classification, fuzzy clustering,
unsupervised classification,
nearest neighbors classifier

Tomasz PRZYBYŁA[1], Tomasz PANDER[1],
Krzysztof HOROBA[2], Tomasz KUPKA[2], Adam MATONIA[2]

AN APPROACH TO UNSUPERVISED CLASSIFICATION

Classification methods can be divided into supervised and unsupervised methods. The supervised classifier requires a training set for the classifier parameter estimation. In the case of absence of a training set, the popular classifiers (e.g. K-Nearest Neighbors) can not be used. The clustering methods are considered as unsupervised classification methods. This paper presents an idea of the unsupervised classification with the popular classifiers. The fuzzy clustering method is used to create a learning set. The learning set includes only these patterns that are the best representative of each class in the input dataset. The numerical experiment uses an artificial dataset as well as the medical datasets (PIMA, Wisconsin Breast Cancer) and illustrates the usefulness of the proposed method.

1. INTRODUCTION

Pattern classification methods play a very important role in pattern recognition. The classification methods are successfully applied in the biomedical engineering area, e.g. in the Computer-Brain interface [2], detecting an abnormal brain activity [3], controlling the prostheses [17], as well as in the gender recognition [1]. Generally, pattern recognition methods can be divided into two main categories. The first category contains supervised classification methods, while the second category includes the unsupervised classification methods. One of the most popular supervised classification method is the *K*-Nearest Neighbors (*K*-NN) method [4,10,18]. The designing of a supervised classificatory approach requires a learning (training) set. The learning set is required for an estimation of classifier parameters. However, the designing of a classifier without training set is a very difficult task [14,15].

On the other hand, the unsupervised methods do not require a training set. Most of the unsupervised classification methods are clustering methods [8,9,13]. The clustering aims at assigning a set of objects to clusters in such a way that objects within the same cluster have a high degree of similarity, while objects belonging to different clusters are dissimilar. The clustering methods can be divided into two main categories: hierarchical and partitional [2,5,6]. In the hierarchical clustering a number of clusters need not to be specified a priori. The problems concerning an initialization and an occurrence of local minima are also irrelevant. However, it cannot incorporate a priori knowledge about the global shape or size of clusters since hierarchical methods consider only local neighbors in each step [8].

Prototype-based partitional clustering methods can be classified into two classes: hard (or crisp) methods and fuzzy methods [11,12]. In the hard clustering methods every case belongs to only one cluster. In the fuzzy clustering methods every data point belongs to every cluster. Fuzzy clustering algorithms can deal with overlapping cluster boundaries. The most familiar fuzzy clustering method is the fuzzy *c*-means clustering method proposed by Bezdek [2].

The learning set can be created by an expert. In this case, the expert have to assign labels for each pattern from the unlabeled dataset. For a large dataset (thousands or tens of thousands patterns), such approach is very tedious for the expert. In such case, the expert can assign labels only for randomly selected patterns. The number of selected patterns is much lower than the cardinality of the dataset, but there is uncertain whether the samples have been selected correctly (i.e. that intrinsic structures are represented correctly). Another approach to building the learning set uses clustering methods. The clustering methods discover the internal structure in the dataset. The obtained groups (clusters) represents those patterns that are very similar within group and are very dissimilar to patterns from other groups.

[1] Silesian University of Technology, Institute of Electronics, Akademicka St. 16, 44-100 Gliwice, Poland.
[2] Institute of Medical Technology and Equipment, Biomedical Signal Processing Department, Roosevelt St. 118, 41-800 Zabrze, Poland.

One of the results obtained from the clustering procedure is the partition matrix. For the medical data, the obtained groups can contain similar cases (e.g. one group can represent healthy patients while other group can represent with a disease entity). By analyzing the values of the obtained partition matrix, it is possible to select only these patterns (patients) with high membership degree (patients who are the best representative of each group). In the proposed method, the patterns with membership degree greater than the assumed threshold are chosen. In this way, the learning set consists of patterns that are the best representative for the classes in the input dataset.

The goal of this work is to propose an unsupervised classification method. The proposed method consists of two stages. At the first stage, a fuzzy clustering procedure is applied to the input dataset. At this stage, a learning dataset is created from those patterns which membership degrees meet assumed criteria. At the second stage, the classification method is applied to the remaining patterns from the input dataset.

This paper is organized as follows. The section 2 contains overview of the classification and clustering methods used in proposed approach. The proposed procedure is presented in section 3. Section 4 contains numerical experiments. Conclusions complete the paper.

