Available at www.ComputerScienceWeb.com
POWERED BY SCIENCE DIRECT®

Pattern Recognition Letters 24 (2003) 2195–2207

Pattern Recognition Letters

www.elsevier.com/locate/patrec

Supervised fuzzy clustering for the identification of fuzzy classifiers

Janos Abonyi *, Ferenc Szeifert

Department of Process Engineering, University of Veszprem, P.O. Box 158, H-8201 Veszprem, Hungary

Received 9 July 2001; received in revised form 26 August 2002

Abstract

The classical fuzzy classifier consists of rules each one describing one of the classes. In this paper a new fuzzy model structure is proposed where each rule can represent more than one classes with different probabilities. The obtained classifier can be considered as an extension of the quadratic Bayes classifier that utilizes mixture of models for estimating the class conditional densities. A supervised clustering algorithm has been worked out for the identification of this fuzzy model. The relevant input variables of the fuzzy classifier have been selected based on the analysis of the clusters by Fisher's interclass separability criteria. This new approach is applied to the well-known wine and Wisconsin breast cancer classification problems.

Keywords: Fuzzy clustering; Bayes classifier; Rule-reduction; Transparency and interpretability of fuzzy classifiers

1. Introduction

Typical fuzzy classifiers consist of interpretable if–then rules with fuzzy antecedents and class labels in the consequent part. The antecedents (if-parts) of the rules partition the input space into a number of fuzzy regions by fuzzy sets, while the consequents (then-parts) describe the output of the classifier in these regions. Fuzzy logic improves rule-based classifiers by allowing the use of overlapping class definitions and improves the interpretability of the results by providing more insight into the decision making process. Fuzzy logic, however, is not a guarantee for interpretability, as was also recognized in (Valente de Oliveira, 1999; Setnes et al., 1998). Hence, real effort must be made to keep the resulting rule-base transparent.

The automatic determination of compact fuzzy classifiers rules from data has been approached by several different techniques: neuro-fuzzy methods (Nauck and Kruse, 1999), genetic-algorithm (GA)-based rule selection (Ishibuchi et al., 1999), and fuzzy clustering in combination with GA-optimization (Roubos and Setnes, 2000). Generally, the bottleneck of the data-driven identification of

* Corresponding author. Tel.: +36-88-422-0224290; fax: +36-88-422-0224171.
E-mail address: abonyij@fmt.vein.hu (J. Abonyi).
URL: http://www.fmt.vein.hu/softcomp.

0167-8655/03/$ - see front matter
doi:10.1016/S0167-8655(03)00047-3

fuzzy systems is the structure identification that requires non-linear optimization. Thus for high-dimensional problems, the initialization the fuzzy model becomes very significant. Common initializations methods such as grid-type partitioning (Ishibuchi et al., 1999) and *rule generation on extrema* initialization, result in complex and non-interpretable initial models and the rule-base simplification and reduction steps become computationally demanding. To avoid these problems, fuzzy clustering algorithms (Setnes and Babuška, 1999) were put forward. However, the obtained membership values have to be projected onto the input variables and approximated by parameterized membership functions that deteriorates the performance of the classifier. This decomposition error can be reduced by using eigenvector projection (Kim et al., 1998), but the obtained linearly transformed input variables do not allow the interpretation of the model. To avoid the projection error and maintain the interpretability of the model, the proposed approach is based on the Gath–Geva (GG) clustering algorithm (Gath and Geva, 1989) instead of the widely used Gustafson–Kessel (GK) algorithm (Gustafson and Kessel, 1979), because the simplified version of GG clustering allows the direct identification of fuzzy models with exponential membership functions (Hoppner et al., 1999).

Neither GG nor GK algorithm does not utilize the class labels. Hence, they give suboptimal result if the obtained clusters are directly used to formulate a classical fuzzy classifier. Hence, there is a need for fine-tuning of the model. This GA or gradient-based fine-tuning, however, can result in overfitting and thus poor generalization of the identified model. Unfortunately, the severe computational requirements of these approaches limit their applicability as a rapid model-development tool.

This paper focuses on the design of interpretable fuzzy rule-based classifiers from data with low-human intervention and low-computational complexity. Hence, a new modeling scheme is introduced based only on fuzzy clustering. The proposed algorithm uses the class label of each point to identify the optimal set of clusters that describe the data. The obtained clusters are then used to build a fuzzy classifier.

The contribution of this approach is twofold.

- The classical fuzzy classifier consists of rules each one describing one of the C classes. In this paper a new fuzzy model structure is proposed where the consequent part is defined as the probabilities that a given rule represents the $c_1, \ldots, c_C$ classes. The novelty of this new model is that one rule can represent more than one classes with different probabilities.
- Classical fuzzy clustering algorithms are used to estimate the distribution of the data. Hence, they do not utilize the class label of each data point available for the identification. Furthermore, the obtained clusters cannot be directly used to build the classifier. In this paper a new cluster prototype and the related clustering algorithm have been introduced that allows the direct supervised identification of fuzzy classifiers.

The proposed algorithm is similar to the multi-prototype classifier technique (Biem et al., 2001; Rahman and Fairhurst, 1997). In this approach, each class is clustered independently from the other classes, and is modeled by few components (Gaussian in general). The main difference of this approach is that each cluster represents different classes, and the number of clusters used to approximate a given class have to be determined manually, while the proposed approach does not suffer from these problems.

Using too many input variables may result in difficulties in the prediction and interpretability capabilities of the classifier. Hence, the selection of the relevant features is usually necessary. Generally, there is a very large set of possible features to compose feature vectors of classifiers. As ideally the training set size should increase exponentially with the feature vector size, it is desired to choose a minimal subset among it. Some generic tips to choose a good feature set include the facts that they should discriminate as much as possible the pattern classes and they should not be correlated/redundant. There are two basic feature-selection approaches: The *closed-loop* algorithms are based

on the classification results, while the *open-loop* algorithms are based on a distance between clusters. In the former, each possible feature subset is used to train and to test a classifier, and the recognition rates are used as a decision criterion: the higher the recognition rate, the better is the feature subset. The main disadvantage of this approach is that choosing a classifier is a critical problem on its own, and that the final selected subset clearly depends on the classifier. On the other hand, the latter depends on defining a distance between the clusters, and some possibilities are Mahalanobis, Bhattacharyya and the class separation distance (Campos and Bloch, 2001).

