JOURNAL OF MEDICAL INFORMATICS & TECHNOLOGIES Vol. 17/2011, ISSN 1642-6037

fuzzy clustering,
classification,
medical data

Michal JEZEWSKI[1]

APPLICATION OF MODIFIED FUZZY CLUSTERING TO MEDICAL DATA CLASSIFICATION

Classification plays very important role in medical diagnosis. This paper presents fuzzy clustering method dedicated to classification algorithms. It focuses on two additional sub-methods modifying obtained clustering prototypes and leading to final prototypes, which are used for creating the classifier fuzzy if-then rules. The main goal of that work was to examine a performance of the classifier which uses such rules. Commonly used including medical benchmark databases were applied. In order to validate the results, each database was represented by 100 pairs of learning and testing subsets. The obtained classification quality was better in relation to the one of the best classifiers – Lagrangian SVM and suggests that presented clustering with additional sub-methods are appropriate to application to classification algorithms.

1. INTRODUCTION

Classification, determined also as pattern recognition [3], consisting in assigning objects to predefined classes, plays very important role in many fields of science, including medical diagnosis support. A trained classifier enables detection of signs of distress basing on some patient's input data (examination results, medical images etc.). One of many examples may be the prediction of newborn condition done during pregnancy [2,5,6]. Classification represents supervised learning, which means that learning subset includes assignment of objects to classes, in case of medical databases – to diagnosis. In case of clustering [1], which represents unsupervised learning, only objects features are taken into consideration. In other words, clustering processes search structures in dataset. There are many different classifiers: neural networks, linear, statistical etc. Nowadays the nonlinear SVM (Support Vector Machines) classifier is regarded as one of the best classification algorithms. It represents kernel-based methods whose idea consists in transformation of objects from original feature space to the new high-dimensional feature space, where objects may be linearly separable. In [10] a combination of linear classifiers leading to nonlinear classifier in original feature space was proposed. Beside its advantages, like using fuzzy if-then rules enabling the interpretation of classification, the necessity of separate clustering (by the fuzzy c-means algorithm) of objects from both classes may be regarded as an inconvenience. The goal of the presented work is to examine a performance of the classifier which uses fuzzy if-then rules created basing on the special fuzzy clustering method dedicated to classification algorithms and additional sub-methods.

2. FUZZY CLUSTERING DEDICATED TO CLASSIFICATION ALGORITHMS

The fuzzy clustering method (FCB) dedicated to classification algorithms and enabling a clustering of dataset including objects from both classes was proposed [9]. It is based on minimization of the following criterion:

$$J(\mathbf{U},\mathbf{V}) = \sum_{i=1}^{c}\sum_{k=1}^{N}(u_{ik})^m d_{ik}^2 + \alpha\sum_{i=1}^{c}\left(\sum_{k=1}^{N}u_{ik}y_k\right)^2 \tag{1}$$

with the constraints

[1] Institute of Electronics, Silesian University of Technology, Akademicka 16, 44-100 Gliwice, Poland,
email: michal.jezewski@polsl.pl.

$$\underset{i=1,2,\ldots,c}{\forall} \quad \sum_{k=1}^{N} u_{ik} = L, \quad L > 0 \tag{2}$$

The above criterion takes into consideration assignment of N objects to one of two classes ($y_k = +1$ or $y_k = -1$) so it may be regarded as a kind of clustering with partial supervision. The partition (prototype) matrix is denoted by $\boldsymbol{U}$ ($\boldsymbol{V}$), d_{ik} represents Euclidean distance between ith prototype and kth object, α determines the proportion between both components of the criterion. According to the assumed constraints (2), the fuzzy cardinality of each of c groups should be equal to L. The second component of (1) plays the most important role because it is responsible for creating the clusters, which include objects from both classes with similar memberships. Prototypes of such clusters should be placed near boundary between two classes. Details of the criterion minimization procedure, obtained solution and implementation can be found in [9].

