Chapter 12

Collective Solutions on Sets of Stable Clusterings

Vladimir Vasilevich Ryazanov

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.76189

Abstract

Two clustering problems are considered. We consider a lot of different clusters of the same data for a given number of clusters. Data clustering is understood as their stable partition into a given number of sets. Clustering is considered stable if the corresponding partitioning remains unchanged with its minimum change. How to create a new cluster based on ensemble clusterings? The second problem is the following. A definition of the committee synthesis as ensemble clustering is introduced. The sets of best and worst matrices of estimates are considered. Optimum clustering is built on the basis of the clusterings obtained as being closest to the set of the best estimation matrices or as the most distant from the set of worst-case matrices of estimates. As a result, the problem of finding the best committee clustering is formulated as a discrete optimization problem on permutations.

Keywords: clustering, algorithm, ensemble, collective, stability, optimality, construction

1. Introduction

There are many different approaches to solving the problems of clustering multidimensional data: based on the optimization of internal criteria (indices) [1, 2], hierarchical clustering [3], centroid-based clustering [4], density-based clustering [5], distribution-based clustering [6], and many others. There are well-known books and papers on clustering [7–10].

This section is devoted to one approach to the creation of stable clusterings and the processing of their sets. A natural criterion is considered, which is applicable to any clustering method. In work [11], various criteria (indices) are proposed, optimizing which clustering is built with a definite look "what is clustering?" In this chapter, we use a criterion based on stability. If we really got clustering, that is, a solution for the whole sample, the partitioning should not

change with a small change in the data. Criteria are introduced for the quality of the partition obtained. If the criterion value is less than one, then the partition is unstable. Let us obtain for the same data N clusterings. How to create a new ensemble clustering based on the N partitions? Previously, a committee method for building ensemble clusterings was proposed [12–15]. Let there be N results of cluster analysis of the same data for l clusters. The committee method of building ensemble clustering makes it possible to build such l clusters, each of which is the intersection of "many" initial clusters. In other words, we find such l clusters whose objects are "equivalent" to each other according to several principles. As initial N clusterings, one can take stable ones. Finally, we consider a video-logical approach to building the initial N coarse clusterings.

2. Criteria for stability of clustering

Let the sample of objects $X = \{\mathbf{x}_i, i = 1, 2, \ldots, m\}$, $\mathbf{x}_i \in R^n$ be given and $K = \{K_1, K_2, \ldots, K_l\}$ is the clustering of the sample into l clusters obtained by some method, $K_i \subseteq X$, $i = 1, 2, \ldots, l$, $\cup_1^l K_i = X$, $K_i \cap K_j = \varnothing$, $i \neq j$. Speaking of clustering, we mean applying a method to a sample without focusing on the method itself. Is partition K of a sample by this method clustering or here some kind of stopping criterion is satisfied? For example, an extremum of some functional is obtained or the maximum number of operations in the iterative process is fulfilled. We will use the following thesis as the main one. If the resulting partition K is indeed clustering, then it must be the same clustering for any minimal change in the sample X. Let $\mathbf{x}_i$ be arbitrary, $\mathbf{x}_i \in K_\alpha$ ensemble then the sample $X \backslash \{\mathbf{x}_i\}$ partition $K^*(\mathbf{x}_i) = \{K_1^*, K_2^*, \ldots, K_l^*\}$, $K_j^* = K_j$, $j = 1, 2, \ldots, l$, $j \neq \alpha$, $K_\alpha^* = K_\alpha \backslash \{\mathbf{x}_i\}$, $i = 1, 2, \ldots, m$ must be clustering. The fact of "coincidence" of clusterings $K = \{K_1, K_2, \ldots, K_l\}$ and $K^*(\mathbf{x}_i) = \{K_1^*, K_2^*, \ldots, K_l^*\}$ will be called identity, the clusterings themselves are identical and denoted it as $K^*(\mathbf{x}_i) \approx K$. In this case, it is natural to call a partition K as stable clustering if the partitions $K^*(\mathbf{x}_i)$ and K coincide for all $\mathbf{x}_i$, $i = 1, 2, \ldots, m$. In the case of non-identity of some individual $K^*(\mathbf{x}_i)$ with K, we will call K as quasi-clustering.

Definition 1. The quality of quasi-clustering (of unstable clustering) is the quantity $\Phi(K) = |\{x_i, i = 1, 2, \ldots, m : K^*(\mathbf{x}_i) \approx K\}|/m$.

If $\Phi(K) = 1$, then in this case, we will talk about stable clustering K or simply clustering. Suppose that for some $i, i = 1, 2, \ldots, m$ the condition $K^*(\mathbf{x}_i) \approx K$ is not satisfied, and $K^\circ(\mathbf{x}_i) = \{K_1^\circ, K_2^\circ, \ldots, K_l^\circ\}$ is the clustering of the sample $X \backslash \{\mathbf{x}_i\}$ obtained from the partition $K^*(\mathbf{x}_i)$ using $K^*(\mathbf{x}_i)$ as the initial approximation. Then $K^\circ(\mathbf{x}_i)$ can significantly differ from $K^*(\mathbf{x}_i)$. We will use as a function of the proximity between clustering $K^\circ(\mathbf{x}_i)$ and partitioning K the value $d(K^\circ(\mathbf{x}_i), K) = \max_{\boldsymbol{\alpha}} \sum_{i=1}^l \left|K_i^\circ \cap K_{\alpha_i}\right|/(m-1)$. Note that to calculate proximity it is required to find the maximum matching in a bipartite graph, for which there is a polynomial algorithm [16]. If $K^\circ(\mathbf{x}_i)$ does not exist, we will assume that $d(K^\circ(\mathbf{x}_i), K) = 0$.

Definition 2. The quality $F_{\min}(K)$ of the quasi-clustering K will be called the quantity $F_{\min}(K) = \min_i d(K^\circ(\mathbf{x}_i), K)$.

Definition 3. The quality $F_{avr}(\mathrm{K})$ of the quasi-clustering K will be called the quantity $F_{avr}(\mathrm{K}) = \sum_{i=1}^{m} d(\mathrm{K}^{\circ}(\mathbf{x}_i), \mathrm{K})/m$.

