Fuzzy Relational Classifier Trained by Fuzzy Clustering

Magne Setnes, *Student Member, IEEE*, and Robert Babuška

***Abstract*—A novel approach to nonlinear classification is presented. In the training phase of the classifier, the training data is first clustered in an unsupervised way by fuzzy c-means or a similar algorithm. The class labels are not used in this step. Then, a fuzzy relation between the clusters and the class identifiers is computed. This approach allows the number of prototypes to be independent of the number of actual classes. For the classification of unseen patterns, the membership degrees of the feature vector in the clusters are first computed by using the distance measure of the clustering algorithm. Then, the output fuzzy set is obtained by relational composition. This fuzzy set contains the membership degrees of the pattern in the given classes. A crisp decision is obtained by defuzzification, which gives either a single class or a "reject" decision, when a unique class cannot be selected based on the available information. The principle of the proposed method is demonstrated on an artificial data set and the applicability of the method is shown on the identification of live-stock from recorded sound sequences. The obtained results are compared with two other classifiers.**

***Index Terms*—Classification, fuzzy clustering, fuzzy relations, pattern recognition, recognition of sound sequences.**

I. Introduction

THE objective of pattern recognition is the identification of structures in data similar to known structures. In the statistical approach to numerical pattern recognition [1] the known structures are based on mathematical models, and the usefulness of such methods depends on the availability of sufficiently accurate models of the objects generating the data.

Methods based on clustering and techniques recently developed in the field of computational intelligence such as neural networks, fuzzy logic and genetic algorithms are becoming increasingly popular in the pattern recognition community [2]–[5]. Such methods offer an attractive alternative to statistical approaches as they do not require *a priori* assumptions of statistical models. They are able to learn the mapping of functions and systems, and can perform classification from labeled training data as well as explore structures and classes in unlabeled data.

This article presents a new approach to pattern classification which uses a fuzzy logic relation to establish the correspondence between structures in the feature space and the class identifiers (labels). This approach can effectively deal with classes that cannot be described by a single construct in the feature space. This is especially useful for problems where one does not *a priori* know which features should be selected in order to yield well-separated classes. By using the fuzzy logic relation, one avoids the problem of labeling the prototypes which can be particularly difficult when classes are characterized by partially shared structures or when the training data contains classification errors (typical for subjective classification). This partial sharing of structures among several classes is naturally captured by the fuzzy relation. Moreover, class labels may be fuzzy distributions as well. The fuzzy relation-based classification scheme represents a transparent alternative to conventional black-box techniques like artificial neural networks for complex nonlinear classification problems. The transparency of the relational classifier allows for the analysis of both the trained classifier and of the classification result for unseen patterns.

In the training of the classifier, two steps are distinguished:

1) exploratory data analysis (unsupervised fuzzy clustering);
2) construction of a logical relation between the structures found in the previous step and the class labels.

In the exploratory step, the available data objects are clustered in groups by the fuzzy c-means (FCM) or a similar algorithm. Clustering results in a fuzzy partition matrix, which specifies for each training sample a c-tuple of membership degrees in the obtained clusters. In the second step, a fuzzy relation is computed, using the memberships obtained in the first step and the target membership of the pattern in the classes (which may be crisp or fuzzy). This relation is built by means of the φ-composition (a fuzzy implication) and conjunctive aggregation. It specifies the logical relationship between the cluster membership and the class membership.

To classify new patterns, the membership of each pattern in the clusters (fuzzy prototypes) is computed from its distance to the cluster centers, giving a fuzzy set of prototype membership. Then, relational composition of this fuzzy set with the fuzzy relation is applied to compute an output fuzzy set. This set gives a fuzzy classification in terms of membership degrees of the pattern in the given classes. When a crisp decision is required, defuzzification has to be applied to this fuzzy set. Typically, the maximum defuzzification method is used.

The rest of this paper is organized in three sections. First, the training of the classifier is explained in Section II. The classification of new patterns is described in Section III. A simple example is presented throughout these two sections in order to illustrate the individual steps. Section IV reports

Manuscript received May 12, 1998; revised November 22, 1998. This work was supported in part by the Research Council of Norway. This paper was recommended by Associate Editor L. O. Hall.

The authors are with the Faculty of Information Technology and Systems, Control Laboratory, Delft University of Technology, 2600 GA Delft, The Netherlands (e-mail: m.setnes@its.tudelft.nl).

Publisher Item Identifier S 1083-4419(99)05270-X.