Detailed analysis of clustering results using different datasets showed, that sometimes not all of obtained prototypes were located appropriately. For this reason, given ith learning subset (LS_i) is clustered by the FCB to obtain *ci* intermediate prototypes. The c final prototypes (**V_i_c**) determining c classifier fuzzy if-then rules are obtained basing on intermediate prototypes with a help of one of two additional sub-methods (Fig. 1). Designing the classifier which uses from two to eight rules (c = 2, 3, …, 8) was assumed.

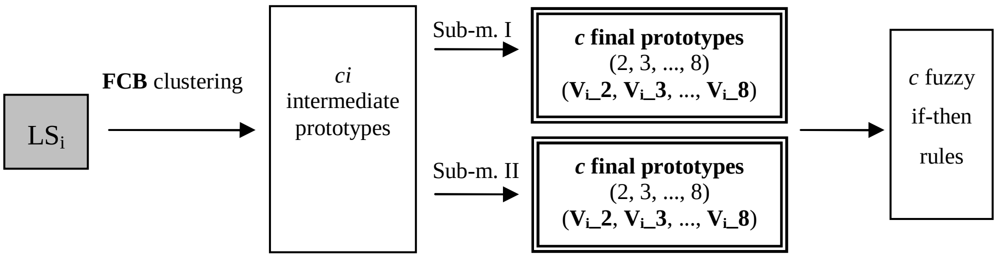

Fig. 1. Role of the FCB clustering in classifier designing process.

Sub-method I: Clustering of intermediate prototypes. In case of this method, each of 10 first learning subsets is clustered by the FCB into *ci* clusters (Fig. 2). For a given database, each clustering is started from the same initial prototype matrix (one matrix for one database, randomly chosen at the beginning). In the following step, obtained *ci* intermediate prototypes, separate for each learning subset, are clustered by the fuzzy c-means (FCM) algorithm [1] into 2, 3, …, 8 clusters. Prototypes of these clusters become final prototypes used to determine the classifier if-then rules. The fuzzy c-means clustering into given number of clusters (2, 3, …, 8) is always (for each database and each learning subset) performed starting from one of seven the same initial partition matrices (sizes $2 \times ci$, $3 \times ci$, …, $8 \times ci$), randomly chosen at the beginning.

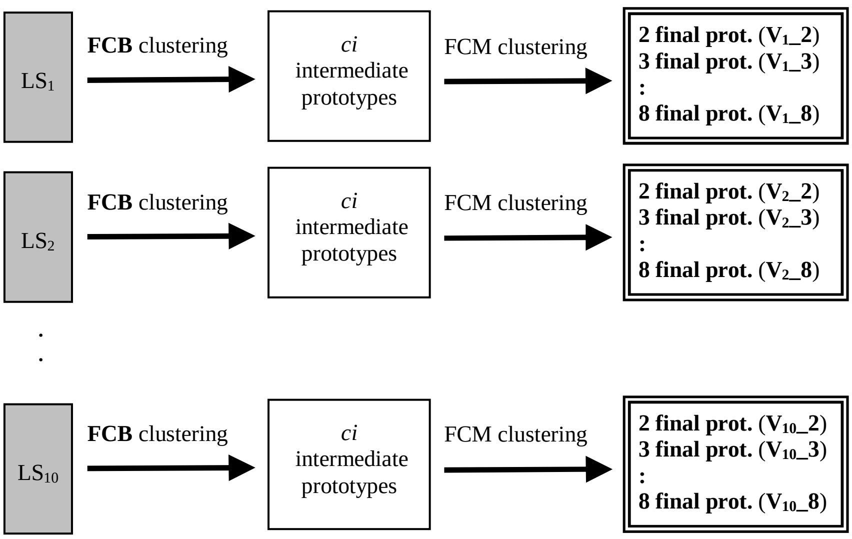

Fig. 2. Sub-method I: Clustering of intermediate prototypes.