For some clustering algorithms, there are simple economical rules for computing $\Phi(\mathrm{K})$. Let us bring them (see also in [3, 17, 18]).

2.1. Method of minimizing the dispersion criterion

It is known that in order to minimize the dispersion criterion, it suffices to satisfy inequalities

$$\frac{n_j}{(n_j-1)}\|\mathbf{x}^{\times}-\mathbf{m}_j\|^2 - \frac{n_k}{(n_k+1)}\|\mathbf{x}^{\times}-\mathbf{m}_k\|^2 \le 0 \tag{1}$$

for any clusters K_j and K_k, arbitrary $\mathbf{x}^{\times} \in K_j$, where $n_j = \left|K_j\right|$, $\mathbf{m}_j = \frac{1}{n_j}\sum_{\mathbf{x}_t \in K_j}\mathbf{x}_t$.

We establish the conditions for the identity $\mathrm{K}^*(\mathbf{x}_i) \approx \mathrm{K}$ of the partitions $\mathrm{K}^*(\mathbf{x}_i)$ and K. In the case $\mathbf{x}^{\times} \in K_j$ [considering (Eq. (1))] to satisfy the condition $\mathrm{K}^*(\mathbf{x}_i) \approx \mathrm{K}$ inequalities.

$\frac{(n_j-1)}{(n_j-2)}\|\mathbf{x}^{\times}-\mathbf{m}_j\|^2 + \frac{2}{(n_j-2)}(\mathbf{x}^{\times}-\mathbf{m}_j, \mathbf{x}_i-\mathbf{m}_j) + \frac{1}{(n_j-1)(n_j-2)}\|\mathbf{x}_i-\mathbf{m}_j\|^2 - \frac{n_k}{n_k+1}\|\mathbf{x}^{\times}-\mathbf{m}_k\|^2 \le 0$ must be satisfied. In the case $\mathbf{x}^{\times} \in K_k$ inequalities $\frac{n_k}{(n_k-1)}\|\mathbf{x}^{\times}-\mathbf{m}_k\|^2 - \frac{(n_j-1)}{n_j}\|\mathbf{x}^{\times}-\mathbf{m}_j\|^2 - \frac{2}{n_j}(\mathbf{x}^{\times}-\mathbf{m}_j, \mathbf{x}_i-\mathbf{m}_j) - \frac{1}{n_j(n_j-1)}\|\mathbf{x}_i-\mathbf{m}_j\|^2 \le 0$ must be satisfied.

2.2. *k*-means method

Let the clustering K be obtained by *k*-means method, that is, $\|\mathbf{x}^{\times}-\mathbf{m}_j\| \le \|\mathbf{x}^{\times}-\mathbf{m}_k\|$, $\forall j \ne k$, $\forall \mathbf{x}^{\times} \in K_j$. In the case of equality, the object is considered to belong to a cluster with a lower number. Then, $\mathrm{K}^*(\mathbf{x}_i) \approx \mathrm{K}$ is satisfied if $\|\mathbf{x}^{\times}-\mathbf{m}_j\|^2 + \frac{2}{(n_j-1)}(\mathbf{x}^{\times}-\mathbf{m}_j, \mathbf{x}_i-\mathbf{m}_j) + \frac{1}{(n_j-1)^2}\|\mathbf{x}_i-\mathbf{m}_j\|^2 \le \|\mathbf{x}^{\times}-\mathbf{m}_k\|^2$ under $\mathbf{x}^{\times} \in K_j, \mathbf{x}^{\times} \ne \mathbf{x}_i$ and $\|\mathbf{x}^{\times}-\mathbf{m}_k\|^2 \le \|\mathbf{x}^{\times}-\mathbf{m}_j\|^2 + \frac{2}{(n_j-1)}(\mathbf{x}^{\times}-\mathbf{m}_j, \mathbf{x}_i-\mathbf{m}_j) + \frac{1}{(n_j-1)^2}\|\mathbf{x}_i-\mathbf{m}_j\|^2$ under $\mathbf{x}^{\times} \in K_k$.

2.3. Method of hierarchical agglomeration grouping

We confine ourselves to the case of an agglomeration hierarchical grouping. To find the value of the criterion $\Phi(\mathrm{K})$, you can calculate the partitioning K, partitions $\mathrm{K}^{\circ}(\mathbf{x}_i)$, $i = 1, 2, \ldots, m$, and compare K with each $\mathrm{K}^{\circ}(\mathbf{x}_i)$, $i = 1, 2, \ldots, m$. Here it is possible to save in the calculation of $\Phi(\mathrm{K})$ without carrying through the clustering for some of "i". Indeed, let there $\mathrm{K}^t(\mathbf{x}_i) = \{K_1^t, K_2^t, \ldots, K_{m-t}^t\}$ be clustering of the sample $X\backslash\{\mathbf{x}_i\}$ into $m-t$ clusters, $t \le m-l$. K is a partition obtained by the clustering algorithm X. The main property of the hierarchical grouping is that for any $k = 1, 2, \ldots, m-t$ there is $j = 1, 2, \ldots, m-t-1$ for which $K_k^t \subseteq K_j^{t+1}$. In this case, if at some step t, $t \le m-l$ for some k the condition $K_k^t \subseteq K_j$ does not hold for all $j = 1, 2, \ldots, l$, then the condition $\mathrm{K}^*(\mathbf{x}_i) \approx \mathrm{K}$ will not be fulfilled.

2.4. Examples

We give some examples illustrating the stability criteria introduced.

1. Below are the results obtained for model samples. The method of clustering based on the minimization of the dispersion criterion [3] has been used. As the initial data, we used samples of a mixture of two two-dimensional normal distributions with independent features, different $\mathbf{a}$, and $\boldsymbol{\sigma}$. Examples are shown in **Figures 1–3** (images of the samples in question) and in **Tables 1** and **2**. **Figure 1** represents a sample of 200 objects for which all the criteria $\Phi(\mathrm{K})$, $F_{\min}(\mathrm{K})$, $F_{avr}(\mathrm{K})$ are equal to 1, and the resulting clustering into two clusters is stable clustering. Here we used distributions with parameters $\mathbf{a}_1 = (0,0)$, $\mathbf{a}_2 = (9,9)$, and $\boldsymbol{\sigma}_1 = \boldsymbol{\sigma}_2 = (3,3)$.

