Clustering of Incomplete Data and Evaluation of Clustering Quality

Vladimir V. Ryazanov

Institution of Russian Academy of Sciences Dorodnicyn Computing Centre of RAS
Vavilov st. 40, 119333 Moscow, Russia
http://www.ccas.ru

Abstract. Two approaches to solving the problem of clustering with gaps for a specified number of clusters are considered. The first approach is based on restoring the values of unknown attributes and solving the problem of clustering of calculated complete data. The second approach is based on solving a finite set of tasks of clustering of corresponding to incomplete data complete sample descriptions and the construction of collective decision. For both approaches, the clustering quality criteria have been proposed as functions of incomplete descriptions. Results of practical experiments are considered.

Keywords: clustering, missing data, gaps, clustering estimation.

1 Introduction

The problem of cluster analysis of incomplete data has high interest of researchers, since the real practical problems are usually with missing data.

Clustering is usually performed in two stages. Incomplete data is first converted to the full data, and then there is clustering of complete data. In recent years, various algorithms have been developed to the construction of complete data. Conventionally, they can be divided into two types: marginalization and imputation. In the first case, the objects with missing data (gaps) are removed simply from the sample. In the second case, the unknown values of features are replaced by best match estimates [1,2]. The simplest imputation methods are replacement gaps on statistical estimates of the average values of attributes (means, random, the nearest neighbor method, etc.). Their generalizations involve averaging of feature values in the neighborhood of the objects with gaps [3]. Many algorithms use regression models. Unknown feature value is calculated using the regression function was found from the known characteristics (linear regression, SVR [4]). Estimation minimization (EM) algorithm is based on probability model of dataset. It is the well-known and popular in this field. Based on the use of imputation \marginalization technique clustering has both advantages and disadvantages. It should be noted that the rate of missing values of features is usually assumed to be low. In constructing the regression model, it is assumed there are a sufficient number of objects without gapes. In these approaches some

L. Alvarez et al. (Eds.): CIARP 2012, LNCS 7441, pp. 146–153, 2012.

information is lost. The advantage of these methods is their simplicity and the possibility of further use of standard software clustering of complete data. Frequently, the finding the estimates of unknown data is of independent interest.

Second approach to the clustering of incomplete data is to adapt the clustering methods to cases of incomplete data. This case does not require reconstruction of missing data. The paper [5] proposed the modification of fuzzy k-means clustering in case of missing data. There are some assumptions in this approach. The attribute with missing data linearly depends on the other features. Some parameters in distances calculation are the independent and identically distributed. The proposed method performs better results for some medical task in comparison with other imputation technique. Two methods for partitioning incomplete data set including missing values into linear fuzzy clusters by using local principal components have been proposed in [6]. One is the direct extension of fuzzy c- varieties method to an incomplete data set. It uses the least square criterion as the objective function. The other method is a hybrid technique of fuzzy clustering and principal component analysis with missing values. The direct clustering method has been proposed in [7]. For the constraining features the set of constraints based on known values is generated. Although there are already different approaches to solving the problem of clustering of incomplete data, the creation of new algorithms is till now an urgent task.

Another important aspect of missing data clustering is to assess the clustering quality as a function of the degree of incompleteness of data. Suppose that for some sample $X = \{\bar{x}_1, \bar{x}_2, ..., \bar{x}_m\}$ of incomplete data clustering $K = \{K_1, K_2, ..., K_l\}$ has been obtained. Let the sample $X' = \{\bar{x}'_1, \bar{x}'_2, ..., \bar{x}'_m\}$ of full descriptions corresponds to an initial sample X , and $K' = \{K'_1, K'_2, ..., K'_l\}$ is its clustering. What will be the "scatter" of the set of all admissible clusterings K' regarding clustering K? It is clear that the " scatter " must depend on many factors such as the clustering algorithm, data, rate of unknown characteristics, information content of missing data, etc.

In this paper we consider two problems associated with clustering of incomplete data: algorithms for clustering of incomplete data,and estimation of the quality of clustering as a function of data incompleteness.

In the first approach, some imputation is used. Degree of clustering certainty is calculated as an estimation of the stability of the obtained clustering result with respect to some sets of admissible complete sample. The second approach does not provide for reconstruction of features. At first, N complete samples which correspond to a given sample of partial descriptions are constructed. Next we solve independently N tasks of clustering and N clusterings are found. Finally, a collective solution of clustering task is computed. The degree of certainty of clustering is computed on the basis of estimation of the scatter difference of partial solutions with respect collective solution. The results of the comparison of degree of clustering certainty for different samples at different levels of data incompleteness are considered.

