman, On approximations of pseudo-Boolean functions, ZOR-Methods Models Oper. Res. 36 (1992) 3–21.
[16] H. Ichihashi, H. Tanaka, K. Asai, Fuzzy integrals based on pseudo-addition and multiplication, J. Math. Anal. Appl. 130 (1988) 354–364.
[17] K. Ishii, M. Sugeno, A model of human evaluation process using fuzzy measure, Internat. J. Man–Mach. Stud. 22 (1985) 19–38.
[18] R.L. Keeney, H. Raiffa, Decisions with Multiple Objectives, Wiley, New York, 1976.
[19] J.M. Keller, P. Gader, H. Tahani, J.-H. Chiang, M. Mohamed, Advances in fuzzy integration for pattern recognition, Fuzzy Sets and Systems 65 (1994) 273–283.

[20] J. Kittler, A review of feature extraction methods based on probabilistic separability measures, Proc. SITEL-ULG Conf. on Pattern Recognition, Societe Belge des Ingenieurs des Telecommunications et d'Electronique, Ophain, B.S.I., Belgium, 1977.

[21] R. Krishnapuram, J. Keller, A possibilistic approach to clustering, IEEE Trans. Fuzzy Systems 1 (1993) 98–110.

[22] Z. Michalewicz, Genetic Algorithms + Data Structures = Evolutionary Programs, Springer, Berlin, 1992.

[23] T. Mori, T. Murofushi, An analysis of evaluation model using fuzzy measure and the Choquet integral, 5th Fuzzy System Symp., Kobe, 1989, pp. 207–212 (in Japanese).

[24] A.N. Mucciardi, E.E. Gose, A comparison of seven techniques for choosing subsets of pattern recognition properties, IEEE Trans. Comput. C-20 (1971) 1023–1031.

[25] T. Murofushi, A technique for reading fuzzy measures (1): the Shapley value with respect to a fuzzy measure, 2nd Fuzzy Workshop, Nagaoka, Japan, October 1992, pp. 39–48 (in Japanese).

[26] T. Murofushi, S. Soneda, Techniques for reading fuzzy measures (3): interaction index, 9th Fuzzy System Symp., Sapporo, May 1993, pp. 693–696 (in Japanese).

[27] T. Murofushi, M. Sugeno, Fuzzy t-conorm integrals with respect to fuzzy measures: generalization of Sugeno integral and Choquet integral, Fuzzy Sets and Systems 42 (1991) 57–71.

[28] T. Murofushi, M. Sugeno, Some quantities represented by the Choquet integral, Fuzzy Sets and Systems 56 (1993) 229–235.

[29] G.C. Rota, On the foundations of combinatorial theory. I. theory of Möbius functions, Z. Wahrscheinlichkeitstheorie Verw. Gebiete 2 (1964) 340–368.

[30] L.S. Shapley, A value for n-person games, in: H.W. Kuhn, A.W. Tucker (Eds.), Contributions to the Theory of Games, vol. 2 (28) in Annals of Mathematics Studies, Princeton University Press, Princeton, NJ, 1953, pp. 307–317.

[31] M. Sugeno, Theory of fuzzy integrals and its applications, Ph.D. Dissertation, Tokyo Institute of Technology, 1974.

[32] M. Sugeno, T. Murofushi, Pseudo-additive measures and integrals, J. Math. Anal. Appl. 122 (1987) 197–222.

[33] M. Sugeno, T. Murofushi, An interpretation of fuzzy measure and the Choquet integral as an integral with respect to a fuzzy measure, Fuzzy Sets and Systems 29 (1989) 201–227.

[34] M. Sugeno, K. Fujimoto, T. Murofushi, A hierarchical decomposition of Choquet integral model, Internat. J. Uncertainty, Fuzziness Knowledge-Based Systems 3 (1995) 1–15.

[35] C.W. Therrien, Decision Estimation and Classification: An Introduction to Pattern Recognition and Related Topics, Wiley, New York, 1989.

[36] P. Wakker, A behavioral foundation for fuzzy measures, Fuzzy Sets and Systems 37 (1990) 327–350.

[37] Z. Wang, G.J. Klir, Fuzzy Measure Theory, Plenum Press, New York, 1992.

[38] S. Weber, $\perp$-decomposable measures and integrals for Archimedian t-conorms $\perp$, J. Math. Anal. Appl. 101 (1984) 114–138.

[39] H.-J. Zimmermann, Fuzzy Set Theory – and its Applications, 3rd ed., Kluwer, Boston, Dordrecht, London, 1996.

[40] H.-J. Zimmermann, Fuzzy sets in pattern recognition, in: P.A. Devijver, J. Kittler (Eds.), Pattern Recognition Theory and Applications, NATO ASI Series, Springer, Berlin, Heidelberg, 1987, pp. 383–391.