Further, with the same parameters $\mathbf{a}_1$, $\mathbf{a}_2$, experiments were carried out for $\boldsymbol{\sigma}_1 = \boldsymbol{\sigma}_2 = (5,5)$.

Then, we used distributions with parameters $\mathbf{a}_1 = (0,0)$, $\mathbf{a}_2 = (9,9)$, $\boldsymbol{\sigma}_1 = \boldsymbol{\sigma}_2 = (10,10)$, $m = 200$. In this case, we have the case of strongly intersecting distributions. Formally, the clustering method gives a quasi-clustering, approximately corresponding to the partitioning of the original sample (**Figure 3**) into two sets by a diagonal from the upper left corner of the picture to the lower right. The values of the criteria in **Table 2** were obtained.

2. Data clustering of [19] and criteria values $\Phi(\mathrm{K})$, $F_{\min}(\mathrm{K})$, $F_{avr}(\mathrm{K})$. The following data from classification problem of electromagnetic signals were considered: $n = 34$, $m_1 = 225$, $m_2 = 126$, $l = 2$. We give the values of the stability criteria obtained. **Figure 4** shows the visualization [3] of the sample. The accuracy of the supervised classification methods was about 87% of the correct answers. However, the clustering of data turned out to be only quasi-clustering (**Table 3**).

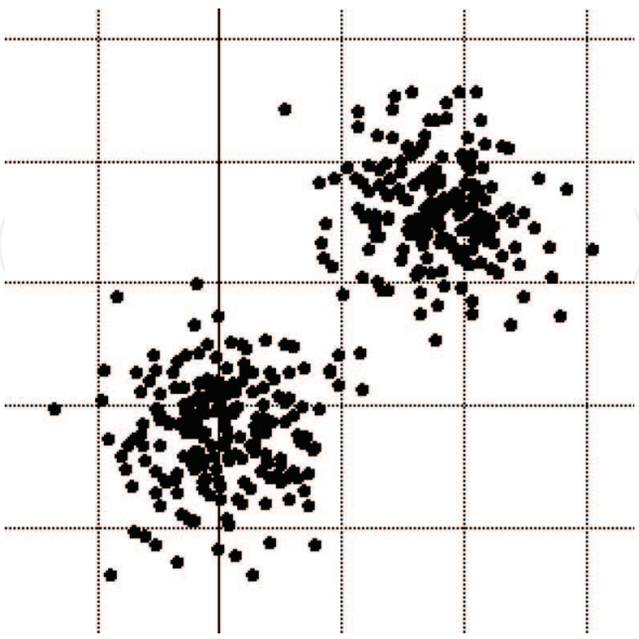

Figure 1. Clustering in a task with parameters $\mathbf{a}_1 = (0,0)$, $\mathbf{a}_2 = (9,9)$, $\boldsymbol{\sigma}_1 = \boldsymbol{\sigma}_2 = (3,3)$, $m = 200$.

http://dx.doi.org/10.5772/intechopen.76189

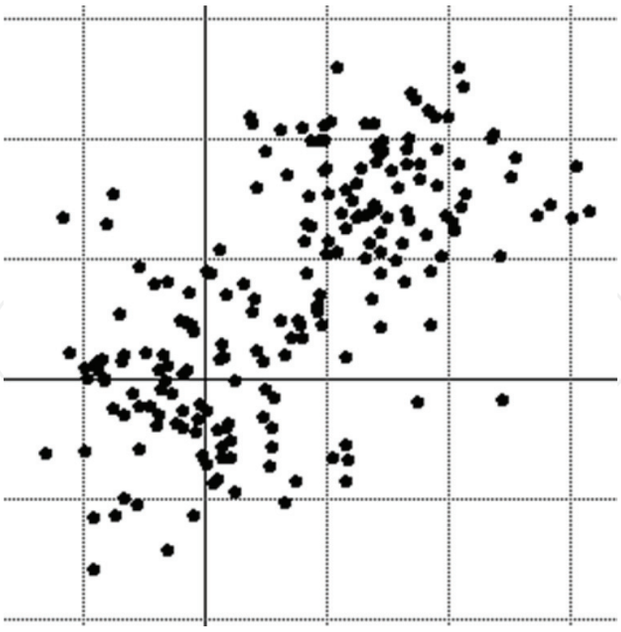

Figure 2. Clustering in a task with parameters $\mathbf{a}_1 = (0,0)$, $\mathbf{a}_2 = (9,9)$, $\boldsymbol{\sigma}_1 = \boldsymbol{\sigma}_2 = (5,5)$, $m = 200$.

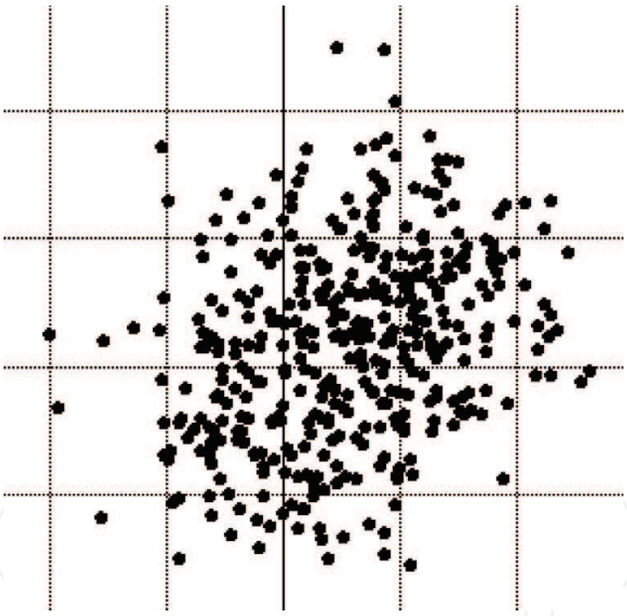

Figure 3. Data with parameters $\mathbf{a}_1 = (0,0)$, $\mathbf{a}_2 = (9,9)$, $\boldsymbol{\sigma}_1 = \boldsymbol{\sigma}_2 = (10,10)$, $m = 200$.