Sub-method II: Choice from intermediate prototypes. This method is characterized by two-stage choice of final prototypes (Fig. 3). In the first stage, 10 first learning subsets are clustered by the FCB 700 times in total. Each FCB clustering is performed starting from different randomly chosen initial prototype matrix. During the first stage, initially selected prototypes (ISP) are chosen. The algorithm used for this choice is based on distances between prototypes and assignment (of objects the closest to prototypes) to classes. Prototypes which:

- have the highest distances between each other (are the most scattered),
- are surrounded by objects with the highest diversity of assignment to both classes.

are chosen as initially selected prototypes. As a result of the first stage, 100 sets of 2 ISP, 100 sets of 3 ISP, …, 100 sets of 8 ISP are obtained. In the second stage, each set of ISP is used to determine the classifier if-then rules (described later). Basing on classification of 10 first pairs of learning and testing (TS) subsets, finally one set (ensuring the lowest classification error for the first 10 testing subsets) for a given number of final prototypes is chosen. In opposite to the previously described sub-method I, the final prototypes obtained using the sub-method II are common for all learning subsets of a given database.

The duration time of sub-method II is much higher than in case of sub-method I, but classification results turned out to be better when using the final prototypes from sub-method II. Detailed description of both sub-methods can be found in [4], they were developed to be different in three aspects described in Table 1.

The above description of sub-methods takes into consideration 10 first learning and testing subsets because cross-validation procedure using 100 pairs of learning and testing subsets of each database was applied to obtain a good generalization ability and to validate results. In the first stage, the values of classifier parameters ensuring the lowest classification error for the first 10 testing subsets are chosen. Using these values, in the second stage, the final results (mean value and standard deviation of classification error) for all 100 testing subsets are obtained.

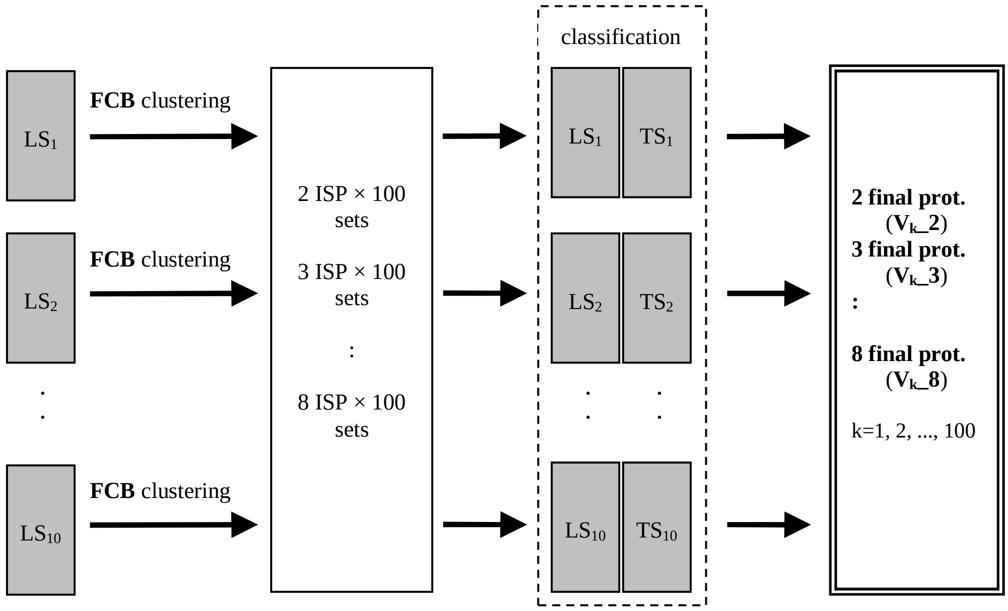

Fig. 3. Sub-method II: Choice from intermediate prototypes.

Experiments for ci = 50, 200, 1000 intermediate prototypes were performed. Final prototypes were created with a help of sub-method I or sub-method II, so as a result six final classification results for each database were obtained.

Table 1. Comparison of both sub-methods.