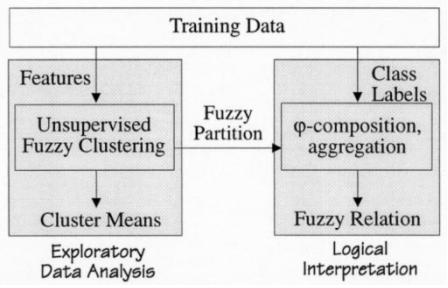


Fig. 1. Training of the relational classifier.

a practical application to the identification of livestock from sound signals. The results obtained with the proposed scheme are compared to the neuro-fuzzy classifier NEFCLASS [4], and to a multireference minimum-distance classifier [6].

II. Training of the Classifier

The training of the classifier proceeds in two main steps. First, exploratory data analysis by means of unsupervised fuzzy clustering is performed in the feature space. Then, a fuzzy relation is built from the obtained fuzzy partition of the feature space and the target vectors (labels) of the training data objects. These two steps are illustrated in Fig. 1 and described in detail below.

A. Exploratory Data Analysis

The aim of the exploratory data analysis is to discover the substructures in the feature space of the available training data. This is done in an unsupervised manner, i.e., class labels are not used. Using this approach, the number of prototype structures is independent of the number of classes. This results in a partition of the feature space that more closely represent the natural structures in the data. These substructures can be discovered by means of unsupervised cluster analysis [6]. In the proposed approach, fuzzy clustering is applied. Most natural phenomena do not lend themselves to crisp classification and the membership of the data samples into the subclasses is often a matter of degree, rather than a yes-or-no decision. Uncertainty and inaccuracy of the features (caused, for instance, by data acquisition) is another reason for using fuzzy rather than crisp clustering. Iterative optimization algorithms such as the fuzzy c-means scheme and its modifications can be applied to obtain the subgroups. A number of cluster validity measures developed for the c-means algorithm can be used in order to assess the goodness of the obtained partition and to estimate the number of subgroups in the data.

Note that other clustering methods than the c-means can be used as well. Examples are the Gustafson–Kessel algorithm [7] or the maximum likelihood clustering method [8], which employ an adaptive distance measure to fit the actual shape and orientation of the cluster to the data. Since the extension of the presented methods to these algorithms is straightforward, in this article, we restrict ourselves to the fuzzy c-means algorithm, which is given in the Appendix.

Let $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\}$ be the set of N training data objects to be classified. Each object is represented by a p-dimensional *feature vector* $\mathbf{x}_k = [x_{1k}, x_{2k}, \cdots, x_{pk}]^T \in \mathbb{R}^p$. A set of N feature vectors in the training data set is represented as a $p \times N$ *feature matrix*:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1N} \\ x_{21} & x_{22} & \cdots & x_{2N} \\ \vdots & \vdots & \vdots & \vdots \\ x_{p1} & x_{p2} & \cdots & x_{pN} \end{bmatrix}. \tag{1}$$

At the exploratory analysis step, the training data set $\mathbf{X}$ is partitioned into c fuzzy subsets (clusters). The membership of the data samples in the clusters is described by the fuzzy partition matrix $\mathbf{U}$ and each cluster is characterized by its center $\mathbf{v}_i$. Prior to clustering, the user must define several parameters: the number of clusters c, the fuzziness exponent m, the termination tolerance and the norm-inducing matrix. These choices are described in the Appendix.

An illustrative example is used throughout the paper to highlight the individual steps. We begin with the exploratory data analysis.

Example II.1: Fig. 2(a) shows four groups of synthetic data. Each group consist of 50 samples normally distributed around a group center with variance according to Table I. The number of classes is 3, and the samples are labeled as belonging to class 1, 2, or 3.

The FCM algorithm was applied to the data several times with values of c and m between the estimated lower and upper bounds $2 \leq c \leq 10$ and $1.3 \leq m \leq 2.5$, respectively. The resulting partitions were evaluated with the Xie–Beni index (16). Fig. 2(b) shows some of the results. The Xie–Beni index detects the correct number of clusters, and has a distinctive minimum at $c = 4$.