$\Phi(\mathrm{K})$	0.995
$F_{\min}(\mathrm{K})$	0.995
$F_{avr}(\mathrm{K})$	0.999

Table 1. Values of quasi-clustering criteria.

Φ(K)	0.770
$F_{\min}$(K)	0.995
F_{avr}(K)	0.998

Table 2. Values of quasi-clustering criteria. Case of very intersecting distributions

Figure 4. Data visualization.

Φ(K)	0.966
$F_{\min}$(K)	0.997
F_{avr}(K)	0.999

Table 3. The values of the criteria in the problem "ionosphere" Φ(K), $F_{\min}$(K), F_{avr}(K).

3. Committee synthesis of ensemble clustering

The problem is as follows. There are N clusterings for the same number of clusters. How to choose from them the only one or build a new clustering from the available ones? In the supervised classification problem (with the help of a collective solution of a set of algorithms) there is a criterion according to which one can choose an algorithm from existing ones or build a new algorithm. This is a supervised classification error. This direction in the theory of classification appeared in the early 1970s of the last century [20, 21], then was created an algebraic approach [22], various correctors were appeared. The key in the algebraic approach is the creation in the form of special algebraic polynomials of a correct (error-free) algorithm based on a set of supervised classification algorithms. Some algebraic operations on matrices of "degrees of belonging" of recognized objects are used. Various types of correctors were also created [22–25], when the problem of constructing (and applying) the best algorithm is also solved in two stages. First, the supervised classification algorithms are determined, and then the corrector. This can be, for example, the problem of approximating a given partial Boolean function by some monotonic function. In recent decades, there are conferences on multiple classifier systems, these issues are reflected in the books [21, 10]. How to choose or create the

best clustering using a finite set of given solutions? Here, all problems are connected primarily with the absence of a single generally accepted criterion. Each clustering algorithm finds such "source" clusters of objects that are "equivalent" to each other. In this chapter, it is proposed to build such a clustering of the initial data, the cluster solutions of which have a large intersection with the initial clusters.

Let the sample of objects $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\}$, $\mathbf{x}_i \in R^n$ for supervised classification and l classes are given. In the theory of supervised classification, the following definition of the supervised classification algorithm exists [21]. Let $\alpha_{ij} \in \{0, 1\}$ be equal to 1 when the object $\mathbf{x}_i$, $i = 1, 2, \ldots, m$ is classified by the algorithm A^r as $\mathbf{x}_i \in K_j$ and 0 otherwise: $A^r(X) = \|\alpha_{ij}\|_{m\times l}$. Here the intersection of classes is allowed. Unlike the supervised classification problem, when clustering a sample, we have freedom in the designation of clusters.

Definition 4. The matrices $I = \|\alpha_{ij}\|_{m\times l}$, $\alpha_{ij} \in \{0, 1\}$ and $I' = \|\alpha'_{ij}\|_{m\times l}$, $\alpha'_{ij} \in \{0, 1\}$ are said to be equivalent if they are equals to within a permutation of the columns.

It is clear that this definition defines a class of equivalent matrices for some matrix.

Definition 5. A clustering algorithm is an algorithm that maps a sample X to a set of equivalent information matrices $A^c(X) = K(\|\alpha_{ij}\|_{m\times l})$.

The number of clusters and the length of the control sample are considered to be given. This definition emphasizes the fact that in an arbitrary partition of a sample into l clusters, we have complete freedom in the numbering of clusters. In what follows we shall always consider matrices of dimension $m \times l$.

Let there be given N algorithms $A_1^c, A_2^c, \ldots, A_N^c$ for clustering and their solutions $A_\nu^c(X) = K\left(\|\alpha_{ij}^\nu\|_{m\times l}\right)$ for sample X. We denote $I_\nu = \|\alpha_{ij}^\nu\|_{m\times l}$ an arbitrary element of the clustering $K\left(\|\alpha_{ij}^\nu\|_{m\times l}\right)$.

Therefore, we have $I = K(I_1) \times K(I_2) \times \ldots \times K(I_N)$ or set $I = \left\{\left(I'_1, I'_2, \ldots, I'_N\right), I'_\nu \in K(I_\nu)\right\}$, $I'_\nu = \|\alpha'^\nu_{ij}\|_{m\times l}$.

There are two problems.

1. Construction of the mapping I on, K_c, $I \to K_c = \left\{K\|c_{ij}\|_{m\times l}\right\}$, $c_{ij} \in \{0, 1\}$ (that is, the construction of some kind of clustering).
2. Finding the optimal element in K_c (i.e. finding the best clustering in K_c).

Definition 6. An operator $B\left(I'_1, I'_2, \ldots, I'_N\right) = B = \|b_{ij}\|_{m\times l}$ is called an adder if $b_{ij} = \sum_{\nu=1}^N \alpha'^\nu_{ij}$.

It is clear that $0 \le b_{ij} \le N$, $b_{ij} \in \{0, 1, 2, \ldots, N\}$.

Definition 7. An operator r is called a threshold decision rule, if $r(B) = C = \|c_{ij}\|_{m\times l}$, $c_{ij} = \begin{cases} 1, & b_{ij} \geq \delta_i, \\ 0, & \text{otherwise}, \end{cases}$ where $\delta_i \in R$.

Definition 8. By the committee synthesis of an information matrix C on an element $\tilde{I}' = \left(I'_1, I'_2, \ldots, I'_N\right)$ let us call it a computation by the formula $C = r\mathrm{B}\left(\tilde{I}'\right)$, provided that B is the adder and r is the threshold decision rule.

The general scheme of collective synthesis is shown in **Figure 5**.