Aspect	Sub-method I	Sub-method II
Initial prototype matrix	Constant for a given database	Randomly changed
Final prototypes creating	New, created based on intermediate prototypes	Chosen from intermediate prototypes
Final prototypes assignment	Separate for each learning subset	Common for all learning subsets

Seven commonly used benchmark databases were applied to verify the obtained classification quality: *banana*, *breast-cancer*, *diabetis*, *heart*, *Ripley data*, *titanic*, *thyroid*. All of them consist of objects assigned to two classes. They represent synthetic (*banana*, *Ripley data*), non-medical (*titanic*) as well as medical data (all four remaining). The structure of data is also various – different number of features, different proportions between learning and testing subsets. Databases were obtained on December 2009 from http://ida.first.fraunhofer.de /projects/bench/benchmarks.htm where they were represented by 100 pairs of learning and testing subsets. This source does not include the *Ripley data*, so the original *Ripley* learning and testing subsets were joined and randomly divided into 100 pairs of subsets.

Fuzzy if-then rules in the Takagi-Sugeno-Kang form were chosen. Fuzzy sets in antecedents had Gaussian membership functions, whereas consequents had linear functions. Antecedents parameters were determined directly basing on the final prototypes, in particular the final prototypes were established as centers of Gaussian functions. Consequents parameters were determined with a help of modified Ho-Kashyap algorithm [10]. The Ho-Kashyap algorithm is a method of determining the weight vector of linear classifier. The minimized criterion of its modified version includes, controlled by regularization parameter, additional component responsible for reducing the classifier complexity.

3. RESULTS

The FCB clustering and sub-methods results. Figure 4 presents objects from the first learning subset of synthetic database *banana.* Objects from both classes are marked by star and plus symbols. The FCB clustering into 8 clusters is shown. Circles represent the initial, randomly chosen prototypes to start clustering, prototype moving process during iterations was marked by black lines and points. As it may be seen, most of prototypes moved towards boundary between classes. Eight prototypes obtained directly as a result of the FCB clustering are marked by triangles. Figures 5 and 6 show the same dataset and also eight, but final prototypes obtained using sub-method I (Figure 5) and sub-method II (Figure 6).

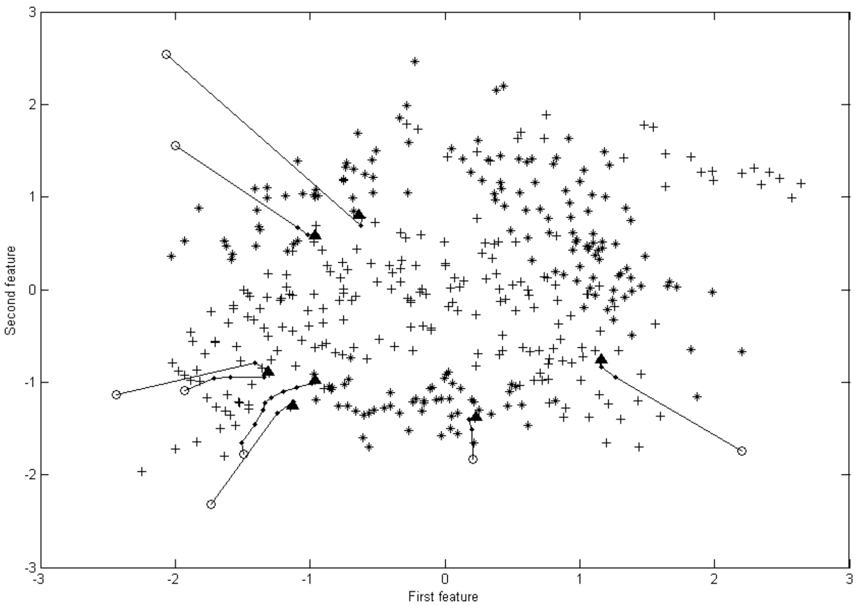

Fig. 4. The FCB clustering prototypes.