B. Computing the Fuzzy Relation

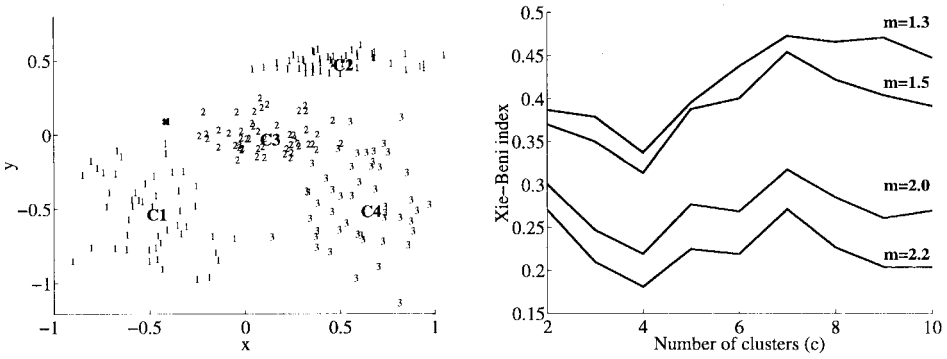
The aim of this step is to compute a fuzzy relation, which will encode the logical relationship between the *cluster membership* and the *class membership*. This relation is computed from the information in the fuzzy partition matrix and in the target vectors containing membership of the pattern in the classes. The kth *target vector* is denoted by

$$\boldsymbol{\omega}_k = [\omega_{1k}, \omega_{2k}, \cdots, \omega_{Lk}]^T \tag{2}$$

where $\omega_{jk} \in [0, 1]$ and $L \in \mathbb{N}$ is the number of classes. For the training data, where the classification is exactly known, $\omega_{jk} \in \{0, 1\}$. The target vector $\boldsymbol{\omega}_k$ is then a fuzzy singleton set, i.e., a vector of all zeros except for a one at the place of the known class index. If the data consist of four classes ($L = 4$), and training data object $\mathbf{x}_k$ belongs to class 3, the target vector is represented as $\boldsymbol{\omega}_k = [0, 0, 1, 0]^T$.

The cluster memberships are directly available in the fuzzy partition matrix. After clustering, for each training sample $\mathbf{x}_k$, the vector of cluster membership degrees μ_{ik} is contained in the kth column of $\mathbf{U}$. We denote this vector by:

$$\boldsymbol{\mu}_k = [\mu_{1k}, \mu_{2k}, \cdots, \mu_{ck}]^T. \tag{3}$$



(a) (b)

Fig. 2. Data belonging to three classes forming (a) four groups in the pattern space and (b) cluster validity measure.

TABLE I
SYNTHETIC DATA WITH THREE CLASSES IN FOUR GROUPS

Group	Class label	Group center (x, y)	Variance (σ_x, σ_y)
1	1	$(-0.5, -0.5)$	$(0.2, 0.2)$
2	1	$(0.5, 0.5)$	$(0.2, 0.05)$
3	2	$(0.1, 0)$	$(0.2, 0.1)$
4	3	$(0.6, -0.4)$	$(0.2, 0.25)$

The binary fuzzy relation, $\mathbf{R}$, is a mapping $\mathbf{R}$: $[0, 1]^c \times [0, 1]^L \rightarrow [0, 1]$. It can be represented as a $c \times L$ matrix

$$\mathbf{R} = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1L} \\ r_{21} & r_{22} & \cdots & r_{2L} \\ \vdots & \vdots & \vdots & \vdots \\ r_{c1} & r_{c2} & \cdots & r_{cL} \end{bmatrix}. \tag{4}$$

The relation $\mathbf{R}$ is obtained by aggregating the partial relations $\mathbf{R}_k$, computed for the training samples. Each $\mathbf{R}_k$ is obtained by the φ-composition operator [9]. We choose this operator to be the Łukasiewicz implication:

$$(r_{ij})_k = \min(1, 1 - \mu_{ik} + \omega_{jk}), \quad j = 1, 2, \cdots, L, \; i = 1, 2, \cdots, c. \tag{5}$$

As this is a generalization of the classical implication [10], the aggregation of the relations R_k is computed by means of a fuzzy conjunction operator

$$\mathbf{R} = \bigcap_{k=1}^{N} \mathbf{R}_k \tag{6}$$

implemented element-wise by the minimum function:

$$r_{ij} = \min_{k=1, 2, \cdots, N} [(r_{ij})_k]. \tag{7}$$

Other types of residuated fuzzy implications could be chosen as well, but empirical results have shown that the above method performs well. Some authors also use t-norms (such as minimum or product) instead of implications [9]. This is, in principle possible, but the interpretation of the trained relation is not based on logic anymore (it is rather a kind of correlation matrix). In the crisp case (when using nonfuzzy clustering), the fuzzy implication reduces to classical implication, which is another advantage.