We note that the total number of possible values B is bounded from above by a quantity $(l!)^N$. Let s be the operator that performs permutation of columns of matrices $m \times l$ with the help of a substitution $< j_1, j_2, \ldots, j_l >$, $S = \{s\}$ is the set of all operators s. We believe that $rs = sr, \forall s \in S$. We continue $s \in S$ to the n-dimensional case $\sigma\left(\tilde{I}'\right) = \left(s\left(I'_1\right), s\left(I'_2\right), \ldots, s\left(I'_n\right)\right)$. We denote $\Sigma = \{\sigma\}, \sigma$ is the extension of s. From the definition of the adder it follows that $\sigma\mathrm{B} = \mathrm{B}\sigma, \forall\sigma \in \Sigma$. Further, $\forall\tilde{I}' \in \mathrm{I}, \forall\sigma \in \Sigma$ we have $r\mathrm{B}\left(\sigma\left(\tilde{I}'\right)\right) = r\sigma\left(\mathrm{B}\left(\tilde{I}'\right)\right) = s\left(r\mathrm{B}\left(\tilde{I}'\right)\right)$ and finally $\left\{\sigma\left(\tilde{I}'\right), \sigma \in \Sigma\right\} \overset{r\mathrm{B}}{\to} \left\{s\left(r\mathrm{B}\left(\tilde{I}'\right)\right),\ s \in S\right\} = \mathrm{K}\left(r\mathrm{B}\left(\tilde{I}'\right)\right) = \mathrm{K}(\|c_{ij}\|_{m\times l})$. Therefore, the product rB defines the desired mapping and specifies some ensemble clustering. It is necessary to determine the optimal element from K_c, find it and $\tilde{I}'$.

$$\mathrm{I} \overset{r\mathrm{B}}{\to} \mathrm{K}_c,\ A^c_{\tilde{I}'}(\mathrm{X}) = \mathrm{K}\left(r\mathrm{B}\left(\tilde{I}'\right)\right).$$

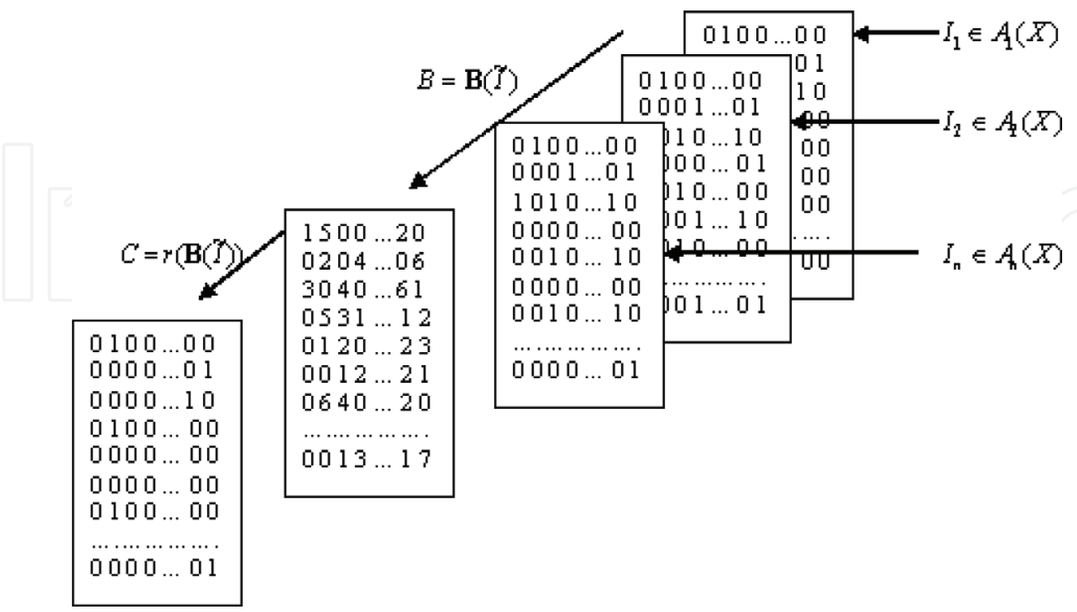

Figure 5. Scheme of committee synthesis.

We introduce definitions of potentially best and worst-case solutions. As the "ideal" of the collective solution, we will consider the case when all algorithms give us essentially the same partitions or coverings.

Definition 9. A numerical matrix $\|b_{ij}\|_{m\times l}$ is called contrasting if $b_{ij} \in \{0, N\}$. A numeric matrix $\|b_{ij}\|_{m\times l}$ is called blurred if $b_{ij} = \delta_i \in R$.

As the distance between two numerical matrices, we consider the function

$$\rho\left(B^1, B^2\right) = \sum_{i=1}^{m}\sum_{j=1}^{l}\left|b_{ij}^1 - b_{ij}^2\right|.$$

Denote by M the set of all contrast matrices, and by $\tilde{M}$ the set of all blurred matrices. We introduce definitions for estimating the quality of matrices.

Definition 10.

$$\Phi(B) = \rho(B, \mathrm{M}) \underset{B}{\rightarrow} \min. \tag{2}$$

Definition 11.

$$\tilde{\Phi}(B) = \rho\left(B, \tilde{\mathrm{M}}\right) \underset{B}{\rightarrow} \max. \tag{3}$$

The set $\tilde{\mathrm{M}}' = \left\{\tilde{B}\right\}$ (where $\tilde{B} = \|\tilde{b}_{ij}\|_{m\times l}$, $\tilde{b}_{ij} = \frac{N}{2}$) is called the mean blurred matrix.

Definition 12.

$$\tilde{\Phi}'(B) = \rho\left(B, \tilde{B}\right) \underset{B}{\rightarrow} \max \tag{4}$$

We note that the optimums according to the criteria (Eq. (2)) and (Eq. (3)) do not have to coincide. The sets M and $\tilde{\mathrm{M}}$ intersect.

Figure 6 illustrates the sets of contrasting and blurred matrices. Arrows indicate some elements of sets.

Theorem 1. The sets of optimal solutions by criteria Eqs. (2) and (4) coincide.