In case of prototypes from the Figure 4, two of them are very similar, so only seven triangles may be seen. Such situations was one of reasons for applying the sub-methods. Taking into consideration the correctness of prototypes location it should be stated that most of them are located near boundary between two classes, but they are not distant from each other. Another situation occurs in case presented in the Figure 5 – final prototypes are near boundary as well as they are far away from each other. There are not very similar prototypes. Analysis of prototypes from the Figure 6 may rise doubts about correctness of their location. However, it should be emphasized, that in case of sub-method II the final prototypes are common for all learning subsets, whereas Figures from 4 to 6 present only first subset. What is more important, the classification results are better when using final prototypes from sub-method II. Taking into account, that the presented FCB clustering and both sub-methods are dedicated to classification algorithms, the correctness of prototypes location should be evaluated mainly basing on classification results.

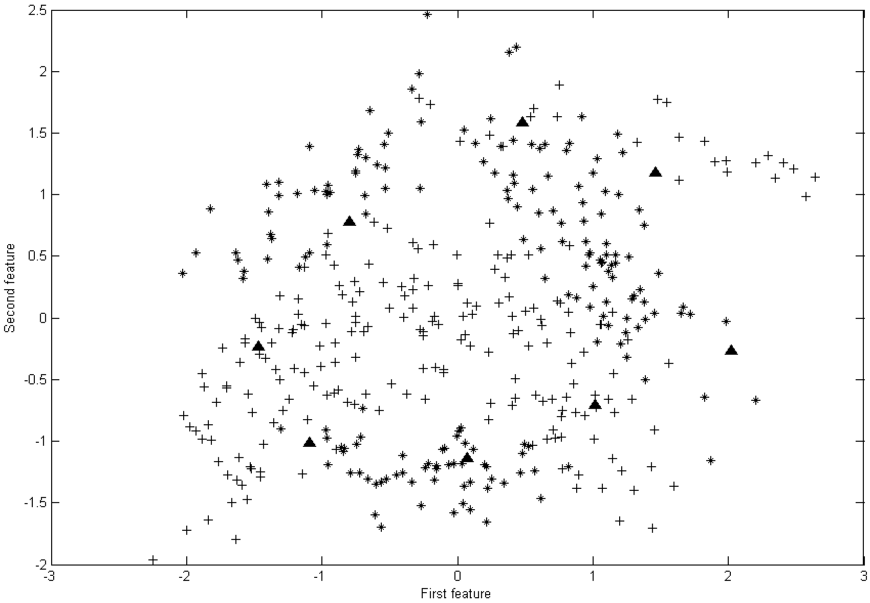

Fig. 5. Final prototypes from sub-method I.

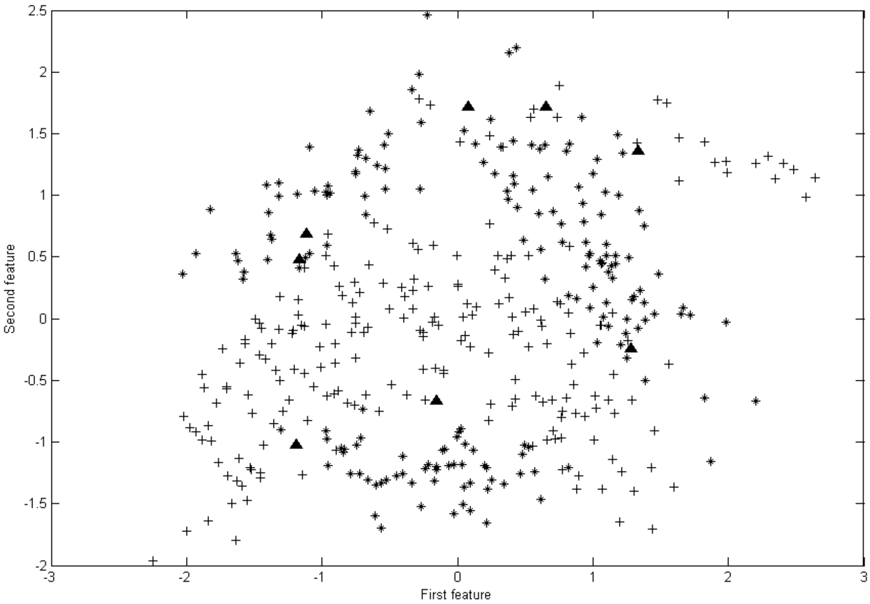

Fig. 6. Final prototypes from sub-method II.