Example II-2: Consider the data from Example II-1. A fuzzy partition is obtained by FCM with $c = 4$ and $m = 2$. Fig. 2(a) shows the cluster centers C_1, C_2, C_3, and C_4. Using the approach described above, a fuzzy relation is calculated that relates the cluster membership, $\boldsymbol{\mu}_k = [\mu_{1k}, \cdots, \mu_{4k}]^T$, to the class membership $\boldsymbol{\omega}_k = [\omega_{1k}, \cdots, \omega_{3k}]^T$. The relation is given by the following 4×3 matrix:

$$\mathbf{R} = \begin{bmatrix} 0.7499 & 0.0095 & 0.0095 \\ 0.5255 & 0.0004 & 0.0004 \\ 0.0010 & 0.1769 & 0.0010 \\ 0.0404 & 0.0404 & 0.6922 \end{bmatrix}. \tag{8}$$

From the relation it is seen that there is strong evidence for, e.g., correspondence between membership in clusters 1 and 2, and class label 1.

III. CLASSIFICATION OF NEW PATTERNS

The classification of new patterns proceeds in three steps. First, the cluster membership $\hat{\mu}$ is computed from the distance to the cluster centers. Then, the fuzzy relational composition is applied to compute the vector of class memberships $\hat{\omega}$. Finally, defuzzification is applied to obtain a crisp decision. These steps are illustrated in Fig. 3, and described in more detail below.

A. Calculation of Cluster Membership

For a new feature vector $\mathbf{x}$, the vector of cluster membership degrees, $\hat{\mu} = [\hat{\mu}_1, \cdots, \hat{\mu}_c]^T$, is computed by measuring the similarity of the feature vector $\mathbf{x}$ to each of the cluster prototypes $\mathbf{v}_i$:

$$\hat{\mu}_i = S(\mathbf{x}, \mathbf{v}_i). \tag{9}$$

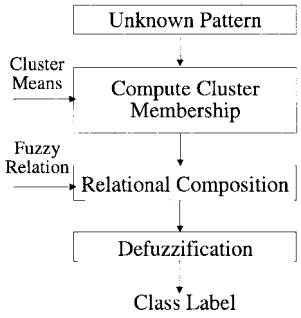


Fig. 3. Classification of new data.

The similarity S is computed using the distance of $\mathbf{x}$ from $\mathbf{v}_i$

$$\hat{\mu}_i = \frac{1}{\sum_{j=1}^{c} (d(\mathbf{x}, \mathbf{v}_i)/d(\mathbf{x}, \mathbf{v}_j))^{2/(m-1)}} \tag{10}$$

where the distance measure $d(\mathbf{x}, \mathbf{v}_i)$ is the same distance measure as used by the FCM algorithm (see Appendix). Note that the membership degrees in (10) are computed relative to each other, and the sum of the membership degrees equals 1.

B. Relational Composition

Given the cluster membership $\hat{\mu}$, the class membership vector $\hat{\omega}$ can be computed by fuzzy relational composition

$$\hat{\omega} = \hat{\mu} \circ_{\mathrm{T}} \mathbf{R}, \tag{11}$$

where $\circ_{\mathrm{T}}$ is the sup-t composition operator [11]. The t-norm corresponds to the implication used in the φ-composition. For the Łukasiewicz implication, the Łukasiewicz t-norm is used, which is given by $\max(a + b - 1, 0)$, where $a, b \in [0, 1]$. In our case, the involved domains are discrete and hence the supremum is replaced by maximum, yielding

$$\hat{\omega}_j = \max_{1 \le i \le c} [\max(\hat{\mu}_i + r_{ij} - 1, 0)], \qquad j = 1, 2, \cdots, L. \tag{12}$$

C. Defuzzification

The fuzzy set $\hat{\omega}$ containing the class membership degrees is eventually defuzzified, using the maximum method

$$\hat{y} = \arg \max_{1 \le j \le L} \hat{\omega}_j \tag{13}$$

where $\hat{y}$ is the class index. If multiple j satisfy (13), the classification is seen as undecidable (reject decision). The value of the maximum membership degree

$$\hat{\omega}_{\max} = \max_{1 \le j \le c} (\hat{\omega}_j) \tag{14}$$

gives an indication as to whether the decision cannot be made because of a conflict or because of insufficient information in the training data set. Low values of $\hat{\omega}_{\max}$ indicate that a conflict occurred in the training data set [note that the intersection of the partial relations (5) is the minimum operator]. Overall high values of the fuzzy set of class memberships $\hat{\omega}$, on the other hand, indicate that insufficient evidence for the classification is available from the training data set (all outcomes are possible to a high degree). Existence of multiple maxima in the output fuzzy set leads to a reject decision, indicating a possible logical conflict in the training data (due to for instance noise in the features, misclassifications of training samples or inappropriate feature selection) or alternatively absence of information in a particular region of the feature space (no or very few samples available in the training set for the particular feature values).

