clustering, classification, class boundaries

Jacek ŁĘSKI[1], Michał JEŻEWSKI[1]

CLUSTERING ALGORITHM FOR CLASSIFICATION METHODS

Classification plays an important role in many fields of life, including medical diagnosis support. In the paper, fuzzy clustering algorithm dedicated to classification methods is proposed. Its goal is to find pairs of prototypes located near boundaries of both classes of objects. The minimization procedure of the proposed criterion function is described. The algorithm for determining the value of the clustering parameter is also presented. Presented results (synthetic dataset) confirm correctness of clustering – most of final prototypes, determined based on obtained pairs, are located between boundary of two classes.

1. INTRODUCTION

Classification methods play an important role in many fields of science, including medical diagnosis support. Appropriately developed and tested classifier may help in detection of signs of distress basing on some patient's input data (examination results, medical images etc.). Many examples may be found [10, 11], one of them may be the prediction of newborn condition done during pregnancy [2, 3, 6]. The presented paper describes algorithm of clustering, but dedicated to classification methods. Therefore, the topic of classification [4] is also crucial in the presented work.

Clustering consists in finding groups (clusters) and their centers (prototypes) of similar objects in dataset. The role of clustering is also important in many fields of science. In case of computational intelligence methods, clustering may be applied to classification algorithms. For example, the extraction of fuzzy if-then rules may be done with a help of fuzzy clustering methods [5, 7, 8]. There are many types of clustering methods represented by various algorithms [8]. A popular type of clustering is clustering by minimization of the criterion function. In that case clustering results are presented by two matrices: partition matrix **U**, which describes membership degrees of objects to clusters, and prototype matrix **V**, which describes location of prototypes (**v**). Clustering by minimization of the criterion function consists in iteratively updating values of **U** and **V** matrices. The process starts from the randomly established values of one of matrices and stops after the stop condition (reaching the maximum number of iterations or lack of significant changes of the criterion value) is fulfilled. In case of fuzzy clustering, object (**x**) may belong to several clusters, with the membership degree value from 0 to 1.

The goal of the proposed algorithm of fuzzy clustering is to find pairs of prototypes located near boundaries of both classes of objects. The final prototypes, which should be located near boundary between classes, are determined based on obtained pairs. The outline of the proposed method was presented in [9].

2. CLUSTERING WITH PAIRS OF PROTYTYPES

The proposed method is based on minimization of the following criterion function

$$J,(\mathbf{U}^{(1)},\mathbf{U}^{(2)},\mathbf{V}^{(1)},\mathbf{V}^{(2)}) = \sum_{i=1}^{c}\sum_{\{k|\mathbf{x}_k\in\omega_1\}}(u_{ik}^{(1)})^m(d_{ik})^2 + \sum_{i=1}^{c}\sum_{\{k|\mathbf{x}_k\in\omega_2\}}(u_{ik}^{(2)})^m(d_{ik})^2 + \eta\sum_{i=1}^{c}\left\|\mathbf{v}_i^{(1)}-\mathbf{v}_i^{(2)}\right\|^2 \quad (1)$$

with the constraints

$$\forall_{\{k|\mathbf{x}_k\in\omega_1\}}\ \sum_{i=1}^{c}u_{ik}^{(1)} = 1, \quad (2)$$

[1] Institute of Electronics, Silesian University of Technology, Akademicka 16, 44-100 Gliwice, Poland

$$\forall_{\{k|\mathbf{x}_k\in\omega_2\}} \sum_{i=1}^{c} u_{ik}^{(2)} = 1, \tag{3}$$

The goal of the method is to find pairs of prototypes located near boundaries of both classes (ω_1 and ω_2) of objects. It is realized by clustering of objects from both classes separately, but with minimization of distances between prototypes in pairs – prototypes in pairs should move closer to each other and as a result should be located near boundaries of classes. Two first component represent classic fuzzy *c*-means clustering [1] and are responsible for clustering in both classes – upper indices (1) and (2) denote the first and second class. The third component ensures minimization of Euclidean distances between prototypes in pairs. Symbols have the following meaning: c denotes number of clusters (prototypes), m influences a fuzziness of clusters (usually m=2 is chosen and such value was assumed), $d_{ik}=\|\mathbf{x}_k-\mathbf{v}_i\|$ denotes Euclidean distance between the *i*th prototype and *k*th object. The η determines the proportion between clustering and minimizing distances between prototypes.