Let us show that $\Phi(B)+\tilde{\Phi}'(B) = \frac{Nml}{2}$ for any B. We write $\tilde{\Phi}'(B) = \sum_{i=1}^{m}\sum_{j=1}^{l}\tilde{\alpha}_{ij}$, $\tilde{\alpha}_{ij} = \left|b_{ij} - \frac{N}{2}\right|$, $\Phi(B)$ $= \sum_{i=1}^{m}\sum_{j=1}^{l}\alpha_{ij}^*$, $\alpha_{ij}^* = \min\left(b_{ij}, N - b_{ij}\right)$. If $b_{ij} \geq \frac{N}{2}$ then $\tilde{\alpha}_{ij} = b_{ij} - \frac{N}{2}$, $\alpha_{ij}^* = N - b_{ij}$, and $\tilde{\alpha}_{ij}+\alpha_{ij}^* = \frac{N}{2}$. If $b_{ij} < \frac{N}{2}$ then $\tilde{\alpha}_{ij} = \frac{N}{2} - b_{ij}$, $\alpha_{ij}^* = b_{ij}$, and $\tilde{\alpha}_{ij}+\alpha_{ij}^* = \frac{N}{2}$.

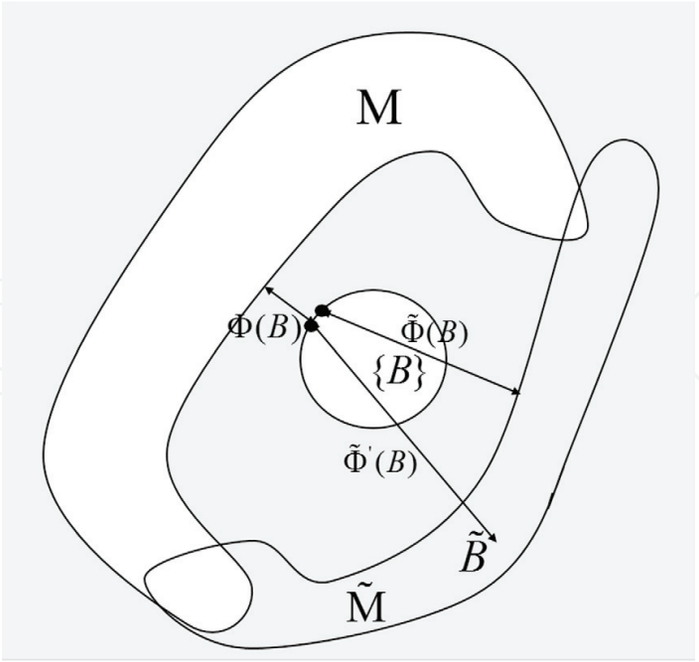

Figure 6. The sets of contrasting M, blurred $\tilde{M}$ matrices, and the set of matrices $\{B\}$.

Summing over all the set of values of pairs of indices i, j, we get that $\Phi(B)+\tilde{\Phi}'(B) = \frac{Nml}{2}$.

We consider the problem of finding optimal ensemble clusterings for the criterion (2). It is clear that $\Phi(B) = \sum_{i=1}^{m}\sum_{j=1}^{l}\min\left(b_{ij}, N - b_{ij}\right)$.

We introduce the notations $M = \{1, 2, \ldots, m\}$, $X_j = \left\{i | b_{ij} \geq \frac{N}{2}, i = 1, 2, \ldots, m\right\}$, $Y_j = M \backslash X_j$, $j = 1, 2, \ldots, l$. Let $\pi^{\nu} =< \mu_1^{\nu}, \mu_2^{\nu}, \ldots, \mu_l^{\nu} >$, $\nu = 1, 2, \ldots, N$ be some permutation of the set $\pi^0 =< 1, 2, \ldots, l >$. A set of permutations $\pi =< \pi^1, \pi^2, \ldots, \pi^N >$ uniquely determines the matrix of estimates.

$$B' = \|b'_{ij}\|_{m\times l}, b'_{ij} = b_{ij}(\pi) = \sum_{\nu=1}^{N} \alpha'^{\nu}_{ij}.$$

We will further assume that the "initial" matrix $\|\alpha_{ij}^{\nu}\|_{m\times l}$ of the algorithm A_{ν}^{c} corresponds to the permutation π^0. $\|\alpha'^{\nu}_{ij}\|_{m\times l}$ is the matrix of the algorithm A_{ν}^{c} corresponding to some permutation π^{ν}. Then $\alpha'^{\nu}_{ij} = \alpha^{\nu}_{i\mu_j^{\nu}}$.

Consider $\tilde{\Delta}_{\nu} = \sum_{j=1}^{l}\left(\sum_{i\in X_j}\overline{\alpha}^{\nu}_{ij} + \sum_{i\in Y_j}\alpha^{\nu}_{ij}\right)$, $\tilde{\Delta}'_{\nu} = \sum_{j=1}^{l}\left(\sum_{i\in X_j}\overline{\alpha}'^{\nu}_{ij} + \sum_{i\in Y_j}\alpha'^{\nu}_{ij}\right)$.

Then $\Delta_{\nu} = \tilde{\Delta}'_{\nu} - \tilde{\Delta}_{\nu} = \sum_{j=1}^{l}\left(\sum_{i\in X_j}\left(\alpha^{\nu}_{ij} - \alpha'^{\nu}_{ij}\right) + \sum_{i\in Y_j}\left(\alpha'^{\nu}_{ij} - \alpha^{\nu}_{ij}\right)\right)$. We convert this expression.

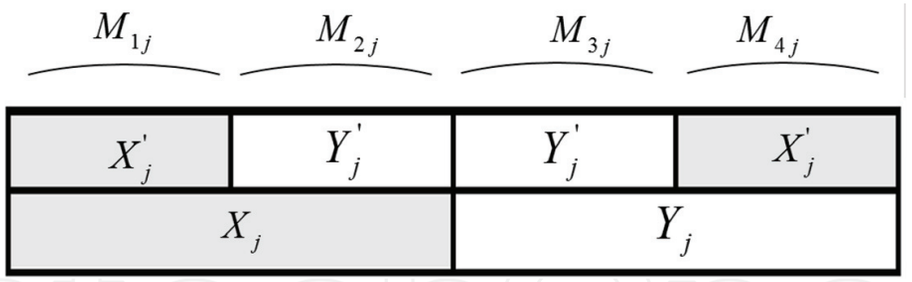

Figure 7. Sets X_j, Y_j, $j = 1, 2, \ldots, l$ are changed.