2 Clustering of Incomplete Data Based on the Features Imputation

Let a standard sample X of incomplete descriptions $\bar{x}_i = (x_{i1}, x_{i2}, ..., x_{in})$ of the objects in terms of features is given. We suppose the set $M_j \subseteq R, j = 1, 2, ..., n$ to be a finite set of values of $j-$th feature. It can be calculated by known feature values from training data. The unknown feature values (slips, gaps) will be denoted as Δ. We believe that $x_{ij} = \Delta, \forall j \in \Omega_i \subseteq \{1, 2, ..., n\}, i = 1, ..., m$. The set of unknown feature values is denoted as the set of pairs $J = \{\langle i, j\rangle, i = 1, 2, ..., m, j \in \Omega_i\}$. We use the local method of filling the gaps [8]. Obtained as a result sample of full descriptions will be denoted as $X^* = \{\bar{x}_1^*, \bar{x}_2^*, ..., \bar{x}_m^*\}$.

Let we solve a task of clustering of the sample X^* to l clusters using an algorithm A: $K = \{K_1, K_2, ..., K_l\}$, $K_i \subseteq X^*, i = 1, 2, ..., l$, $\bigcup_{i=1}^{l} K_i = X^*$,$K_i \bigcap K_j = \emptyset, i \neq j$. Denote $D_t = \{\bar{x}'_t\}$ the set of all possible $\bar{x}'_t$ corresponding to vector $\bar{x}_t$ (i.e. $x'_{tj} = \begin{cases} x_{tj}, & x_{tj} \neq \Delta, \\ \in M_j, & x_{tj} = \Delta. \end{cases}$).

Consider an arbitrary $\bar{x}_t^*$. Let $\bar{x}_t^* \in K_i$. Consider the partition $K' = \{K'_1, K'_2, ..., K'_l\}$ of sample $X' = X^* \backslash \{\bar{x}_t^*\} \bigcup \{\bar{x}'_t\}$, where $K'_j = K_j$, $j \neq i$, and $K'_i = K_i \backslash \{\bar{x}_t^*\} \bigcup \{\bar{x}'_t\}$, $\bar{x}'_t \in D_t$. Let $f_t(K)$ is the proportion of objects $\bar{x}'_t$ from D_t for which the partition K' is the result of clustering .

Definition 1. Degree of certainty $f(K)$ of the clustering K is the quantity $f(K) = \frac{1}{m} \sum_{t=1}^{m} f_t(K)$.

Consider the problem of calculating of $f(K)$ on the example of two well-known algorithms.

2.1 Clustering of Incomplete Data as the Minimization of Variance Criterion

It is known [9] that the condition for local optimality of clustering $K = \{K_1, K_2, ..., K_l\}$ with minimum value of the variance criterion is execution of the inequality

$$\frac{n_i}{(n_i - 1)} \left\|\bar{x}^* - \bar{m}_i^*\right\|^2 - \frac{n_j}{(n_j + 1)} \left\|\bar{x}^* - \bar{m}_j^*\right\|^2 \leq 0 \tag{1}$$

for any pair K_i, K_j, and any $\bar{x}^* \in K_i$ (here $n_i = |K_i|$, $\bar{m}_i^* = \frac{1}{n_i} \sum_{\bar{x}^* \in K_i} \bar{x}^*$). We will use $\|\bar{x} - \bar{y}\| = \rho(\bar{x}, \bar{y}) = \sqrt{\sum_{j=1}^{n} (x_j - y_j)^2}$).

Let $\bar{x}'_t = (x'_{t1}, x'_{t2}, ..., x'_{tn}) \in D_t$ is an arbitrary admissible vector corresponding to the vector $\bar{x}_t^* = (x_{t1}^*, x_{t2}^*, ..., x_{tn}^*) \in K_i$. We obtain the conditions under which the partition K' is the clustering. To do this, let's write the conditions (1) for all objects from X'. Denote $\delta\bar{x}_t = \bar{x}_t^* - \bar{x}'_t$, then

$$\bar{m}'_i = \bar{m}_i^* - \frac{\delta\bar{x}_t}{n_i} \tag{2}$$

Partition K' is the clustering if the following conditions are satisfied:

$$\bar{x}'_t \in K'_i, \frac{n_i}{(n_i - 1)} \left\|\bar{x}'_t - (\bar{m}_i^* - \frac{\delta\bar{x}_t}{n_i})\right\|^2 - \frac{n_j}{(n_j + 1)} \left\|\bar{x}'_t - \bar{m}_j^*\right\|^2 \leq 0; \tag{3}$$

$$\forall \bar{x}'_\alpha \in K_i, \alpha \neq t, \frac{n_i}{(n_i-1)}\left\|\bar{x}^*_\alpha - (\bar{m}^*_i - \frac{\delta\bar{x}_t}{n_i})\right\|^2 - \frac{n_j}{(n_j+1)}\left\|\bar{x}^*_\alpha - \bar{m}^*_j\right\|^2 \leq 0; \quad (4)$$

$$\forall \bar{x}'_\alpha \in K_j, j \neq i, \frac{n_j}{(n_j-1)}\left\|\bar{x}^*_\alpha - \bar{m}^*_j)\right\|^2 - \frac{n_i}{(n_i+1)}\left\|\bar{x}^*_\alpha - \bar{m}^*_i + \frac{\delta\bar{x}_t}{n_i}\right\|^2 \leq 0. \quad (5)$$

Given (2),(3 - 5) can be rewritten as

$$\frac{n_i}{n_i-1}\left\|\bar{x}^*_t - \bar{m}^*_i\right\|^2 - \frac{n_j}{(n_j+1)}\left\|\bar{x}^*_t - \bar{m}^*_j\right\|^2 +$$

$$+ 2(\delta\bar{x}_t, \bar{m}^*_i - \frac{n_j}{(n_j+1)}\bar{m}^*_j - \frac{1}{(n_j+1)}\bar{x}^*_t) + \left\|\delta\bar{x}_t\right\|^2 \frac{(n_i - n_j - 1)}{n_i(n_j+1)} \leq 0, \quad (6)$$

$$\frac{n_i}{n_i-1}\left\|\bar{x}^*_\alpha - \bar{m}^*_i\right\|^2 - \frac{n_j}{(n_j+1)}\left\|\bar{x}^*_\alpha - \bar{m}^*_j\right\|^2 +$$

$$+ 2(\delta\bar{x}_t, \frac{1}{(n_i-1)}(\bar{x}^*_\alpha - \bar{m}^*_i)) + \left\|\delta\bar{x}_t\right\|^2 \frac{1}{n_i(n_i-1)} \leq 0, \quad (7)$$

$$\frac{n_j}{n_j-1}\left\|\bar{x}^*_\alpha - \bar{m}^*_j\right\|^2 - \frac{n_i}{(n_i+1)}\left\|\bar{x}^*_\alpha - \bar{m}^*_i\right\|^2 -$$

$$- 2(\delta x_t, \frac{(x^*_\alpha - m^*_i)}{(n_i+1)}) - \left\|\delta x_t\right\|^2 \frac{1}{n_i(n_i+1)} \leq 0, \quad (8)$$

System (6 - 8) can be written as (9). Thus, the partition K' is the clustering, if for fixed $\bar{x}'_t$ system of m inequalities (9) is performed,

$$a_\lambda + \sum_{i \in \Omega_t} y_i c_{\lambda i} + b_\lambda \sum_{i \in \Omega_t} y_i^2 \leq 0, \lambda = 1, 2, ..., m, \quad (9)$$

where $a_\lambda, b_\lambda, c_{\lambda i}, i = 1, 2, ..., k, \lambda = 1, 2, ..., m$ are constants for found K, and $y_i = \{x^*_{ti} - x'_{ti} : x'_{ti} \in M_i\}$. To calculate $f_t(K)$ we make enumeration for all admissible y_i (the systems (9) and calculate the number of executed systems (9). With a large enumeration, we estimate $f_t(K)$ on a random sample of allowed values of $f_t(K)$.

2.2 Clustering of Incomplete Data Using *k*-means Algorithm

Let K be the clustering X^* using k-means algorithm [9]. This means, $\forall \bar{x}^*_t \in K_i$ there is

$$\left\|\bar{x}^*_t - \bar{m}^*_i\right\| \leq \left\|\bar{x}^*_t - \bar{m}^*_j\right\|, \forall j \neq i, \quad (10)$$