2.1. CRITERION MINIMIZATION PROCEDURE

Because of the constraints (2) and (3), to obtain necessary conditions for partition matrices $\mathbf{U}^{(1)}$ and $\mathbf{U}^{(2)}$, the Lagrange multipliers method was applied. If $\mathbf{V}^{(1)}$ and $\mathbf{V}^{(2)}$ are fixed, then columns of $\mathbf{U}^{(1)}$ and $\mathbf{U}^{(2)}$ are independent, and the minimization of (1) can be performed term by term

$$J(\mathbf{U}^{(1)},\mathbf{U}^{(2)}) = \sum_{\{k|\mathbf{x}_k\in\omega_1\}} g_k^{(1)}(\mathbf{U}^{(1)}) + \sum_{\{k|\mathbf{x}_k\in\omega_2\}} g_k^{(2)}(\mathbf{U}^{(2)}) + \eta\sum_{i=1}^{c}\left\|\mathbf{v}_i^{(1)}-\mathbf{v}_i^{(2)}\right\|^2, \tag{4}$$

where

$$\forall_{\{k|\mathbf{x}_k\in\omega_1\}} \quad g_k^{(1)}(\mathbf{U}^{(1)}) = \sum_{i=1}^{c}(u_{ik}^{(1)})^m (d_{ik})^2, \tag{5}$$

$$\forall_{\{k|\mathbf{x}_k\in\omega_2\}} \quad g_k^{(2)}(\mathbf{U}^{(2)}) = \sum_{i=1}^{c}(u_{ik}^{(2)})^m (d_{ik})^2, \tag{6}$$

The Lagrangian of (5) with constraints from (2) is

$$\forall_{\{k|\mathbf{x}_k\in\omega_1\}} \quad G_k^{(1)}(\mathbf{U}^{(1)},\lambda^{(1)}) = \sum_{\{k|\mathbf{x}_k\in\omega_1\}}(u_{ik}^{(1)})^m (d_{ik})^2 - \lambda^{(1)}\left[\sum_{i=1}^{c} u_{ik}^{(1)} - 1\right], \tag{7}$$

where $\lambda^{(1)}$ is the Lagrange multiplier. The Lagrangian of (6) takes the similar form

$$\forall_{\{k|\mathbf{x}_k\in\omega_2\}} \quad G_k^{(2)}(\mathbf{U}^{(2)},\lambda^{(2)}) = \sum_{\{k|\mathbf{x}_k\in\omega_2\}}(u_{ik}^{(2)})^m (d_{ik})^2 - \lambda^{(2)}\left[\sum_{i=1}^{c} u_{ik}^{(2)} - 1\right], \tag{8}$$

Setting the Lagrangian's gradients to 0 we obtain

$$\forall_{\{k|\mathbf{x}_k\in\omega_1\}} \quad \frac{\partial G_k^{(1)}(\mathbf{U}^{(1)},\lambda^{(1)})}{\partial\lambda^{(1)}} = \sum_{i=1}^{c} u_{ik}^{(1)} - 1 = 0, \tag{9}$$

$$\forall_{\{k|\mathbf{x}_k\in\omega_1\}} \forall_{1\le s\le c} \quad \frac{\partial G_k^{(1)}(\mathbf{U}^{(1)},\lambda^{(1)})}{\partial u_{sk}^{(1)}} = m(u_{sk}^{(1)})^{m-1} d_{sk}^2 - \lambda^{(1)} = 0, \tag{10}$$