In this paper the Fisher-interclass separability method is utilized, which is an open-loop feature selection approach (Cios et al., 1998). Other papers focused on feature selection based on similarity analysis of the fuzzy sets (Campos and Bloch, 2001; Roubos and Setnes, 2000). Differences in these reduction methods are: (i) Feature reduction based on the similarity analysis of fuzzy sets results in a closed-loop feature selection because it depends on the actual model while the applied open-loop feature selection can be used beforehand as it is independent from the model. (ii) In similarity analysis, a feature can be removed from individual rules. In the interclass separability method the feature is omitted in all the rules (Roubos et al., 2001). In this paper the simple Fisher interclass separability method have been modified, but in the future advanced multiclass data reduction algorithms like weighted pairwise Fisher criteria (Loog et al., 2001) could be also used.

The paper is organized as follows. In Section 2, the structure of the new fuzzy classifier is presented. Section 3 describes the developed clustering algorithm that allows for the direct identification of fuzzy classifiers. For the selection of the important features of the fuzzy system a Fisher interclass separability criteria based method will be presented in Section 4. The proposed approach is studied for the Wisconsin breast cancer and the wine classification examples in Section 5. Finally, the conclusions are given in Section 6.

2. Structure of the fuzzy rule-based classifier

2.1. Classical Bayes classifier

The identification of a classifier system means the construction of a model that predicts the class $y_k = \{c_1, \ldots, c_C\}$ to which pattern $\boldsymbol{x}_k = [x_{1,k}, \ldots, x_{n,k}]$ should be assigned. The classic approach for this problem with C classes is based on Bayes' rule. The probability of making an error when classifying an example $\boldsymbol{x}$ is minimized by Bayes' decision rule of assigning it to the class with the largest a posteriori probability:

$$\boldsymbol{x} \text{ is assigned to } c_i \Longleftrightarrow p(c_i|\boldsymbol{x}) \geqslant p(c_j|\boldsymbol{x}) \quad \forall j \neq i \tag{1}$$

The a posteriori probability of each class given a pattern $\boldsymbol{x}$ can be calculated based on the $p(\boldsymbol{x}|c_i)$ class conditional distribution, which models the density of the data belonging to the class c_i, and the $P(c_i)$ class prior, which represents the probability that an arbitrary example out of data belongs to class c_i

$$p(c_i|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|c_i)P(c_i)}{p(\boldsymbol{x})} = \frac{p(\boldsymbol{x}|c_i)P(c_i)}{\sum_{j=1}^{C} p(\boldsymbol{x}|c_j)P(c_j)} \tag{2}$$

As (1) can be rewritten using the numerator of (2)

$$\boldsymbol{x} \text{ is assigned to } c_i \Longleftrightarrow p(\boldsymbol{x}|c_i)P(c_i) \geqslant p(\boldsymbol{x}|c_j)P(c_j) \quad \forall j \neq i \tag{3}$$

we would have an optimal classifier if we would perfectly estimate the class priors and the class conditional densities.

In practice one needs to find approximate estimates of these quantities on a finite set of training data $\{\boldsymbol{x}_k, y_k\}$, $k = 1, \ldots, N$. Priors $P(c_i)$ are often estimated on the basis of the training set as the proportion of samples of class c_i or using prior knowledge. The $p(c_i|\boldsymbol{x})$ class conditional densities can be modeled with non-parametric methods like histograms, nearest-neighbors or parametric methods such as mixture models.

A special case of Bayes classifiers is the quadratic classifier, where the $p(\boldsymbol{x}|c_i)$ distribution generated by the class c_i is represented by a Gaussian function

$$p(\boldsymbol{x}|c_i) = \frac{1}{|2\pi \boldsymbol{F}_i|^{n/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{v}_i)^{\mathrm{T}}(\boldsymbol{F}_i)^{-1}(\boldsymbol{x}-\boldsymbol{v}_i)\right) \tag{4}$$

where $\boldsymbol{v}_i = [v_{1,i}, \ldots, v_{n,i}]^{\mathrm{T}}$ denotes the center of the ith multivariate Gaussian and $\boldsymbol{F}_i$ stands for a covariance matrix of the data of the class c_i. In this case, the (3) classification rule can be reformulated based on a distance measure. The sample $\boldsymbol{x}_k$ is classified to the class that minimizes the $D^2_{i,k}(\boldsymbol{x}_k)$ distance, where the distance measure is inversely proportional to the probability of the data:

$$D^2_{i,k}(\boldsymbol{x}_k) = \left(\frac{P(c_i)}{|2\pi \boldsymbol{F}_i|^{n/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{v}_i)^{\mathrm{T}}(\boldsymbol{F}_i)^{-1}(\boldsymbol{x}-\boldsymbol{v}_i)\right)\right)^{-1} \tag{5}$$

2.2. Classical fuzzy classifier

The classical fuzzy rule-based classifier consists of fuzzy rules each one describing one of the C classes. The rule antecedent defines the operating region of the rule in the n-dimensional feature space and the rule consequent is a crisp (non-fuzzy) class label from the $\{c_1, \ldots, c_C\}$ label set:

$$r_i: \quad \text{If } x_1 \text{ is } A_{i,1}(x_{1,k}) \text{ and } \ldots x_n \text{ is } A_{i,n}(x_{n,k}) \text{ then } \hat{y} = c_i, [w_i] \tag{6}$$

where $A_{i,1}, \ldots, A_{i,n}$ are the antecedent fuzzy sets and w_i is a certainty factor that represents the desired impact of the rule. The value of w_i is usually chosen by the designer of the fuzzy system according to his or her belief in the accuracy of the rule. When such knowledge is not available, w_i is fixed to value 1 for any i.