2. METHODS

Selected methods used in proposed approach to unsupervised classification are presented in this section. First, the fuzzy clustering method is introduced. In the next subsection, two classification methods are presented: the classification method based on the Fisher linear discriminant analysis and the *K*-nearest neighbors method. The minimum class-mean distance classifier was not used. The obtained results of this classifier are the same as the result from the clustering stage [19]. In a such case, there is no need to use the classification step.

2.1. FUZZY CLUSTERING

For the fuzzy clustering methods, the fuzzy partition matrix is defined in the following way:

$$M_{fc} = \left\{ \mathbf{U} \in \Re^{c \times N} \,|\, \underset{i,j}{\forall} u_{ik} \in [0,1]; \underset{1 \le k \le N}{\forall} \sum_{i=1}^{c} u_{ik} = 1; \underset{1 \le i \le c}{\forall} \sum_{k=1}^{N} u_{ik} < N \right\},$$

where: N is the number of objects, and c is the number of clusters.

The FCM method is the prototype-based method, where the objective function has been defined as follows:

$$J_m(\mathbf{U}, \mathbf{V}) = \sum_{k=1}^{N} \sum_{i=1}^{c} u_{ik}^m \left\| \mathbf{x}_k - \mathbf{v}_i \right\|^2, \tag{1}$$

where: $\boldsymbol{U}$ is the fuzzy partition matrix, $\boldsymbol{V}=\{\boldsymbol{v_1}, \boldsymbol{v_2}, \ldots \boldsymbol{v_c}\}$ is the set of prototype vectors and $\underset{1 \le i \le c}{\forall} v_i \in \Re^p$, $\boldsymbol{x}_k$ is the feature vector $\underset{1 \le k \le N}{\forall} x_k \in \Re^p$, p is the number of features describing the clustering objects, and m is the fuzzyfying exponent.

The optimization of the objective function (1) is completed with respect to partition matrix $\boldsymbol{U}$ and prototypes of the clusters $\boldsymbol{V}$. The optimal values of the partition matrix can be calculated as follows:

$$\underset{1 \le i \le c}{\forall} \underset{1 \le k \le N}{\forall} u_{ik} = \begin{cases} \left[\sum_{j=1}^{c} \left(\dfrac{\left\| \mathbf{x}_k - \mathbf{v}_i \right\|}{\left\| \mathbf{x}_k - \mathbf{v}_j \right\|} \right)^{2/(m-1)} \right]^{-1} & \text{if} \quad \Im_k = \varnothing \\ 0 & \text{if} \quad \underset{i \in \Im_k}{\forall}, \\ 1 & \text{if} \quad \Im_k \neq \varnothing \end{cases} \tag{2}$$

where: the sets $\Im_k$ and $\widetilde{\Im}_k$ are defined in the following way:

$$\underset{1\le k\le N}{\forall}\Im_k=\left\{i\,|1\le i\le c;\left\|\mathbf{x}_k-\mathbf{v}_i\right\|^2=0\right\}$$
$$\widetilde{\Im}=\{1,2,\ldots,c\}-\Im_k$$

The optimal values of the cluster prototypes can be computed using the formula:

$$\underset{1\le i\le c}{\forall}\mathbf{v}_i=\frac{\sum_{k=1}^{N}u_{ik}^{m}\mathbf{x}_k}{\sum_{k=1}^{N}u_{ik}^{m}}. \tag{3}$$

2.2. CLASSIFICATION METHODS

2.2.1. FISHER LINEAR DISCRIMINANT ANALYSIS

Let us consider a set of N training samples $\{\boldsymbol{x_1},\ldots\boldsymbol{x_N}\}$ taking values in a p-dimensional space. Let c denotes the number of classes and c_i be the number of training samples of class i ($1 \le i \le c$). Then the between-class scatter matrix S_b has the following expression [10,15]

$$\mathbf{S}_b=\sum_{i=1}^{c}(m_i-m_0)(m_i-m_0)^T,$$

where m_i is the mean vector of training samples in class i, and m_0 is the mean vector of all training samples. Similarly, the within-class scatter matrix can be defined as follows:

$$\mathbf{S}_w=\sum_{i=1}^{c}\sum_{k=1}^{c_i}\left(x_i^{(k)}-m_i\right)\left(x_i^{(k)}-m_i\right)^T,$$

where $x_i^{(k)}$ denotes kth sample from class i, and m_i denotes the mean vector of samples from class i.