Transformations of (9) and (10) lead to formula known from the fuzzy *c*-means algorithm

$$\bigvee_{1\le s\le c}\ \bigvee_{\{k|\mathbf{x}_k\in\omega_1\}} u_{sk}^{(1)}=\frac{(d_{sk})^{2/(1-m)}}{\sum_{j=1}^{c}(d_{jk})^{2/(1-m)}}, \tag{11}$$

After transformations of gradients of the Lagrangian for the second class (8) we obtain similar equation

$$\bigvee_{1\le s\le c}\ \bigvee_{\{k|\mathbf{x}_k\in\omega_2\}} u_{sk}^{(2)}=\frac{(d_{sk})^{2/(1-m)}}{\sum_{j=1}^{c}d_{jk}^{\ 2/(1-m)}}, \tag{12}$$

To obtain necessary conditions for prototype matrices $\mathbf{V}^{(1)}$ and $\mathbf{V}^{(2)}$, we calculate gradients of the criterion (1). For prototypes for the first class we obtain

$$\bigvee_{1\le s\le c}\ \frac{\partial J(\mathbf{U}^{(1)},\mathbf{U}^{(2)},\mathbf{V}^{(1)},\mathbf{V}^{(2)})}{\partial\mathbf{v}_s^{(1)}}=-2\sum_{\{k|\mathbf{x}_k\in\omega_1\}}(u_{ik}^{(1)})^m(\mathbf{x}_k-\mathbf{v}_s^{(1)})+2\eta(\mathbf{v}_s^{(1)}-\mathbf{v}_s^{(2)})=0, \tag{13}$$

and after transformations

$$\bigvee_{1\le s\le c}\ \mathbf{v}_s^{(1)}=\frac{\sum_{\{k|\mathbf{x}_k\in\omega_1\}}(u_{ik}^{(1)})^m\mathbf{x}_k+\eta\mathbf{v}_s^{(2)}}{\sum_{\{k|\mathbf{x}_k\in\omega_1\}}(u_{ik}^{(1)})^m+\eta}, \tag{14}$$

The formula for prototypes for the second class takes the similar form

$$\bigvee_{1\le s\le c}\ \mathbf{v}_s^{(2)}=\frac{\sum_{\{k|\mathbf{x}_k\in\omega_2\}}(u_{ik}^{(2)})^m\mathbf{x}_k+\eta\mathbf{v}_s^{(1)}}{\sum_{\{k|\mathbf{x}_k\in\omega_2\}}(u_{ik}^{(2)})^m+\eta}, \tag{15}$$

There are two equations (14) and (15) with two unknowns. Solving them leads to similar formulas defining prototypes for both classes

$$\bigvee_{1\le s\le c}\ \mathbf{v}_s^{(1)}=\frac{\sum_{\{k|\mathbf{x}_k\in\omega_1\}}(u_{ik}^{(1)})^m\mathbf{x}_k\left(\sum_{\{k|\mathbf{x}_k\in\omega_2\}}(u_{ik}^{(2)})^m+\eta\right)+\eta\sum_{\{k|\mathbf{x}_k\in\omega_2\}}(u_{ik}^{(2)})^m\mathbf{x}_k}{\left(\sum_{\{k|\mathbf{x}_k\in\omega_1\}}(u_{ik}^{(1)})^m+\eta\right)\left(\sum_{\{k|\mathbf{x}_k\in\omega_2\}}(u_{ik}^{(2)})^m+\eta\right)-\eta^2}, \tag{16}$$

$$\bigvee_{1\le s\le c}\ \mathbf{v}_s^{(2)}=\frac{\sum_{\{k|\mathbf{x}_k\in\omega_2\}}(u_{ik}^{(2)})^m\mathbf{x}_k\left(\sum_{\{k|\mathbf{x}_k\in\omega_1\}}(u_{ik}^{(1)})^m+\eta\right)+\eta\sum_{\{k|\mathbf{x}_k\in\omega_1\}}(u_{ik}^{(1)})^m\mathbf{x}_k}{\left(\sum_{\{k|\mathbf{x}_k\in\omega_1\}}(u_{ik}^{(1)})^m+\eta\right)\left(\sum_{\{k|\mathbf{x}_k\in\omega_2\}}(u_{ik}^{(2)})^m+\eta\right)-\eta^2}, \tag{17}$$