Partition K' is the clustering if

$$\left\|\bar{x}^*_t - \delta\bar{x}_t - \bar{m}^*_i + \frac{\delta\bar{x}_t}{n_i}\right\|^2 \leq \left\|\bar{x}^*_t - \delta\bar{x}_t - \bar{m}^*_j\right\|^2, \bar{x}'_t \in K'_i, j \neq i, \quad (11)$$

$$\left\|\bar{x}^*_\alpha - \bar{m}^*_i + \frac{\delta\bar{x}_t}{n_i}\right\|^2 \leq \left\|\bar{x}^*_\alpha - \bar{m}^*_j\right\|^2, j \neq i, \forall \bar{x}'_\alpha \in K_i, \alpha \neq t, \quad (12)$$

$$\left\|\bar{x}^*_\alpha - \bar{m}^*_j\right\|^2 \leq \left\|\bar{x}^*_\alpha - \bar{m}^*_i + \frac{\delta\bar{x}_t}{n_i}\right\|^2, \forall \bar{x}'_\alpha \in K_j, j \neq i. \quad (13)$$

After elementary transformations we obtain a system similar to (9). The calculation of $f_t(K)$ is also carried out similarly.

3 Clustering of Sample with Missing Data without Imputation

By using X we make N samples of full descriptions $X'^{(i)} = \{\bar{x}_1'^{(i)}, \bar{x}_2'^{(i)}, ..., \bar{x}_m'^{(i)}\}, i = 1, 2, ..., N$, where $x_{tj}'^{(i)} = \begin{cases} x_{tj}, & x_{tj} \neq \varDelta, \\ \in M_j, & x_{tj} = \varDelta \end{cases}$ (probability of assigning a value from M_j to $x_{tj}'^{(i)}$ is equal to its frequency of occurrence in the training sample X). For each of the resulting complete samples, we solve the problem of clustering on l clusters and find N solutions $K^{(i)} = \{K_1^{(i)}, K_2^{(i)}, ..., K_l^{(i)}\}, i = 1, 2, ..., N$. Further, the collective clustering $K = \{K_1, K_2, ..., K_l\}$ is build and considered as a solution of the clustering task with missing data.

Denote $< t_1, t_2, ..., t_l >$ a permutation of $< 1, 2, ..., l >$.

Definition 2. Degree of certainty $\varPhi(K)$ of clustering $K = \{K_1, K_2, ..., K_l\}$ is the quantity

$$\varPhi(K) = \sum_{i=1}^{N} \max_{<t_1, t_2, ..., t_l>} \sum_{j=1}^{l} \left| K_j \bigcap K_{t_j}^{(i)} \right| / mN.$$

Definition 3. Degree of certainty $F(K)$ of clustering $K = \{K_1, K_2, ..., K_l\}$ is the quantity $F(K) = \min_{i=1,...,N} \max_{<t_1, t_2, ..., t_l>} \sum_{j=1}^{l} \left| K_j \bigcap K_{t_j}^{(i)} \right| / m.$

Quantity $\max_{<t_1, t_2, ..., t_l>} \sum_{j=1}^{l} \left| K_j \bigcap K_{t_j}^{(i)} \right|$ characterizes the proximity of clustering results K and $K^{(i)}, i = 1, 2, ..., N$. Criterion $\varPhi(K)$ characterizes the normalized average the proximity of collective clustering with respect to clustering of admissible samples. Criterion $F(K)$ meets the worst case.

The task of collective clustering construction and the committee algorithm for its solution were proposed in [10,11]. Earlier, a collective clustering for some sample was based on using some algorithm that combines the set of clusterings obtained for the same sample by different clustering methods. In our case, the collective solution will be built as application of some clustering algorithm to the set of clusterings obtained by fixed method for different full samples $X'^{(i)}$. The results of the clustering of samples $X'^{(i)}$ by some clustering method can be written in the form of three-dimensional information matrix

$$\left\| \alpha_{ij}^{\nu} \right\|_{m \times l \times N}, \alpha_{ij}^{\nu} \in \{0, 1\}, \sum_{j=1}^{l} \alpha_{ij}^{\nu} = 1, i = 1, ..., m, j = 1, ..., l, \nu = 1, ..., N.$$

Its submatrix $\left\| \alpha_{ij}^{\nu} \right\|_{l \times N}$, $i = 1, 2, ..., m$, can be regarded as a new description of $\bar{x}_i$. As a collective solution of main cluster analysis task was considered the clustering of given m matrix descriptions by k - means method.