The linear discriminant analysis methods seeks a set of $d<<p$ basis vectors $\varphi=[\varphi_1,\ \ldots\ ,\varphi_d]$ in such way that the ratio between-class and within-class scatter matrices of the training samples is maximized. Fisher criterion has the following form [7,15]

$$J(\varphi)=\frac{\varphi^T\mathbf{S}_b\varphi}{\varphi^T\mathbf{S}_w\varphi},$$

where $\varphi=[\varphi_1,\ \ldots\ ,\varphi_d]$, and $\varphi_i\in\Re^p$.
The optimal vectors φ are defined as follows

$$\varphi=\arg\max_{\varphi}J(\varphi)=\arg\max_{\varphi}\frac{\varphi^T\mathbf{S}_b\varphi}{\varphi^T\mathbf{S}_w\varphi}. \tag{4}$$

Hence, when $\boldsymbol{S_w}$ is non-singular, the basis vectors φ correspond to the first d most significant eigenvectors of $\left(\mathbf{S}_w^{-1}\mathbf{S}_b\right)$. The word "significant" means the eigenvalues corresponding to these eigenvectors are the first d largest ones.
The classification rule is defined as follows. The unknown pattern $\boldsymbol{x}$ is classified to ith class, when the following equation holds true (minimum class-mean distance classifier)

$$\left|\varphi^T \mathbf{x} - \varphi^T m_i\right| < \left|\varphi^T \mathbf{x} - \varphi^T m_k\right| \text{ for } k \neq i\,. \tag{5}$$

2.2.2. K-NEAREST NEIGHBORS METHOD

The *K*-nearest neighbors method is a method for classifying objects based on the closest training examples in the feature space. The *k*-NN method is a type of instance-based learning. The *K*-NN method is among the simplest of all machine learning methods. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its *K* nearest neighbors. If *K=1*, then the object is simply assigned to the class of its nearest neighbor.

3. UNSUPERVISED CLASSIFICATION

The proposed procedure can be described as follows:

1. For the given dataset ***X***, find *c* classes using fuzzy clustering method,
2. Based on the partition matrix **U**, select these patterns with membership degree greater than assumed threshold value U_T.
3. Classify the rest patterns from the dataset ***X*** using a selected classifier.

4. NUMERICAL EXPERIMENTS

In our numerical experiments the value of the fuzzyfing exponent *m=2* and the tolerance limit $\varepsilon=10^{-5}$ are chosen. The Euclidean distance is used as the distance metric. The accuracy is measured as the ratio of incorrect assigned samples and total number of samples in the dataset. The accuracy is expressed as a percentage of misclassified samples, i.e.

$$\varepsilon_0 = \frac{N_0}{N} \times 100\%\,,$$

where: N_0 is the number of misclassified samples, and N is the total number of samples in the dataset.

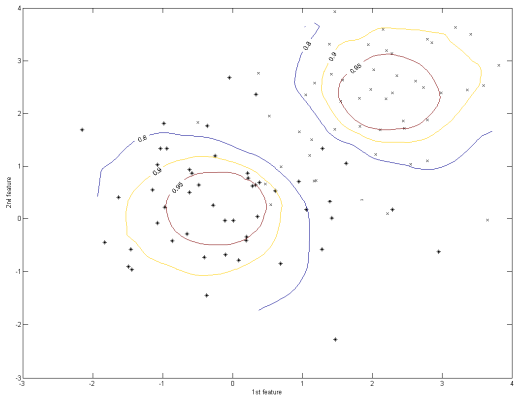

Fig. 1. Original data – the dataset contains two overlapped groups of 50 samples. The contour lines represent different membership degrees after clustering stage.

The purpose of the first experiment is to investigate the ability to correct classification of patterns from the dataset. For this purpose, an artificial dataset is generated by a pseudo-random generator. This dataset contains two overlapped groups, and is presented in Figure 1. Each group includes 50 samples in 2D space. Each sample has an assigned label. Thereby the correctness of the proposed approach could be

confirmed. For the dataset, two classification methods are used. The *K*-NN method as the first classification is used. As the second method, the Fisher's discrimination method is used. For both methods, the threshold value is selected from the set $U_T \in \{0.95, 0.9, 0.8\}$. For the *K*-NN method, number of neighbors was taken from the set $K \in \{5, 7, 11, 17\}$. For the Fisher's method, the number of dimensions is fixed as *d=1* and *d=2*. The obtained results are presented in Table 1.