Analysis of equations (16) and (17) shows, that if $\eta=0$, the formulas known from the fuzzy *c*-means algorithm are obtained. Final prototypes should be located between each two prototypes making pairs. We propose the following solution

$$\bigvee_{1\le s\le c} \mathbf{v}_s = \frac{\sum_{\{k|\mathbf{x}_k\in\omega_1\}} (u_{ik}^{(1)})^m \mathbf{x}_k + \sum_{\{k|\mathbf{x}_k\in\omega_2\}} (u_{ik}^{(2)})^m \mathbf{x}_k}{\sum_{\{k|\mathbf{x}_k\in\omega_1\}} (u_{ik}^{(1)})^m + \sum_{\{k|\mathbf{x}_k\in\omega_2\}} (u_{ik}^{(2)})^m}, \tag{18}$$

The location of final prototypes depends on pairs of prototypes from both classes – the values of partition matrices in the equation above are updated basing on pairs of prototypes, (11) and (12). Another ways of determining final prototypes are the topic for future works.

2.2. DETERMINATION OF η

The η (etha) parameter in the proposed criterion (1) determines the proportion between clustering and minimizing distances between prototypes. The growth of the η should result in decrease of distances between prototypes. Boundary between classes region is characterized by high diversity of assignments of objects to classes. To obtain the value of the η ensuring the best location of final prototypes, the following algorithm was proposed:

Algorithm 1:

For each η value *(changed from 0.1 to 10 with a step 0.1)*:

1) cluster dataset into assumed number of clusters (c)
2) for each of c final prototypes find K nearest objects from dataset *(K=10 was assumed)*
3) calculate absolute sum (abs_sum) of their class labels
   Labels of objects from the first (second) class are converted to +1 (–1).
   The abs_sum takes values from 0 (even K) or 1 (odd K) to K with a step 2,
   for example for K=10 the abs_sum takes the value 0, 2, 4, 6, 8 or 10.
4) calculate diversity=abs_sum/c
   Dividing by c enables comparing quality of the diversity between different number of prototypes – the diversity value is always within the same range (described above).
   The lowest value (0 or 1) denotes the best diversity, the value equals to K denotes lack of diversity (all class labels equal to +1 or –1).

Chose the value of the η for the best diversity (or first for equal values).

3. RESULTS

The banana benchmark synthetic dataset was applied to present clustering results. The original dataset have 5300 objects assigned to two classes. To make graphical presentation of clustering possible, smaller dataset including 530 objects was created (each tenth object starting from the first was chosen). Figure 1 presents clustering into three clusters (η=7.1, basing on algorithm 1). Black (white) squares represent objects from the first (second) class. Starting the clustering into c clusters from c prototypes located in the diagonal (determined by minimum and maximum values of features) was assumed in all experiments. Three white triangles denote prototypes applied to start the clustering. Consecutive locations of prototypes during iterations are marked by black dots, connected by dashed lines. Black and white circles denote prototypes obtained after clustering (connected in pairs by dotted lines). They are located near boundaries of classes, it is especially visible in case of the first class of objects. Final prototypes, located near boundary between classes, are represented by black triangles.