The *and* connective is modeled by the product operator allowing for interaction between the propositions in the antecedent. Hence, the degree of activation of the ith rule is calculated as:

$$\beta_i(\boldsymbol{x}_k) = w_i \prod_{j=1}^{n} A_{i,j}(x_{j,k}) \tag{7}$$

The output of the classical fuzzy classifier is determined by the *winner takes all* strategy, i.e. the output is the class related to the consequent of the rule that gets the highest degree of activation:

$$\hat{y}_k = c_{i^*}, \quad i^* = \arg\max_{1 \leqslant i \leqslant C} \beta_i(\boldsymbol{x}_k) \tag{8}$$

To represent the $A_{i,j}(x_{j,k})$ fuzzy set, we use Gaussian membership functions

$$A_{i,j}(x_{j,k}) = \exp\left(-\frac{1}{2}\frac{(x_{j,k}-v_{j,i})^2}{\sigma^2_{j,i}}\right) \tag{9}$$

where $v_{i,j}$ represents the center and $\sigma^2_{j,i}$ stands for the variance of the Gaussian function. The use of Gaussian membership function allows for the compact formulation of (7):

$$\beta_i(\boldsymbol{x}_k) = w_i \boldsymbol{A}_i(\boldsymbol{x}_k) = w_i \exp\left(-\frac{1}{2}(\boldsymbol{x}_k-\boldsymbol{v}_i)^{\mathrm{T}}(\boldsymbol{F}_i)^{-1}(\boldsymbol{x}_k-\boldsymbol{v}_i)\right) \tag{10}$$

where $\boldsymbol{v}_i = [v_{1,i}, \ldots, v_{n,i}]^{\mathrm{T}}$ denotes the center of the ith multivariate Gaussian and $\boldsymbol{F}_i$ stands for a diagonal matrix that contains the $\sigma^2_{i,j}$ variances.

The fuzzy classifier defined by the previous equations is in fact a quadratic Bayes classifier when $\boldsymbol{F}_i$ in (4) contains only diagonal elements (variances). (For more details, refer to the paper of Baraldi and Blonda (1999), which overviews this issue.)

In this case, the $\boldsymbol{A}_i(\boldsymbol{x})$ membership functions and the w_i certainty factors can be calculated from the parameters of the Bayes classifier following Eqs. (4) and (10) as

$$\boldsymbol{A}_i(\boldsymbol{x}) = p(\boldsymbol{x}|c_i)|2\pi \boldsymbol{F}_i|^{n/2}, \quad w_i = \frac{P(c_i)}{|2\pi \boldsymbol{F}_i|^{n/2}} \tag{11}$$

2.3. Bayes classifier based on mixture of density models

One of the possible extensions of the classical quadratic Bayes classifier is to use mixture of models for estimating the class-conditional densities. The usage of mixture models in Bayes classifiers is not so widespread (Kambhatala, 1996). In these solutions each conditional density is modeled by a separate mixture of models. A possible criticism of such Bayes classifiers is that in a sense they

are modeling too much: for each class many aspects of the data are modeled which may or may not play a role in discriminating between the classes.

In this paper a new approach is presented. The $p(c_i|\boldsymbol{x})$ posteriori densities are modeled by $R > C$ mixture of models (clusters)

$$p(c_i|\boldsymbol{x}) = \sum_{l=1}^{R} p(r_l|\boldsymbol{x})P(c_i|r_l) \tag{12}$$

where $p(r_l|\boldsymbol{x})$ represents the a posteriori probability of $\boldsymbol{x}$ has been generated by the r_lth local model and $P(c_i|r_l)$ denotes the *prior* probability of this model represents the class c_i.

Similarly to (2) $p(r_l|\boldsymbol{x})$ can be written as

$$p(r_i|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|r_i)P(r_i)}{p(\boldsymbol{x})} = \frac{p(\boldsymbol{x}|r_i)P(r_i)}{\sum_{j=1}^{R} p(\boldsymbol{x}|r_j)P(r_j)} \tag{13}$$

By using this mixture of density models the posteriori class probability can be expressed following Eqs. (2), (12) and (13) as

$$\begin{aligned} p(c_i|\boldsymbol{x}) &= \frac{p(\boldsymbol{x}|c_i)P(c_i)}{p(\boldsymbol{x})} \\ &= \sum_{l=1}^{R} \frac{p(\boldsymbol{x}|r_i)P(r_i)}{\sum_{j=1}^{R} p(\boldsymbol{x}|r_j)P(r_j)} P(c_i|r_l) \\ &= \frac{\sum_{l=1}^{R} p(\boldsymbol{x}|r_i)P(r_i)P(c_i|r_l)}{p(\boldsymbol{x})} \end{aligned} \tag{14}$$

The Bayes decision rule can be thus formulated similarly to (3) as

$$\begin{aligned} &\boldsymbol{x} \text{ is assigned to } c_i \\ &\iff \sum_{l=1}^{R} p(\boldsymbol{x}|r_l)P(r_l)P(c_i|r_l) \\ &\geqslant \sum_{l=1}^{R} p(\boldsymbol{x}|r_l)P(r_l)P(c_j|r_l) \quad \forall j \neq i \end{aligned} \tag{15}$$

where the $p(\boldsymbol{x}|r_l)$ distribution is represented by Gaussians similarly to (4).

2.4. Extended fuzzy classifier

A new fuzzy model that is able to represent Bayes classifier defined by (15) can be obtained. The idea is to define the consequent of the fuzzy rule as the probabilities of the given rule represents the $c_1, \ldots, c_C$ classes:

$$\begin{aligned} r_i: \quad &\text{If } x_1 \text{ is } A_{i,1}(x_{1,k}) \text{ and } \ldots x_n \text{ is } A_{i,n}(x_{n,k}) \\ &\text{then } \hat{y}_k = c_1 \text{ with } P(c_1|r_i), \ldots, \hat{y}_k = c_C \\ &\quad \text{with } P(c_C|r_i)[w_i] \end{aligned} \tag{16}$$

Similarly to Takagi–Sugeno fuzzy models (Takagi and Sugeno, 1985), the rules of the fuzzy model are aggregated using the normalized fuzzy mean formula and the output of the classifier is determined by the label of the class that has the highest activation:

$$\hat{y}_k = c_{i^*}, \quad i^* = \arg\max_{1 \leqslant i \leqslant C} \frac{\sum_{l=1}^{R} \beta_l(\boldsymbol{x}_k)P(c_i|r_l)}{\sum_{i=1}^{R} \beta_l(\boldsymbol{x}_k)} \tag{17}$$

where $\beta_l(\boldsymbol{x}_k)$ has the meaning expressed by (7).