The identity $\sum_{j=1}^{l}\left(\sum_{i\in X_j}\alpha_{ij}^{\nu}+\sum_{i\in Y_j}\alpha_{ij}^{\nu}\right)=\sum_{j=1}^{l}\left(\sum_{i\in X_j}\alpha_{i\mu_j^{\nu}}^{\nu}+\sum_{i\in Y_j}\alpha_{i\mu_j^{\nu}}^{\nu}\right)$ is valid. Get $\Delta_{\nu}=\sum_{j=1}^{l}\left(\sum_{i\in X_j}\left(\alpha_{ij}^{\nu}-\alpha_{i\mu_j^{\nu}}^{\nu}\right)+\sum_{i\in Y_j}\left(\alpha_{i\mu_j^{\nu}}^{\nu}-\alpha_{ij}^{\nu}\right)\right)=2\sum_{j=1}^{l}\sum_{i\in X_j}\left(\alpha_{ij}^{\nu}-\alpha_{i\mu_j^{\nu}}^{\nu}\right)=2\sum_{j=1}^{l}\sum_{i\in X_j}\alpha_{ij}^{\nu}-2\sum_{j=1}^{l}\sum_{i\in X_j}\alpha_{i\mu_j^{\nu}}^{\nu}$.

Thus, minimizing a function is equivalent to maximizing the second sum of the expression.

After applying the permutations $\pi=<\pi^1,\pi^2,\ldots,\pi^N>$, the sets X_j, Y_j, $j = 1, 2, \ldots, l$ change. We introduce the notations $M_{1j}=X_j\backslash\left(Y_j^{'}\backslash Y_j\right)$, $M_{2j}=Y_j^{'}\backslash Y_j$, $M_{3j}=Y_j\backslash\left(X_j^{'}\backslash X_j\right)$, $M_{4j}=X_j^{'}\backslash X_j$.

Figure 7 schematically shows the changes in sets X_j, Y_j, $j = 1, 2, \ldots, l$.

Theorem 2

$$\Delta\Phi=\Phi\left(B^{'}\right)-\Phi(B)\le\sum_{\nu=1}^{N}\Delta_{\nu}+\sum_{\nu=1}^{N}\left(|M_{2j}|\begin{cases}-2, & N-even,\\ -1, & N-odd\end{cases}+|M_{4j}|\begin{cases}0, & N-even,\\ -1, & N-odd\end{cases}\right)$$

The proof is given in [12, 13]. Theorem 2 is the basis for creating an effective minimization algorithm of Φ.

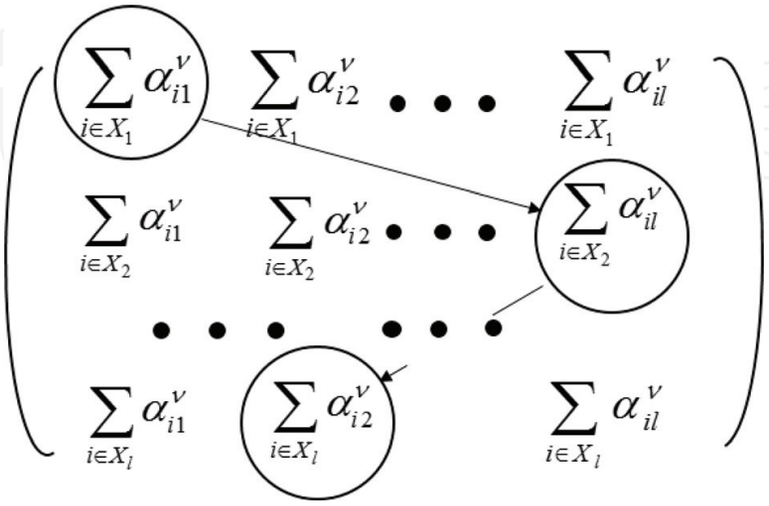

Figure 8. All possible variants of $\sum_{j=1}^{l}\sum_{i\in X_j}\alpha_{i\mu_j^{\nu}}^{\nu}$ for all admissible j and i.

Since the second sum is always not positive, we have an upper bound. We consider the problem of minimizing a function Δ_ν. We write out all possible variants of the function $\sum_{j=1}^{l}\sum_{i\in X_j}\alpha_{i\mu_j^\nu}^{\nu}$ in the form of a table in **Figure 8**. Then the minimum of this function is reduced to finding the maximum matching of the bipartite graph, for finding which we can use the polynomial Hungarian algorithm [16].

It is clear that $\min\limits_{\pi^\nu}\Delta_\nu \le 0$. Now we can propose the following heuristic algorithm for steepest descent.

Algorithm.

1. We calculate X_j, $j = 1, 2, \ldots, l$.

2. We find $\Delta_\nu^* = \min\limits_{\pi^\nu}\Delta_\nu$ for each ν.

If $\sum_{\nu=1}^{N}\Delta_\nu^* < 0$, then apply the found permutations $\pi^\nu =< \mu_1^\nu, \mu_2^\nu, \ldots, \mu_l^\nu >$, $\nu = 1, 2, \ldots, N$ and go to step 1).

If $\sum_{\nu=1}^{N}\Delta_\nu^* = 0$ then the END of algorithm.

NOTE. We note that our algorithm does not even find a local minimum of the criterion $\Phi(B)$. Nevertheless, this algorithm is very fast, its complexity at each iteration is estimated as $O(l^5 mN)$.