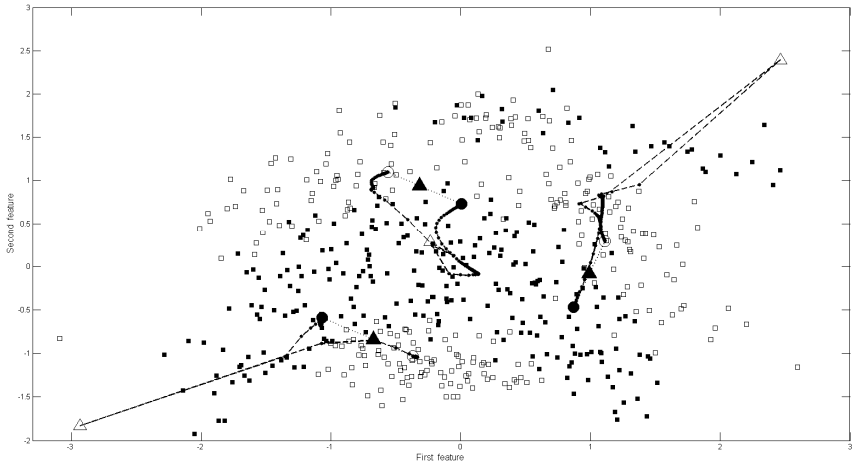

Fig. 1. Clustering into three clusters, illustration of the process

According to the assumed idea, the growth of the η resulted in decrease of distances between prototypes in pairs. It is illustrated in figures 2, 3 and 4, which present pairs of prototypes in clustering into 3 clusters with the η equals to 0.1, 3.5 and 7.1. As a result, the value of the third component (sum of distances, without multiplying by the η) of the criterion (1) also decreases – figure 5 presents this relation for three different number of prototypes. In case of clustering presented in figures 2-4, after reaching η=7.1, pairs of prototypes changed their location (as a result the location of final prototypes is better).

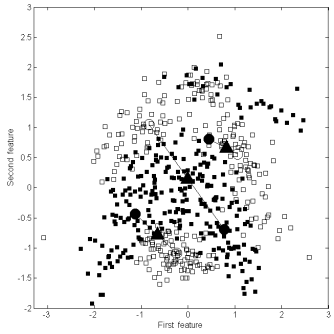

Fig. 2. Clustering with η=0.1

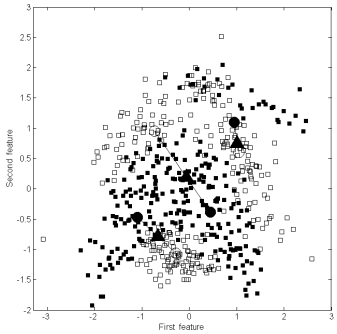

Fig. 3. Clustering with η=3.5

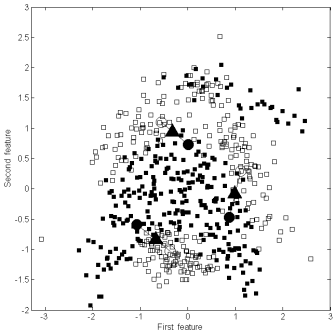

Fig. 4. Clustering with η=7.1

The solid line in figure 6 presents diversity values obtained applying the algorithm 1 to clustering into 3 clusters. The high decrease of diversity value is caused by the mentioned change of location of prototype pairs (fig. 4). Situations, when for given values of the η some of final prototypes were surrounded by all *K* objects belonging to the same class (diversity value=K, lack of diversity), were observed applying the algorithm 1. The number of them (within the range from 1 to no. of clusters) is represented by dashed line.