Example III-1: Consider the relation from Example II-2 and the pattern $\mathbf{x} = [-0.4, 0.09]^T$ to be classified. This pattern is marked with an "x" in Fig. 2(a) and belongs to class 1. By (10), the cluster memberships are computed: $\hat{\mu} = [0.3027, 0.1336, 0.4772, 0.0865]^T$. The composition of the cluster memberships $\hat{\mu}$ with the relation (8), given by (12), yields the class membership vector

$$\begin{aligned} \hat{\omega}^T &= [0.3027, 0.1336, 0.4772, 0.0865] \circ_{\mathrm{T}} \\ &\quad \cdot \begin{bmatrix} 0.7499 & 0.0095 & 0.0095 \\ 0.5255 & 0.0004 & 0.0004 \\ 0.0010 & 0.1769 & 0.0010 \\ 0.0404 & 0.0404 & 0.6922 \end{bmatrix} \\ \hat{\omega}^T &= [0.053, 0, 0]. \end{aligned} \tag{15}$$

Defuzzification by (13) assigns the class index 1 to this data object. Note that applying a multireference minimum distance classifier, based on the cluster centers, would misclassify this data object as belonging to class 2. This is because the object has the highest membership in the fuzzy partitioning cluster C_3, which is a reference for class 2 in this case.

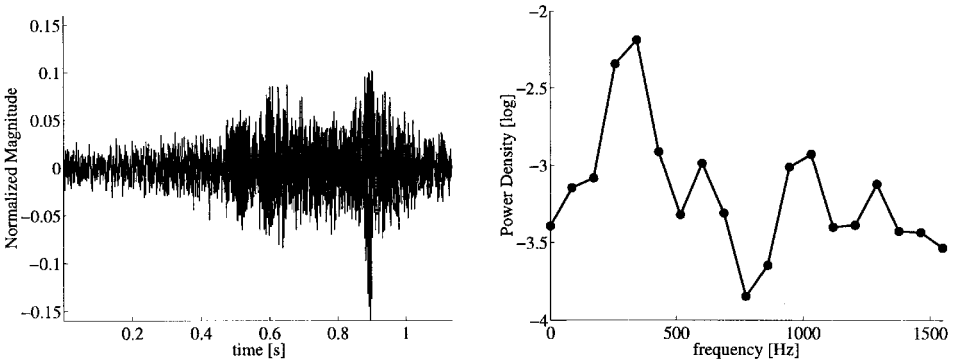
IV. Application to the Recognition of Livestock from Sound Sequences

Increasing level of automation in agriculture offers a wide variety of interesting applications for pattern recognition and classification. This section deals with the identification of animals on the basis of sounds. The data used in this section consists of a training set of 68 sound sequences recorded from four different cows with known labels, and a data set of 31 sequences with "unknown" labels. The training data contains 26, 13, 6, and 23 sequences from cows 1–4, respectively. The length of the sound sequences varies between 12515–42 054 samples, taken at the sampling rate of 11 025 Hz.

A. Feature Selection

Fig. 4(a) shows as example of a cow sound sequence. Two main approaches in the classification of sounds can be considered the *envelope* and the *frequency spectrum* [12]. The latter is more informative and more likely to give good results for the recognition of individual cows. The envelope approach is often used for the classification of different *types* of sounds, e.g., different musical instruments.

For each sound sequence, the power spectrum densities (PSD) are calculated for frequencies <1500 Hz, using the MATLAB PSD routine with 128 points and a nonoverlapping



(a) (b)

Fig. 4. (a) The data set consists of recorded sound sequences. (b) Logarithms of the PSD estimates ($\leq$1500 Hz) are used as feature vectors.