As the previous equation can be rewritten using only its numerator, the obtained expression is identical to the Gaussian mixtures of Bayes classifiers (15) when similarly to (11) the parameters of the fuzzy model are calculated as

$$\boldsymbol{A}_i(\boldsymbol{x}) = p(\boldsymbol{x}|r_i)|2\pi\boldsymbol{F}_i|^{n/2}, \quad w_i = \frac{P(r_i)}{|2\pi\boldsymbol{F}_i|^{n/2}} \tag{18}$$

The main advantage of the previously presented classifier is that the fuzzy model can consist of more rules than classes and every rule can describe more than one class. Hence, as a given class will be described by a set of rules, it should not be a compact geometrical object (hyper-ellipsoid).

The aim of the remaining part of the paper is to propose a new clustering-based technique for the identification of the fuzzy classifier presented above. In addition, a new method for the selection of the antecedent variables (features) of the model will be described.

3. Supervised fuzzy clustering

The objective of clustering is to partition the identification data $\boldsymbol{Z}$ into R clusters. This means, each observation consists of input and output variables, grouped into a row vector $\boldsymbol{z}_k = [\boldsymbol{x}_k^{\mathrm{T}}, y_k]$, where the k subscript denotes the $k = 1, \ldots, N$th row of the $\boldsymbol{Z}$ pattern matrix. The fuzzy partition is represented by the $\boldsymbol{U} = [\mu_{i,k}]_{R \times N}$ matrix, where the $\mu_{i,k}$ element of the matrix represents the degree of

membership, how the $\boldsymbol{z}_k$ observation is in the cluster $i = 1, \ldots, R$.

The clustering is based on the minimization of the sum of weighted $D^2_{i,k}$ squared distances between the data points and the η_i cluster prototypes that contains the parameters of the clusters.

$$J(\boldsymbol{Z}, \boldsymbol{U}, \eta) = \sum_{i=1}^{R} \sum_{k=1}^{N} (\mu_{i,k})^m D^2_{i,k}(\boldsymbol{z}_k, r_i) \tag{19}$$

where m is a fuzzy weighting exponent that determines the fuzziness of the resulting clusters. As m approaches one from above, the partition becomes hard ($\mu_{i,k} \in \{0, 1\}$), and $\boldsymbol{v}_i$ are the ordinary means of the clusters. As $m \to \infty$, the partition becomes fuzzy ($\mu_{i,k} = 1/R$) and the cluster means are all equal to the grand mean of $\boldsymbol{Z}$. Usually, m is often chosen as $m = 2$.

Classical fuzzy clustering algorithms are used to estimate the distribution of the data. Hence, they do not utilize the class label of each data point available for the identification. Furthermore, the obtained clusters cannot be directly used to build the classifier. In the following a new cluster prototype and the related distance measure will be introduced that allows the direct supervised identification of fuzzy classifiers. As the clusters are used to obtain the parameters of the fuzzy classifier, the distance measure is defined similarly to the distance measure of the Bayes classifier (5):

$$\frac{1}{D^2_{i,k}(\boldsymbol{z}_k, r_i)} = \underbrace{P(r_i) \prod_{j=1}^{n} \exp\left(-\frac{1}{2} \frac{(x_{j,k} - v_{i,j})^2}{\sigma^2_{i,j}} \right)}_{\text{Gath–Geva clustering}} \times P(c_j = y_k | r_i) \tag{20}$$

This distance measure consists of two terms. The first term is based on the geometrical distance between the $\boldsymbol{v}_i$ cluster centers and the $\boldsymbol{x}_k$ observation vector, while the second is based on the probability that the r_ith cluster describes the density of the class of the kth data, $P(c_j = y_k | r_i)$. It is interesting to note that this distance measure only slightly differs from the unsupervised GG clustering algorithm which can also be interpreted in a probabilistic framework (Gath and Geva, 1989). However, the novelty of the proposed approach is the second term, which allows the use of class labels.

To get a fuzzy partitioning space, the membership values have to satisfy the following conditions:

$$U \in \boldsymbol{R}^{c \times N} | \mu_{i,k} \in [0, 1] \ \forall i, k;$$
$$\sum_{i=1}^{R} \mu_{i,k} = 1 \ \forall k; \quad 0 < \sum_{k=1}^{N} \mu_{i,k} < N \quad \forall i \tag{21}$$

The minimization of the (22) functional represents a non-linear optimization problem that is subject to constraints defined by (21) and can be solved by using a variety of available methods. The most popular method, is the alternating optimization (AO), which consists of the application of Picard iteration through the first-order conditions for the stationary points of (22), which can be found by adjoining the constraints (21) to J by means of LaGrange multipliers (Hoppner et al., 1999),

$$\overline{J}(\boldsymbol{Z}, \boldsymbol{U}, \eta, \lambda) = \sum_{i=1}^{R} \sum_{k=1}^{N} (\mu_{i,k})^m D^2_{i,k}(\boldsymbol{z}_k, r_i) + \sum_{k=1}^{N} \lambda_k \left(\sum_{i=1}^{R} \mu_{i,k} - 1 \right) \tag{22}$$

and by setting the gradients of $\overline{J}$ with respect to $\boldsymbol{Z}$, $\boldsymbol{U}$, η and λ to zero.

Hence, similarly to the update equations of GG clustering algorithm, the following equations will result in a solution that satisfies the (22) constraints.

Initialization Given a set of data $\boldsymbol{Z}$ specify R, choose a termination tolerance $\epsilon > 0$. Initialize the $\boldsymbol{U} = [\mu_{i,k}]_{R \times N}$ partition matrix randomly, where $\mu_{i,k}$ denotes the membership that the $\boldsymbol{z}_k$ data is generated by the ith cluster.