4 The Results of Experiments on Simulated and Practical Data

Proposed clustering algorithms of incomplete data were tested on simulated and practical problems. As a model problem we used a sample mixture of normal distributions with independent features $(n = 10, l = 4, m_i = 50, i = 1, 2, 3, 4)$. Expectation and variance of the classes are chosen such that the result of their clustering coincided with their a priori classification. Considered training samples were transformed into descriptions of the samples with gaps at various levels of data incompleteness. Unknown values of features in each training object on the uniform law of distribution is set, and $w\%$ feature values were unknown. Separately, we solved the problem of clustering of incomplete samples by means of collective clustering. Visualization of a model example for the four classes (the "projection" of the multidimensional data on the plane of generalized features [9]) is shown in Fig. 1. Fig. 3,4 demonstrate the proposed criteria $f(K)$,$\Phi(K)$, $F(K)$ and parameters $\varphi(K) = \max\limits_{<t_1,t_2,...,t_l>} \sum_{j=1}^{l} |K_j \bigcap K^*| /m$, $\varphi_{avr}(\ K) = \max\limits_{<t_1,t_2,...,t_l>} \sum_{j=1}^{l} \left|K_j'' \bigcap K^*\right| /m$ as functions of incompleteness rate w. Here $K^* = \{K_1^*, K_2^*, ..., K_l^*\}$ is a priori classification of the initial sample, K is a collective clustering of sample with gaps, and $K'' = \{K_1^{''}, K_2^{''}, ..., K_l^{''}\}$ is a sample clustering after replacing the gaps on the average feature values. In Figures 1-2 and 3-4 , respectively, visualizations and graphics of a model and the practical task of "breast cancer" [12] are shown. Task "breast cancer" is a sample of 344 descriptions of patients with benign or malignant tumor $(n = 9, l = 2, m_1 = 218, m_2 = 126)$. The task has a cluster structure that agrees well with a priori classification. The experimental results are the preliminary, but the form of obtained dependences corresponds to a priori expectations. Criterion $F(K)$ corresponds to the worst-case of possible clusterings $K^{(i)}, i = 1, 2, ..., N$, and its value decreases rapidly with w increasing.

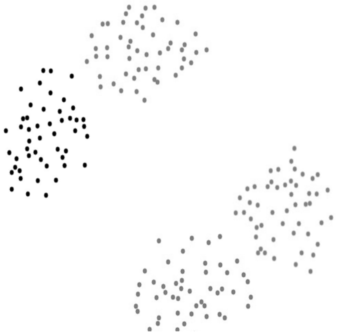

Fig. 1. Mixture of normal distributions

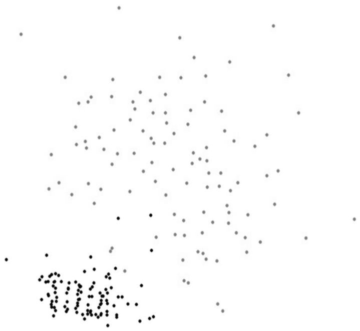

Fig. 2. The problem of breast cancer

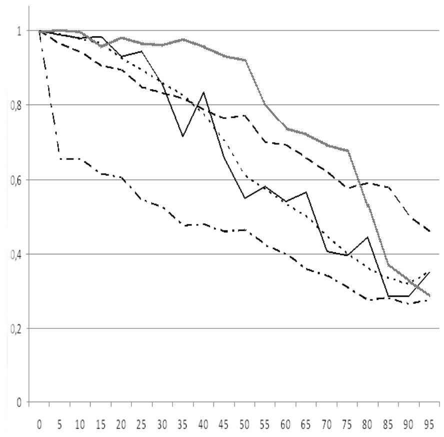

Fig. 3. Dependencies of criteria and parameters for incomplete data in the model task

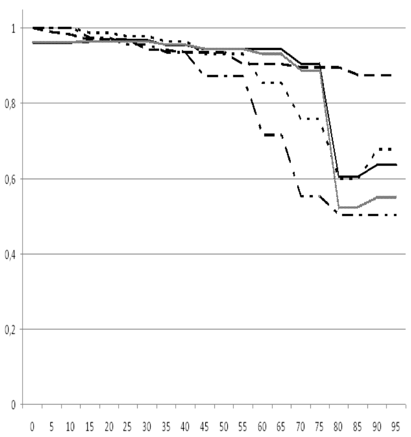

Fig. 4. Dependencies of criteria and parameters for incomplete data in the task of breast cancer

Table 1. Notations of graphics of criteria and indicators

———	index $\varphi(\ K)$	- - - - - - - - -	criterion $\varPhi(K)$
———	index$\varphi_{avr}(\ K)$	▬ ▬ ▬ ▬ ▬	criterion $f(K)$
▬ · ▬ · ▬ · ▬	criterion$F(K)$		