TABLE II
STATISTICAL EVALUATION

Method	Training data		Evaluation data		Cost		
	Mean error	Mean reject	Mean error	Mean reject	Best	Worst	Mean
FRC	0	3.42	2.16	2.42	2	24	8.7
MMDC	2.28	5.88	2.58	7.28	16	30	23.52

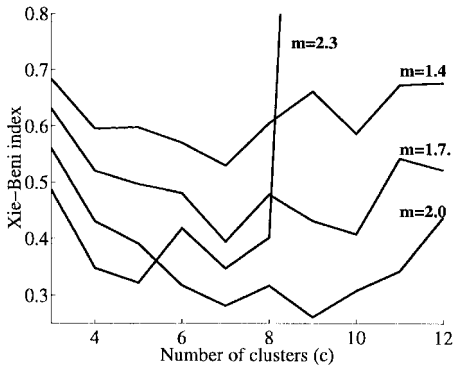


Fig. 5. The Xie–Beni validity index for different values of c and m.

Hanning window [13]. Each sound sequence is now represented by an 18×1 vector of PSD estimates. As features we take the logarithm of the PSD estimates, illustrated in Fig. 4(b). From the training data we get the pattern matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_{68}]$, where $\mathbf{x}_k = [x_{1k}, x_{2k}, \cdots, x_{18k}]^T$ is the feature vector of sample k. The data is clustered by the FCM algorithm for a range of values for c and m ($4 \leq c \leq 20$, $1.1 \leq m \leq 3$). The Xie–Beni index is applied to the resulting partitions, and from Fig. 5 we see that a first distinctive local minimum is located at $c = 7$. From this analysis, the partition obtained with $c = 7$ and $m = 2$ is selected.

B. Statistical Evaluation

For the purpose of evaluating the approach, the training data set was split at random into a new training data set consisting of about half of the training samples available for each cow (total of 35 samples) and a test set containing the other half of the training samples (total of 33 samples). This was done 50 times, and each time a fuzzy relational classifier (FRC) and a multireference minimum distance classifier (MMDC) was trained and evaluated. The MMDC uses the obtained cluster centers as references with labels calculated from the training data using weighted votes [6]. Results from maximum likelihood classification, with costs 2 and 1 for error and reject, respectively, are presented in Table II. The results achieved with FRC are overall better than the results of MMDC.

The costs reported in Table II, are calculated as Cost $= C_e * (\text{number of errors}) + C_r * (\text{number of reject})$, where C_e is the cost assigned to a misclassification, and C_r is the cost of a reject. In this application, the costs are selected as $C_e = 2$ and $C_r = 1$. The costs are calculated for both training and evaluation data. It can be concluded that the best FRC is far better than the best MMDC. It is also interesting to note that the FRC always learns the given training set (zero error) without overtraining. This is due to the fuzzy relation which enables the FRC to learn the correct classifications of samples that are geometrically closer to another group (cluster) than the majority of the samples from their own class.

C. Classification

For the classification of the 31 unknown samples, three classifiers where trained with all the available training data: the NEFCLASS neuro-fuzzy classifier [4], the MMDC and the FRC. NEFCLASS was trained with seven rules and three fuzzy sets in the domain of each feature. For the two other classifiers, the settings were as above. The results are given in Table III. Note that the FRC outperforms both the MMDC and NEFCLASS.

TABLE III
CLASSIFICATION RESULTS

Method	Training data		Evaluation data		Cost
	Error	Reject	Error	Reject	
FRC	0	8	1	4	14
MMDC	6	14	3	9	41
NEFCLASS	5	13	3	11	40

V. CONCLUSION

A new approach to nonlinear pattern classification has been presented. The classifier is constructed in two steps. In the first step, exploration of structures in the feature space is performed by a fuzzy clustering algorithm. This step is unsupervised which means that the class labels are not used. The resulting partition may consist of more clusters than the actual number of classes in the data. This allows for the existence of multiple and partially shared structures in the feature space. In the second step, a fuzzy relation is built that encodes the logical relationship between the cluster membership and the class membership (class labels). This is done by φ-composition using the Łukasiewicz fuzzy implication.

The trained classifier is determined by the cluster centers in the feature space, the fuzzy relation, and a distance measure. For unseen data, a fuzzy set of cluster memberships is calculated by using the distance to the cluster centers. A fuzzy set of class membership is obtained by relational composition with the trained relation. The membership values give an indication about the quality of the decision. In general, if all membership degrees are close to zero, conflicts occurred in the training data. If all membership degrees are close to one, on the other hand, this indicate insufficient evidence for classification. The same reasoning holds for the relation itself, and the result of training can be evaluated without the use of data by inspecting the obtained fuzzy relation. Applying a fuzzy relation for classification will in general decrease the number of misclassifications. However, if no reasonable clusters can be found in the data, the number of unclassified patterns (reject decisions) will increase.