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

Step 1 Calculate the parameters of the clusters

- Calculate the centers and standard deviation of the Gaussian membership functions (the diagonal elements of the $\boldsymbol{F}_i$ covariance matrices):

$$\boldsymbol{v}_i^{(l)} = \frac{\sum_{k=1}^{N} (\mu_{i,k}^{(l-1)})^m \boldsymbol{x}_k}{\sum_{k=1}^{N} (\mu_{i,k}^{(l-1)})^m},$$
$$\sigma_{i,j}^{2,(l)} = \frac{\sum_{k=1}^{N} (\mu_{i,k}^{(l-1)})^m (x_{j,k} - v_{j,k})^2}{\sum_{k=1}^{N} (\mu_{i,k}^{(l-1)})^m} \tag{23}$$

- Estimate the consequent probability parameters,

$$p(c_i|r_j) = \frac{\sum_{k|y_k=c_i} (\mu_{j,k}^{(l-1)})^m}{\sum_{k=1}^{N} (\mu_{j,k}^{(l-1)})^m}, \quad 1 \leqslant i \leqslant C,\ 1 \leqslant j \leqslant R \tag{24}$$

- A priori probability of the cluster and the weight (impact) of the rules:

$$P(r_i) = \frac{1}{N} \sum_{k=1}^{N} (\mu_{i,k}^{(l-1)})^m,$$
$$w_i = P(r_i) \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi\sigma_{i,j}^2}} \tag{25}$$

Step 2 Compute the distance measure $D_{i,k}^2(\boldsymbol{z}_k, r_i)$ by (20).

Step 3 Update the partition matrix

$$\mu_{i,k}^{(l)} = \frac{1}{\sum_{j=1}^{R} \left(D_{i,k}(\boldsymbol{z}_k, r_i)/D_{j,k}(\boldsymbol{z}_k, r_j)\right)^{2/(m-1)}}, \quad 1 \leqslant i \leqslant R,\ 1 \leqslant k \leqslant N \tag{26}$$

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

The remainder of this section is concerned with the theoretical convergence properties of the proposed algorithm. Since, this algorithm is the member of the family of algorithms discussed in (Hathaway and Bezdek, 1993), the following discussion is based on the results of Hathaway and Bezdek (1993). Using LaGrange multiplier theory, it is easily shown that for $D_{i,k}^2(\boldsymbol{z}_k, r_i) \geqslant 0$, (26) defines $\boldsymbol{U}^{(l+1)}$ to be a global minimizer of the restricted cost function (22). From this it follows that the proposed iterative algorithm is a special case of grouped coordinate minimization, and the general convergence theory from (Bezdek et al., 1987) can be applied for reasonable choices of $D_{i,k}^2(\boldsymbol{z}_k, r_i)$ to shown that any limit point of an iteration sequence will be a minimizer, or at worst a saddle point of the cost function J. The local convergence result in (Bezdek et al., 1987) states that if the distance measures $D_{i,k}^2(\boldsymbol{z}_k, r_i)$ are sufficiently smooth and a standard convexity holds at a minimizer of J, then any iteration sequence started with $\boldsymbol{U}^{(0)}$ sufficiently close to $\boldsymbol{U}^*$ will converge to the minima. Furthermore, the rate of convergence of the sequence will be γ-linear. This means that there is a norm $\|*\|$ and constants $0 < \gamma < 1$ and $l_0 > 0$, such that for all $l \geqslant l_0$, the sequence of errors $\{e^l\} = \{\|(\boldsymbol{U}^l) - (\boldsymbol{U}^*)\|\}$ satisfies the inequality $e^{l+1} < \gamma e^l$.

4. Feature selection based on interclass separability

Using too many input variables may result in difficulties in the interpretability capabilities of the obtained classifier. Hence, selection of the relevant features is usually necessary. Others have focused on reducing the antecedent variables by similarity analysis of the fuzzy sets (Roubos and Setnes, 2000), however this method is not very suitable for feature selection. In this section Fischer interclass separability method (Cios et al., 1998) is modified which is based on statistical properties of the data. The interclass separability criterion is based on the $\boldsymbol{F}_{\mathrm{B}}$ between-class and the $\boldsymbol{F}_{\mathrm{W}}$ within-class covariance matrices that sum up to the total covariance of the data $\boldsymbol{F}_{\mathrm{T}} = \boldsymbol{F}_{\mathrm{W}} + \boldsymbol{F}_{\mathrm{B}}$, where

$$\boldsymbol{F}_{\mathrm{W}} = \sum_{l=1}^{R} P(r_l)\boldsymbol{F}_l,$$
$$\boldsymbol{F}_{\mathrm{B}} = \sum_{l=1}^{R} P(r_l)(\boldsymbol{v}_l - \boldsymbol{v}_0)^{\mathrm{T}}(\boldsymbol{v}_l - \boldsymbol{v}_0), \tag{27}$$
$$\boldsymbol{v}_0 = \sum_{l=1}^{R} P(r_l)\boldsymbol{v}_l$$

The feature interclass seperability selection criterion is a trade-off between $\boldsymbol{F}_{\mathrm{W}}$ and $\boldsymbol{F}_{\mathrm{B}}$:

$$J = \frac{\det \boldsymbol{F}_{\mathrm{B}}}{\det \boldsymbol{F}_{\mathrm{W}}} \tag{28}$$

The importance of a feature is measured by leaving out the interested feature and calculating J for the reduced covariance matrices. The feature selection is a step-wise procedure, when in every step the least needed feature is deleted from the model.

In the current implementation of the algorithm after fuzzy clustering and initial model formulation a given number of features are selected by continuously checking the decrease of the performance of the classifier. To increase the classification

performance, the final classifier is identified based on the re-clustering of reduced data which have smaller dimensionality because of the neglected input variables.

5. Performance evaluation

In order to examine the performance of the proposed identification method two well-known multidimensional classification benchmark problems are presented in this section. The studied Wisconsin breast cancer and Wine data come from the UCI Repository of Machine Learning Databases (http://www.ics.uci.edu).

The performance of the obtained classifiers was measured by 10-fold cross validation. The data divided into ten sub-sets of cases that have similar size and class distributions. Each sub-set is left out once, while the other nine are applied for the construction of the classifier which is subsequently validated for the unseen cases in the left-out subset.

5.1. Example 1: the Wisconsin breast cancer classification problem

The Wisconsin breast cancer data is widely used to test the effectiveness of classification and rule extraction algorithms. The aim of the classification is to distinguish between *benign* and *malignant* cancers based on the available nine measurements: x_1 clump thickness, x_2 uniformity of cell size, x_3 uniformity of cell shape, x_4 marginal adhesion, x_5 single epithelial cell size, x_6 bare nuclei, x_7 bland chromatin, x_8 normal nuclei, and x_9 mitosis (data shown in Fig. 1). The attributes have integer value in the range (Baraldi and Blonda, 1999; Hoppner et al., 1999). The original database contains 699 instances however 16 of these are omitted because these are incomplete, which is common with other studies. The class distribution is 65.5% benign and 34.5% malignant, respectively.