Table 1. The percentage numbers of misclassified samples from the first dataset.

U_T=0.95							
Method: *K*NN							
K=5	ε_0=13%	*K*=7	ε_0=12%	*K*=11	ε_0=12%	*K*=17	ε_0=12%
Method: FLDA							
D=1		ε_0=11%		*D*=2		ε_0=11%	
U_T=0.9							
Method: *K*NN							
K=5	ε_0=14%	*K*=7	ε_0=16%	*K*=11	ε_0=14%	*K*=17	ε_0=12%
Method: FLDA							
D=1		ε_0=15%		*D*=2		ε_0=12%	
U_T=0.8							
Method: *K*NN							
K=5	ε_0=13%	*K*=7	ε_0=13%	*K*=11	ε_0=13%	*K*=17	ε_0=15%
Method: FLDA							
D=1		ε_0=12%		*D*=2		ε_0=13%	

The obtained classification error varies from 11% to 15%. The artificial dataset contains two overlapped classes. The misclassification is caused by the use of linear classifiers. When the classified classes overlapped, then the linear classifiers do not provide reliable results.

In the second numerical experiment, the Wisconsin Breast Cancer dataset has been used. This dataset contains 569 cases of breast cancer. Two type of tumor appear in the dataset: 357 cases of malignant tumor, and 212 cases of benign tumor. So, for the clustering process, the number of cluster is fixed at *c*=2. Each case of tumor is represented by 8 features vector. As in the first experiment, the number of neighbors varies from 5 to 17. For the Fisher's method, the maximum number of dimensions is 6 (the covariance matrix of the data has only six nonzero eigenvectors and corresponding eigenvalues). The obtained results are presented in Table 2.

Table 2. The performance of proposed method for the Wisconsin Breast Cancer dataset.

U_T=0.95							
Method: *K*NN							
K=5	ε_0=14.76%	*K*=7	ε_0=14.76%	*K*=11	ε_0=14.76%	*K*=17	ε_0=15.11%
Method: FLDA							
D=2	ε_0=16.52%		*D*=4	ε_0=15.29%		*D*=6	ε_0=15.29%
U_T=0.9							
Method: *K*NN							
K=5	ε_0=15.46%	*K*=7	ε_0=15.46%	*K*=11	ε_0=15.64%	*K*=17	ε_0=15.64%
Method: FLDA							
D=2	ε_0=16.52%		*D*=4	ε_0=16.52%		*D*=6	ε_0=16.7%
U_T=0.8							
Method: *K*NN							
K=5	ε_0=14.76%	*K*=7	ε_0=14.94%	*K*=11	ε_0=14.94%	*K*=17	ε_0=15.46%
Method: FLDA							
D=2	ε_0=16.34%		*D*=4	ε_0=17.22%		*D*=6	ε_0=14.41%

In the last numerical experiment, the Pima database is used. It comprises 768 cases of patients who may show signs of diabetes. In this dataset appear 500 cases of healthy patients and 268 cases of patient who show signs of diabetes. Each case in the Pima dataset is described by 8 features. Similarly as in the previous experiments, the number of neighbors varies from 5 to 17, and the number of dimensions in the Fisher's LDA varies from 2 to 6. The obtained results for different thresholds U_T and different number of dimensions are presented in Table 3.

Table 3. The performance of proposed method for the Pima dataset.

U_T=0.95							
Method: *K*NN							
K=5	ε_0=33.98%	*K*=7	ε_0=33.98%	*K*=11	ε_0=33.98%	*K*=17	ε_0=25.52%
Method: FLDA							
D=2	ε_0=34.76%	*D*=4	ε_0=34.5%	*D*=6	ε_0=34.76%		
U_T=0.9							
Method: *K*NN							
K=5	ε_0=34.11%	*K*=7	ε_0=34.24%	*K*=11	ε_0=34.24%	*K*=17	ε_0=34.24%
Method: FLDA							
D=2	ε_0=34.5%	*D*=4	ε_0=34.76%	*D*=6	ε_0=34.5%		
U_T=0.8							
Method: *K*NN							
K=5	ε_0=34.37%	*K*=7	ε_0=34.24%	*K*=11	ε_0=34.24%	*K*=17	ε_0=33.98%
Method: FLDA							
D=2	ε_0=34.11%	*D*=4	ε_0=34.11%	*D*=6	ε_0=33.98%		

5. CONCLUSIONS

In this paper, an idea of an unsupervised classification is presented. The proposed classification procedure includes two stages. In the first stage, the fuzzy c-means clustering method is used for finding groups in the input dataset. The patterns with high membership degree are chosen for the learning set. At this stage the learning set is created. Such kind of strategy ensures the correct creation of a learning set. In the last step, the classification is performed on the rest of the dataset.