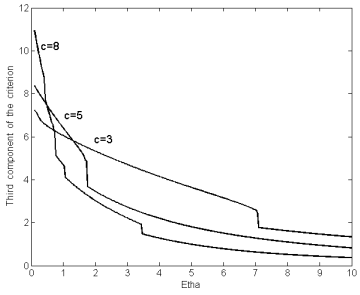

Fig. 5. The value of the third component of the criterion

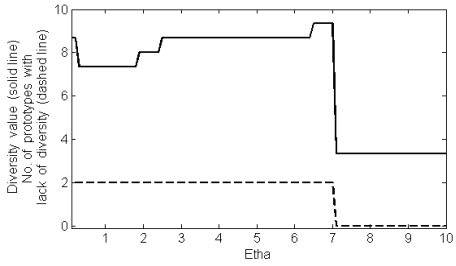

Fig. 6. Determining the value of the η

Lower charts in figures from 7 to 15 illustrate final prototypes obtained as a result of clustering into different number of clusters, from 2 to 20. With the growth of the number of final prototypes, they are located in consecutive boundary regions. The straight majority of them are located near boundary between classes. The determination of appropriate number of prototypes is a problem to be solved in future works. Upper charts in figures from 7 to 15 show values obtained applying the algorithm 1, its results are presented in Table 1.

Table 1. Results of the algorithm 1

Clusters	η	Diversity value	Prototypes with lack of diversity
2	6.5	2.0	0
3	7.1	3.3	0
4	1.4	6.5	1
5	2.2	2.8	0
6	2.9	4.0	1
7	1.2	4.3	0
8	0.8	4.8	1
12	1.2	4.7	2
16	0.5	3.8	2
20	0.5	5.1	3

For number of prototypes > 7, regardless of the η, at least one of them has lack of diversity. On the contrary, for number of prototypes ≤ 7 (except for 4), there were values of the η without prototypes with lack of diversity (for example in fig. 6 for η > 7.1). As a result of the algorithm 1 such values were chosen – the lowest diversity value was for the η without prototypes with lack of diversity. There is an exception for six prototypes, when the lowest diversity value was for the η=2.9 – with one prototype with lack of diversity. In that case the best diversity (diversity value=0) of three prototypes balances lack of diversity of one of them.

It is possible to reject final prototypes with lack of diversity. However, such approach was not assumed, because the lack of diversity indicated by the algorithm 1 may not mean bad location of prototype. What is more, prototype with lack of diversity may be located in