The advanced version of C4.5 gives misclassification of 5.26% on 10-fold cross validation (94.74% correct classification) with tree size 25 ± 0.5 (Quinlan, 1996). Nauck and Kruse (1999) combined neuro-fuzzy techniques with interactive strategies for rule pruning to obtain a fuzzy classifier. An initial rule-base was made by applying two sets for each input, resulting in $2^9 = 512$ rules which was reduced to 135 by deleting the non-firing rules. A heuristic data-driven learning method was applied instead of gradient descent learning, which is not applicable for triangular membership functions. Semantic properties were taken into account by constraining the search space. They final fuzzy classifier could be reduced to two rules with 5–6 features only, with a misclassification of 4.94% on 10-fold validation (95.06% classification accuracy). Rule-generating methods that combine GA and fuzzy logic were also applied to this problem (Peña-Reyes and Sipper, 2000). In this method the number of rules to be generated needs to be determined a priori. This method constructs a fuzzy model that has four membership functions and one rule with an additional *else* part. Setiono (2000) has generated similar compact classifier by a two-step rule extraction from a feedforward neural network trained on preprocessed data.

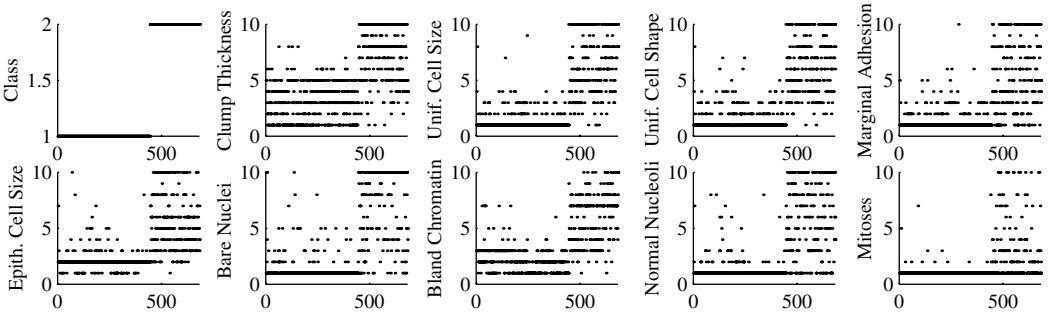

Fig. 1. Wisconsin breast cancer data: two classes and nine attributes (class 1: 1–445, class 2: 446–683).

As Table 1 shows, our fuzzy rule-based classifier is one of the most compact models in the literature with such high accuracy.

In the current implementation of the algorithm after fuzzy clustering an initial fuzzy model is generated that utilizes all the nine information profile data about the patient. A step-wise feature reduction algorithm has been used where in every step one feature has been removed continuously checking the decrease of the performance of the classifier on the training data. To increase the classification performance, the classifier is re-identified in every step by re-clustering of reduced data which have smaller dimensionality because of the neglected input variables. As Table 2 shows, our supervised clustering approach gives better results than utilizing the GG clustering algorithm in the same identification scheme.

The 10-fold validation experiment with the proposed approach showed 95.57% average classification accuracy, with 90.00% as the worst and 95.57% as the best performance. This is really good for such a small classifier as compared with previously reported results (Table 1). As the error estimates are either obtained from 10-fold cross validation or from testing the solution once by using the 50% of the data as training set, the results given in Table 1 are only roughly comparable.

5.2. Example 2: the wine classification problem

The wine data contains the chemical analysis of 178 wines grown in the same region in Italy but derived from three different cultivars. The problem is to distinguish the three different types based on 13 continuous attributes derived from chemical analysis (Fig. 2). Corcoran and Sen (1994) applied all the 178 samples for learning 60 non-fuzzy if–then rules in a real-coded genetic-based-machine learning approach. They used a population of 1500 individuals and applied 300 generations, with full replacement, to come up with the following result for 10 independent trials: best classification rate 100%, average classification rate 99.5% and worst classification rate 98.3% which is three misclassifications. Ishibuchi et al. (1999) applied all the 178 samples designing a fuzzy classifier with 60 fuzzy rules by means of an integer-coded genetic algorithm and grid partitioning. Their population contained 100 individuals and they applied 1000 generations, with full replacement, to come up with the following result for 10 independent trials: best classification rate 99.4% (one misclassification),

Table 1
Classification rates and model complexity for classifiers constructed for the Wisconsin breast cancer problem

Author	Method	♯ Rules	♯ Conditions	Accuracy (%)
Setiono (2000)	NeuroRule 1f	4	4	97.36
Setiono (2000)	NeuroRule 2a	3	11	98.1
Peña-Reyes and Sipper (2000)	Fuzzy-GA1	1	4	97.07
Peña-Reyes and Sipper (2000)	Fuzzy-GA2	3	16	97.36
Nauck and Kruse (1999)	NEFCLASS	2	10–12	95.06♮

♮ Denotes results from averaging a 10-fold validation.

Table 2
Classification rates and model complexity for classifiers constructed for the Wisconsin breast cancer problem

Method	Min Acc.	Mean Acc.	Max Acc.	Min ♯ Feat.	Mean ♯ Feat.	Max ♯ Feat.
GG: $R = 2$	84.28	90.99	95.71	8	8.9	9
Sup: $R = 2$	84.28	92.56	98.57	7	8	9
GG: $R = 4$	88.57	95.14	98.57	9	9	9
Sup: $R = 4$	90.00	95.57	98.57	8	8.7	9

Results from a 10-fold validation. GG: Gath–Geva clustering based classifier, Sup: proposed method.