The future works aims of solving the problem of the linear classification in kernel space. The proposed approach will be developed for better performance.

6. ACKNOWLEDGMENT

This work was in part financed by the Polish Ministry of Science and Higher Education, and by the Polish National Science Centre.

The authors would like to thank the anonymous referees for their valuable suggestions.

BIBLIOGRAPHY

[1] BEKIOS-CALFA J., BUENAPOSADA J. BAUMELA L., Revisiting linear discriminant techniques in gender recognition, IEEE Trans. Patt. An. Mach. Int. 33, 2011, pp. 858-864.
[2] BEZDEK J.C., Pattern Recognition With Fuzzy Objective Function Algorithms. Plenum, New York, 1981.
[3] CECOTTI H., GRASER A., Convolutional neural networks for P300 detection with application to Brain-Computer interface, IEEE Trans Patt. An. Mach. Int. 33, 2011, pp. 433-445.
[4] CHAOVALITWONGSE W., FAN Y.J., SACHDEO R., On the time series K-nearest neighbor classification of abnormal brain activity, IEEE Trans Sys. Man Cyber. A, 37, 2007, pp. 1005-1016.
[5] COVER T.M., HART P.E., Nearest neighbor pattern classification, IEEE Trans. Inf. Theory 13, 1967, pp. 21-27.

[6] DUDA R.O., HART P.E., STORK D.G., Pattern Classification. Wiley-Interscience, New Jersey, 2000.
[7] HASTIE T., TIBSHIRANI R., Discriminant adaptive nearest neighbor classification, IEEE Trans. Patt. An. Mach. Int. 18, 1996, pp. 607-616.
[8] JAIN A.K., Data clustering: 50 years beyond K-means, Patt. Rec. Let. 31, 2010, pp. 651-666.
[9] KAUFMAN L., ROUSSEEUW P., Finding Groups In Data. Wiley-Interscience, New Jersey, 1990.
[10] LIANG Z., LI Y., SHI P., A note on two-dimensional linear discriminant analysis, Patt. Rec. Let. 29, 2008, pp. 2122-2128.
[11] MITANI Y., HAMAMOTO Y., A local mean-based nonparametric classifier, Patt. Rec. Let. 27, 2006, pp. 1151-1159.
[12] PRZYBYLA T., JEZEWSKI J., HOROBA K., ROJ D., Hybrid Fuzzy Clustering Using LP Norms, Intelligent Information and Database Systems, Editors: Ngoc Thanh Nguyen, Chong Gun Kim, Adam Janiak, LNAI 6591/Lecture Notes in Computer Science, Springer Verlag, 2011, pp. 187-196.
[13] PRZYBYLA T., JEZEWSKI J., ROJ D. On Hybrid Fuzzy Clustering Method, Information Technologies in Biomedicine, Editors: PIETKA E., KAWA J., Advances in Soft Computing Series, Vol. 69, Springer Verlag, 2010, pp. 3-14.
[14] PRZYBYLA T., JEZEWSKI J., ROJ D. Unsupervised clustering for fetal state assessment based on selected features of the cardiotocographic signals, Journal of Medical Informatics and Technologies, Vol. 13, 2009, pp. 157-162.
[15] RODRIGUEZ-LUJAN I., SANTA CRUZ S., HUERTA R., On the equivalence of kernel Fisher discriminant analysis and kernel quadratic programming feature selection, Patt. Rec. Let. 32, 2011, pp. 1567-1571.
[16] SCHOELKOPF B., SMOLA A.J., Learning with Kernels, The MIT Press, 2002.
[17] SCHEME E.J., ENGLEHART K.B., HUDGINS B.S., Selective classification for improved robustness of meyelectric control under nonideal conditions, IEEE Trans. Patt. An. Mach. Int. 58, 2011, pp. 1698-1705.
[18] SHAWE-TAYLOR J., CRISTIOANINI N., Kernel Methods for Pattern Analysis, Cambridge University Press, 2004.
[19] ZHENG W., ZHAO L., ZOU C., Locally nearest neighbor classifier for pattern classification, Patt. Rec. 37, 2004, pp. 1307-1309.