Nevertheless, it can be very useful in practice. The beginning of its fall corresponds to the maximal level of missing data, in which the incompleteness of the data does not affect to clustering. In the task "breast cancer" clusters are well separated. They are calculated even with a gaps rate 45% when $F(K)$ begins to decrease sharply. High values of $\varphi_{avr}(\ K)$ are the outcome of the simplicity of the structures of data. Criteria $\varphi(\ K)$ and $\varPhi(K)$ are well correlated, due, apparently, to the use of collective clustering. The graphics show also the correlation of the criteria $f(K)$ and $\varPhi(K)$. Value of the criterion $\varPhi(K)$ seems more objective than of $f(K)$, as the $f(K)$ calculation is based on variations of only the individual objects.

5 Conclusion

In report [13], it was proposed a leave-one-out approach to evaluation of clustering quality that had been based on an estimation of the clustering stability. Clustering quality evaluation belongs to the interval [0,1] and is not associated with any a priori classifications or probability nature of data. It is interesting to study the relationships between the criteria [13] and the criteria of degree of clustering certainty introduced in this paper. The proposed criteria are simple and interpretable. Of course, the results of experiments are preliminary. Nevertheless, we hope that the methods of clustering of incomplete data and criteria for evaluating the degree of clustering certainty proposed here will be useful in solving practical tasks.

Acknowledgments. I would like to thank students of M.V.Lomonosov Moscow State University A.S. Schichko and K.V.Tishin for experiments performed. This work was supported by RAS Presidium program number 15 and "Basic Sciences for Medicine, Program number 2 of Department of Mathematical Sciences of RAS, RFBR 12-01-00912, 11-01-00585.

References

1. Little, R.J.A., Rubin, D.B.: Statistical Analysis with Missing Data. Wiley, New York (1987)
2. Zloba, E.: Statistical methods of reproducing of missing data. J. Computer Modelling & New Technologies 6(1), 51–61 (2002)
3. Zhang, S.: Parimputation: From imputation and null-imputation to partially imputation. IEEE Intelligent Informatics Bulletin 9(1), 32–38 (2008)
4. Honghai, F., Guoshun, C., Cheng, Y., Bingru, Y., Yumei, C.: A SVM Regression Based Approach to Filling in Missing Values. In: Khosla, R., Howlett, R.J., Jain, L.C. (eds.) KES 2005. LNCS (LNAI), vol. 3683, pp. 581–587. Springer, Heidelberg (2005)
5. Sarkar, M., Leong, T.-Y.: Fuzzy k-means Clustering with Missing Values. In. AMIA Symp., pp. 588–592 (2001)
6. Honda, K., Ichihashi, H.: Linear Fuzzy Clustering Techniques With Missing Values and Their Application to Local Principal Component Analysis. IEEE Transactions on Fuzzy Systems 12(2), 183–193 (2004)
7. Wagstaff, K.: Clustering with missing values: No imputation required. In: Meeting of the International Federation of Classification Societies "Classification, Clustering, and Data Mining", pp. 649–658. Springer (2004)
8. Ryazanov, V.: Some Imputation Algorithms for Restoration of Missing Data. In: San Martin, C., Kim, S.-W. (eds.) CIARP 2011. LNCS, vol. 7042, pp. 372–379. Springer, Heidelberg (2011)
9. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification, 2nd edn. Wiley Interscience (2001)
10. Ryazanov, V.V.: The committee synthesis of pattern recognition and classification algorithms, Zh. Vychisl. Mat. i Mat. Fiziki 21(6), 1533–1543 (1981) (in Russian) (Printed in Great Britain, 1982. Pergamon Press. Ltd.)
11. Biryukov, A.S., Ryazanov, V.V., Shmakov, A.S.: Solving Clusterization Problems Using Groups of Algorithms. Zh. Vychisl. Mat. i Mat. Fiziki 48(1), 176–192 (2008) (Printed in Great Britain, 2008. Pergamon Press. Ltd.)
12. Mangasarian, O.L., Wolberg, W.H.: Cancer diagnosis via linear programming. SIAM News 23(5), 1–18 (1990)
13. Arseev, A.S., Kotochigov, K.L., Ryazanov, V.V.: Universal criteria for clustering and stability problems. In: 13th All-Russian Conference "Mathematical Methods for Pattern Recognition", pp. 63–64. S.-Peterburg (2007) (in Russian)