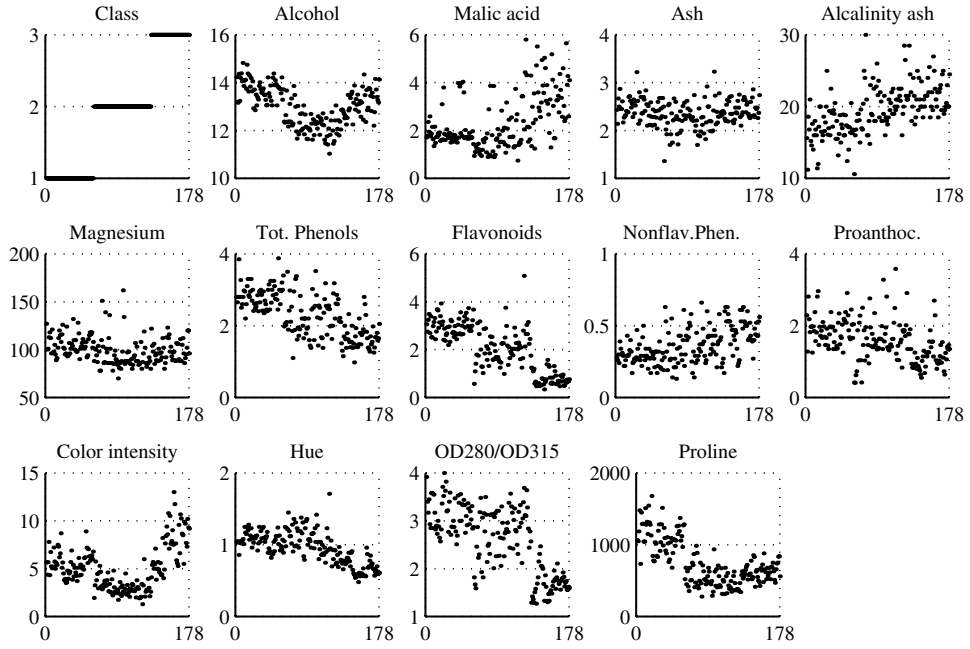

Fig. 2. Wine data: three classes and 13 attributes.

average classification rate 98.5% and worst classification rate 97.8% (four misclassifications). In both approaches the final rule base contains 60 rules. The main difference is the number of model evaluations that was necessary to come to the final result.

Firstly, for comparison purposes, a fuzzy classifier, that utilizes all the 13 information profile data about the wine has been identified by the proposed clustering algorithm based on all the 178 samples. Fuzzy models with three and six rules were identified. The three rule-model gave only two misclassification (correct percentage 98.9%). When a cluster was added to improve the performance of this model, the obtained classifier gave only one misclassification (99.4%). The classification power of the identified models is then compared with fuzzy models with the same number of rules obtained by GG clustering, as GG clustering can be considered the unsupervised version of the proposed clustering algorithm. The GG identified fuzzy model achieves eight misclassifications corresponding to a correct percentage of 95.5%, when three rules are used in the fuzzy model, while six misclassifications (correct percentage 96.6%) in the case of four rules. The results are summarized in Table 3. As it is shown, the performance of the obtained classifiers are comparable to those in

Table 3
Classification rates on the wine data for 10 independent runs

Method	Best result (%)	Average result (%)	Worst result (%)	Rules	Model evaluations
Corcoran and Sen (1994)	100	99.5	98.3	60	150 000
Ishibuchi et al. (1999)	99.4	98.5	97.8	60	6000
GG clustering	95.5	95.5	95.5	3	1
Sup (13 features)	98.9	98.9	98.9	3	1
Sup (13 features)	99.4	99.4	99.4	4	1

Table 4
Classification rates and model complexity for classifiers constructed for the Wine classification problem

Method	Min Acc.	Mean Acc.	Max Acc.	Min ♯ Feat.	Mean ♯ Feat.	Max ♯ Feat.
GG: $R = 3$	83.33	94.38	100	10	12.4	13
Sup: $R = 3$	88.88	97.77	100	12	12.6	13
GG: $R = 3$	88.23	95.49	100	4	4.8	5
Sup: $R = 3$	76.47	94.87	100	4	4.8	5
GG: $R = 6$	82.35	94.34	100	4	4.9	5
Sup: $R = 6$	88.23	97.15	100	4	4.8	5

Results from averaging a 10-fold validation.

(Corcoran and Sen, 1994; Ishibuchi et al., 1999), but use far less rules (3–5 compared to 60) and less features.

These results indicate that the proposed clustering method effectively utilizes the class labels. As can be seen from Table 3, because of the simplicity of the proposed clustering algorithm, the proposed approach is attractive in comparison with other iterative and optimization schemes that involves extensive intermediate optimization to generate fuzzy classifiers.

The 10-fold validation is a rigorous test of the classifier identification algorithms. These experiments showed 97.77% average classification accuracy, with 88.88% as the worst and 100% as the best performance (Table 4). The above presented automatic model reduction technique removed only one feature without the decrease of the classification performance on the training data. Hence, to avoid possible local minima, the feature selection algorithm is used to select only five features, and the proposed scheme has been applied again to identify a model based on the selected five attributes. This compact model with average 4.8 rules showed 97.15% average classification accuracy, with 88.23% as the worst and 100% as the best performance. The resulted membership functions and the selected features are shown in Fig. 3.

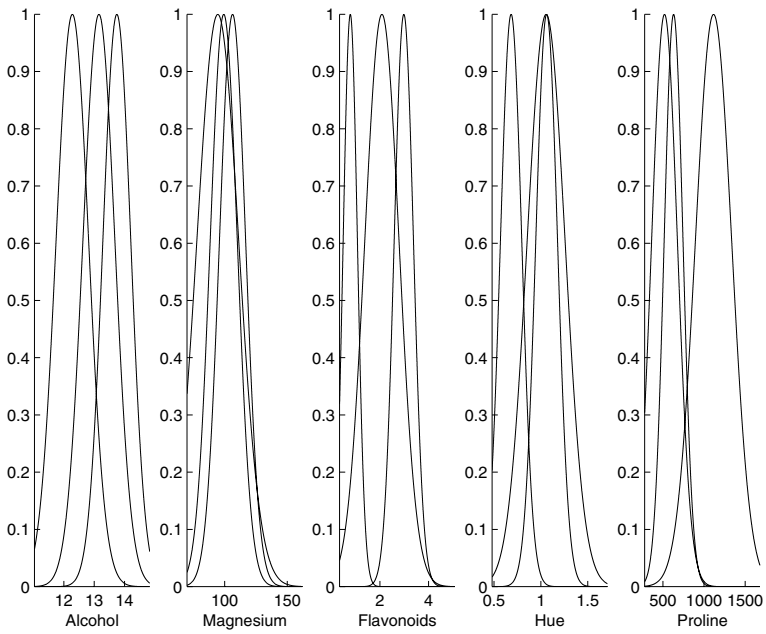

Fig. 3. Membership functions obtained by fuzzy clustering.