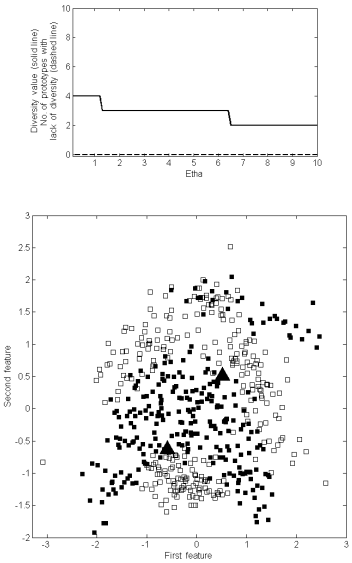

Fig. 7. Clustering into 2 clusters

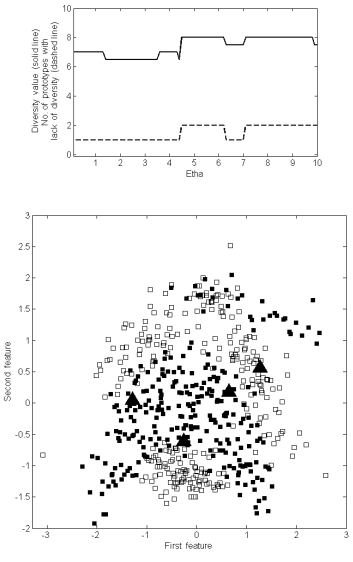

Fig. 8. Clustering into 4 clusters

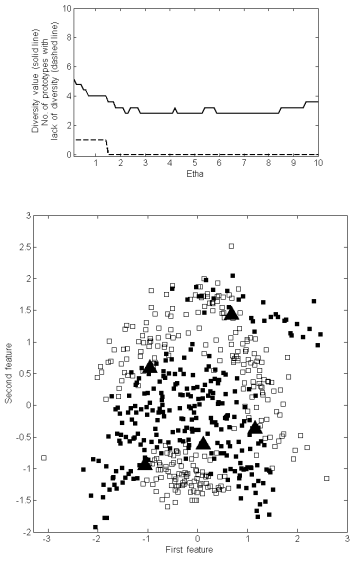

Fig. 9. Clustering into 5 clusters

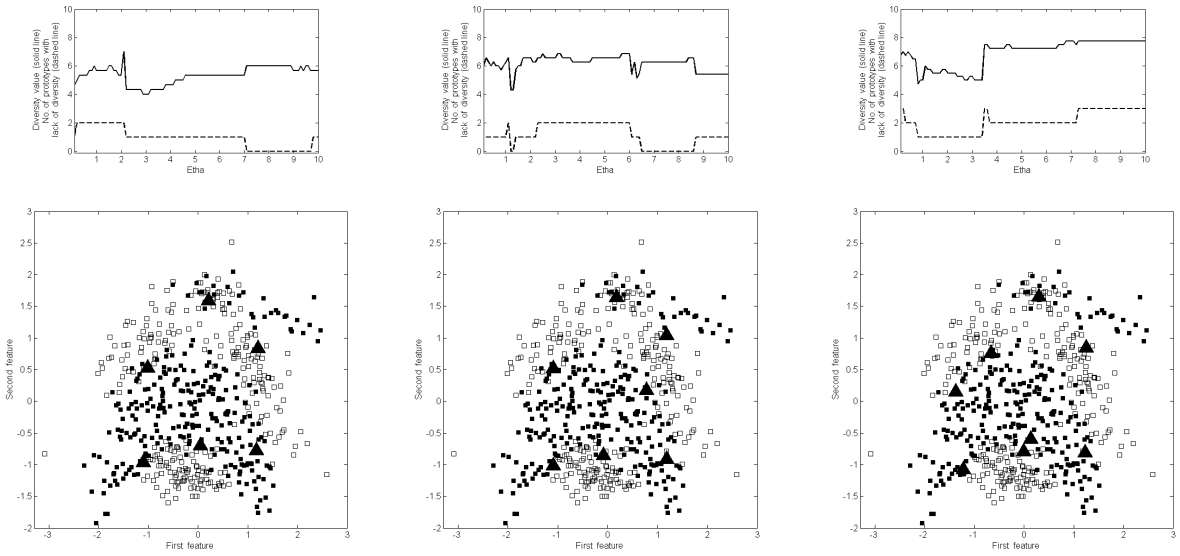

Fig. 10. Clustering into 6 clusters

Fig. 11. Clustering into 7 clusters

Fig. 12. Clustering into 8 clusters

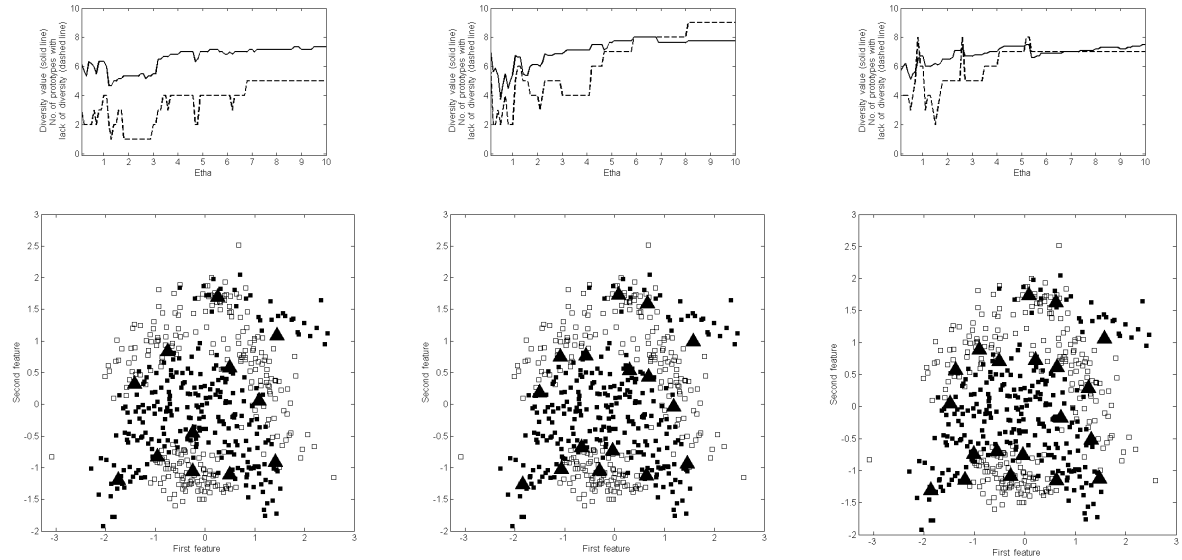

Fig. 13. Clustering into 12 clusters

Fig. 14. Clustering into 16 clusters

Fig. 15. Clustering into 20 clusters

boundary – when two classes are distant from each other and the prototype is between them, but closer to one of them. There is such situation in clusterings into 4, 6 and 8 clusters, and concerns the prototype with coordinates approximately equal to [1,1]. On the other hand, prototypes with coordinates [-2,-1.25] (approximately) in figures 13-15 were correctly indicated as with lack of diversity. For reasons above, another algorithms for determining the value of the η are plans for future works.

4. CONCLUSIONS

The algorithm of fuzzy clustering dedicated to classification methods was proposed. Its goal is to find pairs of prototypes located near boundaries of both classes of objects. Presented results obtained using benchmark synthetic dataset confirm correctness of clustering – most of final prototypes, determined basing on obtained pairs, are located near boundary between classes. The algorithm determining the value of the clustering parameter was proposed, but it should be improved.

Future works will concern modification of proposed clustering and its application to developing the nonlinear classifier. In general, appropriately developed and tested classifier may help in medical diagnosis, but 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 Press1982, New York, London.