The proposed approach has been applied to the identification of livestock from recorded sound sequences. For the used data set, the fuzzy relational classifier is found to outperform two other methods tested on the same problem: the NEFCLASS neuro-fuzzy classifier and a multireference minimum distance classifier. As opposed to (fuzzy) neural network based approached, the fuzzy relational classifier is not sensitive to the order in which the training examples are presented. Further, the problem of selecting an appropriate number of training epochs and the risk of overtraining are eliminated. Also, the obtained classifier is more transparent to the user. The main difference with regard to the minimum distance classification is that the classification is not necessarily that of the nearest reference. Similarly to the minimum distance classification, the relational classifier is sensitive to the result of clustering.

APPENDIX
THE FUZZY c-MEANS ALGORITHM

In this appendix, the fuzzy c-means algorithm is presented and the choice of its parameters is discussed.

Given the data set $\mathbf{X}$, choose the number of clusters $1 < c < N$, the weighting exponent $m > 1$, the termination tolerance $\epsilon > 0$ and the norm-inducing matrix $\mathbf{A}$.
Initialize the partition matrix $\mathbf{U}^{(0)}$ randomly.

Repeat for $l = 1, 2, \cdots$

Step 1: Compute the cluster prototypes (means):

$$\mathbf{v}_i^{(l)} = \frac{\sum_{k=1}^{N} (\mu_{ik}^{(l-1)})^m \mathbf{x}_k}{\sum_{k=1}^{N} (\mu_{ik}^{(l-1)})^m}, \qquad 1 \le i \le c.$$

Step 2: Compute the distances:

$$d_{ik\mathbf{A}}^2 = (\mathbf{x}_k - \mathbf{v}_i^{(l)})^T \mathbf{A} (\mathbf{x}_k - \mathbf{v}_i^{(l)}), \quad 1 \le i \le c, \quad 1 \le k \le N.$$

Step 3: Update the partition matrix:
if $d_{ik\mathbf{A}} > 0$ for $1 \le i \le c, \quad 1 \le k \le N$,

$$\mu_{ik}^{(l)} = \frac{1}{\sum_{j=1}^{c} (d_{ik\mathbf{A}}/d_{jk\mathbf{A}})^{2/(m-1)}},$$

otherwise

$$\mu_{ik}^{(l)} = 0 \text{ if } d_{ik\mathbf{A}} = 0,$$

and

$$\mu_{ik}^{(l)} \in [0, 1] \text{ with } \sum_{i=1}^{c} \mu_{ik}^{(l)} = 1.$$

until $\|\mathbf{U}^{(l)} - \mathbf{U}^{(l-1)}\| < \epsilon$.

The following parameters must be specified.

- The number of clusters c is the most important parameter. Optimally, c would equal the (unknown) number of subgroups present in the data, in which case the chance is high that the underlying structure of the data will be detected. The choice of c can be verified by assessing the validity of the obtained partition, by using validity measures [2], [3], [8], and [14]. In this article, we use the Xie–Beni index [15] which has proven suitable for other classification problems [14]

$$\chi(U, \mathbf{V}; \mathbf{X}) = \frac{\sum_{i=1}^{c} \sum_{k=1}^{N} \mu_{ik}^m \|\mathbf{x}_k - \mathbf{v}_i\|^2}{N \left(\min_{i \ne j} \{\|\mathbf{v}_i - \mathbf{v}_j\|^2\} \right)}. \tag{16}$$

The best partition is the partition that minimizes the value of $\chi(U, \mathbf{V}; \mathbf{X})$. Cluster validity analysis is performed by running the clustering algorithm for different values of c and m, and then several times for each of these settings with a different initialization $\mathbf{U}^{(0)}$. The validity measure is calculated for each run, and the number of clusters which minimizes the measure is chosen as the "correct" number of clusters in the data.