Comparing the fuzzy sets in Fig. 3 with the data in Fig. 2 shows that the obtained rules are highly interpretable. For example, the Flavonoids are divided in low, medium and high, which is clearly visible in the data.

6. Conclusions

In this paper a new fuzzy classifier has been presented to represent Bayes classifiers defined by mixture of Gaussians density model. The novelty of this new model is that each rule can represent more than one classes with different probabilities. For the identification of the fuzzy classifier a supervised clustering method has been worked out that is the modification of the unsupervised GG clustering algorithm. In addition, a method for the selection of the relevant input variables has been presented. The proposed identification approach is demonstrated by the Wisconsin breast cancer and the wine benchmark classification problems. The comparison to GG clustering and GA-tuned fuzzy classifiers indicates that the proposed supervised clustering method effectively utilizes the class labels and able to identify compact and accurate fuzzy systems.

Acknowledgements

This work was supported by the Hungarian Ministry of Education (FKFP-0073/2001) and the Hungarian Science Foundation (OTKA TO37600). Part of the work has been elaborated when J. Abonyi was at the Control Laboratory of Delft University of Technology. J. Abonyi is grateful for the Janos Bolyai Fellowship of the Hungarian Academy of Sciences.

References

Baraldi, A., Blonda, P., 1999. A survey of fuzzy clustering algorithms for pattern recognition—Part I. IEEE Trans. Systems Man Cybernet. Part B 29 (6), 778–785.

Bezdek, J.C., Hathaway, R.J., Howard, R.E., Wilson, C.A., Windham, M.P., 1987. Local convergence analysis of a grouped variable version of coordinate descent. J. Optimization Theory Appl. 71, 471–477.

Biem, A., Katagiri, S., McDermott, E., Juang, B.H., 2001. An application of discriminative feature extraction lo filter-bank-based speech recognition. IEEE Trans. Speech Audio Process. 9 (2), 96–110.

Campos, T.E., Bloch, I., Cesar Jr., R.M., 2001. Feature selection based on fuzzy distances between clusters: First results on simulated data. In: ICAPR'2001—International Conference on Advances in Pattern Recognition, Rio de Janeiro, Brazil, May. In: Lecture Notes in Computer Science. Springer-Verlag, Berlin.

Cios, K.J., Pedrycz, W., Swiniarski, R.W., 1998. Data Mining Methods for Knowledge Discovery. Kluwer Academic Press, Boston.

Corcoran, A.L., Sen, S., 1994. Using real-valued genetic algorithms to evolve rule sets for classification. In: IEEE-CEC, June 27–29, Orlando, USA. pp. 120–124.

Gath, I., Geva, A.B., 1989. Unsupervised optimal fuzzy clustering. IEEE Trans. Pattern Anal. Machine Intell. 7, 773–781.

Gustafson, D.E., Kessel, W.C., 1979. Fuzzy clustering with a fuzzy covariance matrix. In: Proceedings of IEEE CDC, San Diego, USA.

Hathaway, R.J., Bezdek, J.C., 1993. Switching regression models and fuzzy clustering. IEEE Trans. Fuzzy Systems 1, 195–204.

Hoppner, F., Klawonn, F., Kruse, R., Runkler, T., 1999. Fuzzy Cluster Analysis—Methods for Classification, Data Analysis and Image Recognition. John Wiley and Sons, New York.

Ishibuchi, H., Nakashima, T., Murata, T., 1999. Performance evaluation of fuzzy classifier systems for multidimensional pattern classification problems. IEEE Trans. SMC B 29, 601–618.

Kambhatala, N., 1996. Local models and Gaussian mixture models for statistical data processing. Ph.D. Thesis, Oregon Gradual Institute of Science and Technology.

Kim, E., Park, M., Kim, S., Park, M., 1998. A transformed input–domain approach to fuzzy modeling. IEEE Trans. Fuzzy Systems 6, 596–604.

Loog, L.C.M., Duin, R.P.W., Haeb-Umbach, R., 2001. Multiclass linear dimension reduction by weighted pairwise Fisher criteria. IEEE Trans. PAMI 23 (7), 762–766.

Nauck, D., Kruse, R., 1999. Obtaining interpretable fuzzy classification rules from medical data. Artificial Intell. Med. 16, 149–169.

Peña-Reyes, C.A., Sipper, M., 2000. A fuzzy genetic approach to breast cancer diagnosis. Artificial Intell. Med. 17, 131–155.

Quinlan, J.R., 1996. Improved use of continuous attributes in C4.5. J. Artificial Intell. Res. 4, 77–90.

Rahman, A.F.R., Fairhurst, M.C., 1997. Multi-prototype classification: improved modelling of the variability of handwritten data using statistical clustering algorithms. Electron. Lett. 33 (14), 1208–1209.

Roubos, J.A., Setnes, M., 2000. Compact fuzzy models through complexity reduction and evolutionary optimization. In: FUZZ-IEEE, May 7–10, San Antonio, USA. pp. 762–767.

Roubos, J.A., Setnes, M., Abonyi, J., 2001. Learning fuzzy classification rules from data. In: John, R., Birkenhead, R. (Eds.), Developments in Soft Computing. Springer-Verlag, Berlin/Heidelberg, pp. 108–115.

Setiono, R., 2000. Generating concise and accurate classification rules for breast cancer diagnosis. Artificial Intell. Med. 18, 205–219.

Setnes, M., Babuška, R., 1999. Fuzzy relational classifier trained by fuzzy clustering. IEEE Trans. SMC B 29, 619–625.

Setnes, M., Babuška, R., Kaymak, U., van Nauta Lemke, H.R., 1998. Similarity measures in fuzzy rule base simplification. IEEE Trans. SMC B 28, 376–386.

Takagi, T., Sugeno, M., 1985. Fuzzy identification of systems and its application to modeling and control. IEEE Trans. SMC 15, 116–132.

Valente de Oliveira, J., 1999. Semantic constraints for membership function optimization. IEEE Trans. FS 19, 128–138.