[2] CZABANSKI R, JEZEWSKI J., MATONIA A., JEZEWSKI M., Computerized analysis of fetal heart rate signals as the predictor of neonatal acidemia, Expert Systems with Applications, 2012, Vol. 39, pp. 11846-11860.
[3] CZABANSKI R., JEZEWSKI M., WROBEL J., JEZEWSKI J., Predicting the risk of low-fetal birth weight from cardiotocographic signals using ANBLIR system with deterministic annealing and ε-insensitive learning, IEEE Transactions on Information Technology in Biomedicine, 2010, Vol. 14(4), pp. 1062-1074.
[4] DUDA R.O., HART P.E., Pattern classification and scene analysis, John Wiley and Sons, 1973, New York.
[5] JEZEWSKI M., An application of modified fuzzy clustering to medical data classification, Journal of Medical Informatics and Technologies, 2011, Vol. 17, pp. 51-57.
[6] JEZEWSKI M, CZABANSKI R., WROBEL J., HOROBA K., Analysis of extracted cardiotocographic signal features to improve automated prediction of fetal outcome, Biocybernetics and Biomedical Engineering, 2010, Vol. 30(4), pp. 29-47.
[7] LESKI J., An ε-margin nonlinear classifier based on if-then rules, IEEE Transactions on Systems, Man and Cybernetics – Part B: Cybernetics, 2004, Vol. 34, No. 1, pp. 68-76.
[8] LESKI J., Neuro-fuzzy systems, WNT, Warsaw, 2008, (in Polish).
[9] LESKI J., JEZEWSKI M., Fuzzy clustering with prototype pairs, X Jubilee Scientific Seminary "Selected problems of electrotechnology and electronics" (WZEE 2012), (submitted paper).
[10] OGIELA M.R., TADEUSIEWICZ R., Pattern recognition, clustering and classification applied to selected medical images, Studies in Computational Intelligence, Vol. 84, 2007, Springer-Verlag, Berlin Heidelberg.
[11] PANDEY B., MISHRA R.B., Knowledge and intelligent computing system in medicine, Computers in Biology and Medicine, 2009, Vol. 39, pp. 215-230.