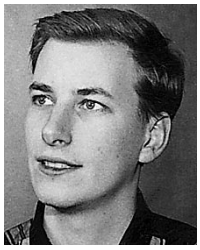
- The weighting exponent m is of a secondary importance, even though it has a significant influence on the shape of the clusters. As m approaches one from above, the partition becomes hard ($\mu_{ik} \in \{0, 1\}$) and $\mathbf{v}_i$ are ordinary means of the clusters. As $m \to \infty$, the partition becomes maximally "fuzzy" ($\mu_{ik} = 1/c$) and the cluster means are all equal to the grand mean of the data. Typically, $m = 2$ is used as a default value, but cluster validity analysis can be applied to search for an alternative value of m.
- The termination criterion is usually set to $\epsilon = 0.001$, but $\epsilon = 0.01$ is often sufficient. Without any prior knowledge about the data distribution and dependencies among the features, the norm-inducing matrix is chosen as $\mathbf{A} = \mathbf{I}$. Alternatively, adaptive distance clustering can be applied [7].
- The choice of the initial partition matrix $\mathbf{U}^{(0)}$ may influence the result of clustering, as the c-means algorithm is only guaranteed to converge to a local optimum. Thus, for a given choice of the remaining parameters, clustering is typically repeated for different $\mathbf{U}^{(0)}$ and a validity measure is applied to choose the best partition.

ACKNOWLEDGMENT

The sound data used in Section IV was supplied by the European Network in Uncertainty Techniques Developments for Use in Information Technology (ERUDIT) as part of the *1996 International Competition for Signal Analysis and Processing by Intelligent Techniques*. The authors would like to thank the anonymous reviewers for their constructive comments.

REFERENCES

[1] K. Fukunaga, *Introduction to Statistical Pattern Recognition.* New York: Academic, 1972.
[2] E. Backer, *Computer-Assisted Reasoning in Cluster Analysis.* Englewood Cliffs, NJ: Prentice-Hall, 1995.
[3] J. C. Bezdek, *Pattern Recognition With Fuzzy Objective Functions.* New York: Plenum, 1981.
[4] D. Nauck and R. Kruse, "Nefclass—A neuro-fuzzy approach for classification of data," in *Proc. 1995 ACM Symp. Applied Computing*, Nashville, TN, 1995, pp. 461–465.
[5] E. A. Wan, "Neural network classification: A Bayesian interpretation," *IEEE Trans. Neural Networks*, vol. 1, pp. 303–304, 1990.
[6] J. Schürmann, *Pattern Classification. A Unified View of Statistical and Neural Approaches.* New York: Wiley, 1996.
[7] D. E. Gustafson and W. C. Kessel, "Fuzzy clustering with a fuzzy covariance matrix," in *Proc. IEEE CDC*, San Diego, CA, 1979, pp. 761–766.
[8] I. Gath and A. B. Geva, "Unsupervised optimal fuzzy clustering," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 11, pp. 773–781, July 1989.
[9] W. Pedrycz, "Reasoning by analogy in fuzzy controllers," *Fuzzy Control Systems*, Kandel and Langholz, Eds. Boca Raton, FL: CRC, 1994, pp. 55–74.
[10] R. Giles, "Łukasiewicz logic and fuzzy set theory," *Int. J. Man–Mach. Stud.*, vol. 8, pp. 313–327, 1976.
[11] G. J. Klir and B. Youan, *Fuzzy Sets and Fuzzy Logic: Theory and Applications.* Englewood Cliffs, NJ: Prentice-Hall, 1995.
[12] R. A. Roberts and C. T. Mullis, *Digital Signal Processing.* New York: Addison-Wesley, 1993.
[13] *MATLAB User's Guide*, The MathWorks Inc., Natick, MA, 4.2 ed., 1994.
[14] N. R. Pal and J. C. Bezdek, "On cluster validity for the fuzzy c-means model," *IEEE Trans. Fuzzy Syst.*, vol. 3, no. 3, pp. 370–379, 1995.
[15] X. L. Xie and G. A. Beni, "Validity measure for fuzzy clustering," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 3, pp. 841–846, Aug. 1991.



Magne Setnes (S'98) was born in 1970 in Bergen, Norway. He received the B.Sc. degree in robotics from the Kongsberg College of Engineering, Norway, in 1992, and the M.Sc. degree in electrical engineering from the Delft University of Technology, Delft, The Netherlands, in 1995. He is currently pursuing the Ph.D. degree at the Control Laboratory, Delft University of Technology.

His research interests include fuzzy sets, fuzzy logic, and neuro-fuzzy systems for modeling, control, and decision making.

Robert Babuška was born in 1967 in Prague, Czechoslovakia. He received the M.Sc. degree in control engineering from the Czech Technical University, Prague, in 1990, and the Ph.D. degree from the Delft University of Technology, Delft, The Netherlands, in 1997.

He is currently an Associate Professor with the Control Laboratory, Electrical Engineering Department, Delft University of Technology. His main research interests include fuzzy set theory, fuzzy systems modeling, identification, and control.