Classification results. Quality of the classification based on the fuzzy if-then rules obtained from the final prototypes was compared with the results provided by the Lagrangian Support Vector Machines

(LSVM) method [11]. It is modified (faster) version of one of the best classifiers - Support Vector Machines. The LSVM parameters values were also determined by using the first ten pairs of learning and testing subsets.

Table 2 presents the classification results (mean classification error and its standard deviation for all 100 testing subsets) obtained by using both sub-methods and for three different numbers of intermediate prototypes (*ci*). Results for the LSVM are also shown. Following values of the FCB parameters were assumed: $m = 2$, $\alpha = 5000$, $L = 10$. Columns describes the following databases (beginning from the left): *banana*, *breast-cancer*, *diabetis*, *heart*, *Ripley data*, *titanic*, *thyroid*.

Table 2. Benchmark databases classification results, mean classification error (standard deviation).

		BAN	BRE	DIA	HEA	RPL	TIT	THY
LSVM		10.34 (0.43)	25.20 (3.96)	23.14 (1.69)	15.68 (3.35)	9.46 (0.55)	22.46 (1.25)	4.15 (2.31)
Sub-method I	***ci* = 50**	10.65 (0.47)	27.27 (4.68)	23.28 (1.76)	**15.61 (3.07)**	9.60 (0.56)	22.53 **(1.22)**	**3.92 (2.15)**
	***ci* = 200**	10.60 (0.47)	26.86 (4.96)	23.26 (1.75)	**15.61 (3.07)**	9.54 (0.56)	22.48 **(1.14)**	**3.27 (1.87)**
	***ci* = 1000**	10.58 (0.47)	25.94 (4.58)	23.21 (1.79)	**15.61 (3.07)**	9.53 (0.57)	22.47 (1.25)	**3.91 (2.20)**
Sub-method II	***ci* = 50**	10.67 (0.49)	**25.12** (4.22)	23.70 (3.63)	18.42* (8.49)	9.60 (0.55)	**21.96 (1.00)**	**3.12 (1.70)**
	***ci* = 200**	10.52 (0.49)	**25.05** (4.25)	**22.62** (1.74)	23.22* (12.99)	11.67* (8.86)	**22.26 (1.00)**	**3.69 (1.86)**
	***ci* = 1000**	10.57 (0.51)	25.29 (4.08)	25.52* (9.49)	**15.61 (3.07)**	9.59 **(0.53)**	**22.30 (1.00)**	**3.88 (2.18)**

The applied modified Ho-Kashyap algorithm requires calculation of the inverse of matrix. In four cases marked by asterisks in Table 2, problems with performing that operation were observed, which led to higher values of classification error and mainly its standard deviation. It should be reminded, that in sub-method II the final prototypes are chosen basing on classification of 10 first pairs of learning and testing subsets, and they are common for all 100 learning subsets. So it may happen, that they are not appropriate for some of the remaining 90 subsets and cause problems with calculating the inverse of matrix. The identical result in 4 cases of *heart* database should be also commented – detailed analysis showed, that in these cases the modified Ho-Kashyap algorithm led to local minimum.

In the presented work, only for synthetic databases – *banana* and *Ripley data*, the better (in relation to the LSVM) results were not obtained. For all other databases, which represent mainly medical data, the better results were achieved. In most cases, beside lower values of classification error, lower values of standard deviations were noticed. The better (in relation to LSVM) values are written in bold in Table 2. For *thyroid* database the better results were obtained in all six variants. For this database the highest improvement was achieved: 3.12% (1.70) in relation to 4.15% (2.31) for the LSVM. Taking this into account, the presented FCB clustering with additional sub-methods should be regarded as appropriate to application to classification algorithms. Presented methodology was also applied to classification of cardiotocographic signals for prediction of newborn condition [7,8], and results were also better than for the LSVM. The higher duration time of sub-method II resulted in high classification quality – as it may be seen in Table 2 – the best results were achieved when using the sub-method II.

4. CONCLUSIONS

In the proposed work, the fuzzy clustering method (FCB) dedicated to classification algorithms was described. Its goal is to find the prototypes which are located near boundary between two classes of objects. The method provides good results, however there are situations, when not all of obtained prototypes are located appropriately. Thus the FCB clustering is applied to obtain intermediate prototypes. Using these prototypes and the introduced additional sub-methods (clustering of intermediate prototypes or choice from intermediate prototypes), final prototypes determining classifier fuzzy if-then rules are obtained. Sub-methods were developed to be different in three important aspects concerning initial prototype matrix, prototypes creating and their assignment to learning subsets.

The usefulness of the presented methods to classification algorithms was proved by achieving high classification quality for different, including medical benchmark databases. In case of all medical databases, the obtained classification error was lower than provided by the one of the best classifiers – Lagrangian SVM (LSVM). The highest improvement was achieved for medical database *thyroid*: 3.12% (1.70) in relation to 4.15% (2.31) for the LSVM. The aim of the presented work was to initially verify usefulness of proposed methods to classification algorithms. Methods were applied only for benchmark datasets. The results for this data are encouraging, but further improvements and tests are necessary. In general, well developed and tested classifier may help in medical diagnosis. However, it is only clinician's support – the clinician's role is irreplaceable.

BIBLIOGRAPHY

[1] BEZDEK J.C., Pattern recognition with fuzzy objective function algorithms, Plenum Press, New York, London, 1982.

[2] CZABANSKI R., JEZEWSKI M., WROBEL J., KUPKA T., LESKI J., JEZEWSKI J., The prediction of the low fetal birth weight based on quantitative description of cardiotocographic signals, Journal of Medical Informatics and Technologies, Vol. 12, 2008, pp. 97-10.

[3] DUDA R.O., HART P.E., Pattern classification and scene analysis, John Wiley and Sons, New York, 1973.

[4] JEZEWSKI M., The prediction of fetal outcome with application of fuzzy clustering and classification methods, PhD Thesis, Silesian Univeristy of Technology, Gliwice, 2011.

[5] JEZEWSKI M., CZABANSKI R., HOROBA K., WROBEL J., LESKI J., JEZEWSKI J., Influence of gestational age on neural networks interpretation of fetal monitoring signals, Journal of Medical Informatics and Technologies, Vol. 12, 2008, pp. 137-142.

[6] JEZEWSKI M., HENZEL N., WROBEL J., LABAJ P., MATONIA A., Application of neural networks for prediction of fetal outcome, Proc. XI Conference on Medical Informatics and Technologies, Vol. 10, 2006, pp. 127-132.

[7] JEZEWSKI M., LESKI J., An application of fuzzy clustering method to cardiotocographic signals classification, Man-Machine Interactions 2, CZACHORSKI T., KOZIELSKI S., STANCZYK U. (Eds.), Advances in Intelligent and Soft Computing, Springer Verlag, Berlin Heidelberg, pp. 315-322.

[8] JEZEWSKI M., LESKI J., Cardiotocographic signals classification based on clustering and fuzzy if-then rules, Proc. 5th European Conference of IFMBE, 121-124.

[9] JEZEWSKI M., LESKI J., Fuzzy clustering finding prototypes on classes boundary, Computer Recognition Systems 4, BURDUK R., KURZYNSKI M., WOZNIAK M., ZOLNIEREK A. (Eds.), Advances in Intelligent and Soft Computing, Springer Verlag, Berlin Heidelberg, 2011, pp. 177-186.

[10] LESKI J., An ε-margin nonlinear classifier based on if-then rules, IEEE Transactions on Systems, Man and Cybernetics – Part B: Cybernetics, Vol. 34, No. 1, 2004, pp. 68-76.

[11] MANGASARIAN O.L., MUSICANT D.R., Lagrangian support vector machines, Journal of Machine Learning Research, Vol. 1, 2001, pp. 161-177.