4. The algorithm of collective k-means

Results of clustering by N algorithms of sampling of m objects to l clusters solutions are obtained, which we can write in the form of a binary matrix $\|\alpha_{ij}^{\nu}\|$, $\nu = 1, 2, \ldots, N$, $i = 1, 2, \ldots, m$, $j = 1, 2, \ldots, l$. We assume that the cluster numbers in each algorithm are fixed. Then any horizontal layer number i of this three-dimensional matrix will denote the results of object $\mathbf{x}_i$ clustering. As an ensemble clustering of the sample X, we can take the result of clustering the "new" descriptions—the layers of the original matrix $\|\alpha_{ij}^{\nu}\|$, $\nu = 1, 2, \ldots, n$. As a method of clustering, we take the method of minimizing the dispersion criterion. Let there be a lot of N clusterings $\|\alpha_{i_1 j}^{\nu}\|, \|\alpha_{i_2 j}^{\nu}\|, \ldots, \|\alpha_{i_N j}^{\nu}\|$ with heuristic clustering algorithms, then we calculate their sample mean $\|\alpha_j^{*\nu}\|$ as the solution of the problem $\sum_{\mu=1}^{t}\left(\alpha_j^{*v} - \alpha_{i_\mu j}^{v}\right)^2 \to \min\limits_{\alpha_j^{*v}}$. Where do we obtain $\alpha_j^{*v} = \frac{1}{N}\sum_{\mu=1}^{N}\alpha_{i_\mu j}^{v}$. Note that this method makes it possible to calculate such ensemble clusterings $\mathrm{K} = \left\{K_1^*, K_2^*, \ldots, K_l^*\right\}$ that the sets of heuristic clustering of the objects of some cluster of the collective solution will be close to each other in the Euclidean metric. The committee synthesis of collective decisions provides more interpretable solutions. Indeed, if $\mathrm{K}^\nu = \left\{K_1^\nu, K_2^\nu, \ldots, K_l^\nu\right\}$, $\nu = 1, 2, \ldots, N$ are separate solutions of heuristic clustering algorithms, then the cluster of collective solution will be the "intersection" of many some original clusters $K_{i_1}^1, K_{i_2}^2, \ldots, K_{i_l}^N$.

5. Man-machine (video-logical) clustering method

In the problems of ensemble clustering synthesis considered earlier, we did not consider the number of initial clustering algorithms, their quality and their proximity. Ensemble clustering was built and reflected only the opinion of the collective decisions that we used. "Internal" indices [9] reflect the person's ideas about clustering. You can think up examples of data when known internal criteria lead to degenerate solutions.

At the same time, a person has the ability to cluster visual sets on a plane without using any proximity functions, criteria and indices. The following idea was realized. A person can personally cluster projections of sets of points from R^n into R^2. Having made such clusterings under different projections, we can construct generally speaking various N clusterings, which we submit to the input of the construction of the collective solution. The person himself "does not see" the objects in R^n, but can exactly solve the clustering tasks on the plane. Thus, here we use N precise solutions, but of various partial information about the data. Consider this video-logical method on one model example.

A sample of two normal distributions with independent characteristics was considered. The first feature of the first distribution (200 objects) had zero expectation and the standard deviation, the first attribute of the second distribution (200 objects) had these values equal to 5. All the other 49 attributes for all objects had $\mathbf{a}_i = 5$, $\boldsymbol{\sigma}_i = 5, i = 2, 3, \ldots, 50$. That is, the two sets had equal distributions for 49 features and one informative feature. Clustering of the entire sample by minimizing dispersion is shown in **Figure 9**. Black and gray points on sample visualization represent the objects of the first and second clusters. Here the fact of informative character of the first feature is lost.

The program of the video-logical approach worked as follows. With the help of a single heuristic approach, all C_n^2 projections are automatically ordered according to the descending criteria of the presence of two clusters. Next we as experts consider some projections and with the help of the mouse we select in each of them two clusters. **Figure 10** shows two such examples. Note that the first feature was present in all projections. It was used "manually" as the defining area for the dense location of objects. Then 10 "manual" clustering went to the program entrance for the committee synthesis of the collective solution. Note that only two objects were erroneously clustered.

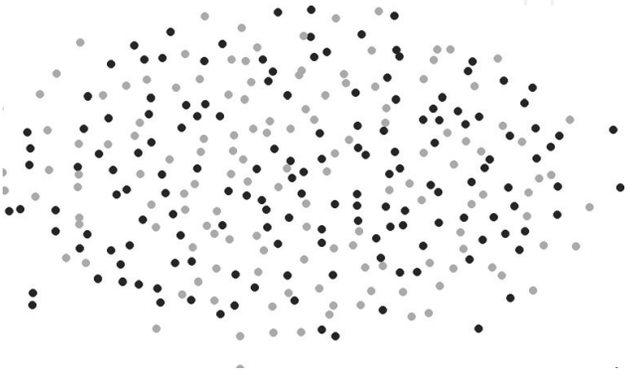

Figure 9. Clustering of a sample of model objects by the method of minimizing variance.

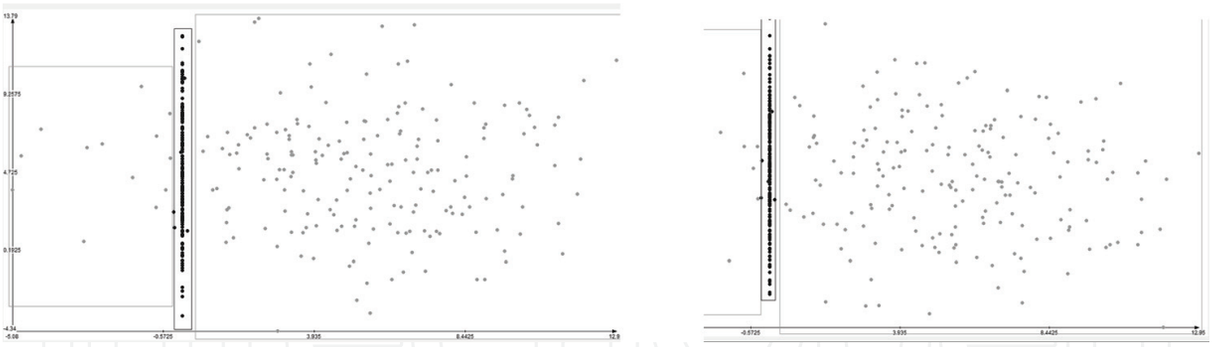

Figure 10. Allocation of clusters by mouse on the (1.4) and (1.6) features.

6. Conclusion

This chapter consists of two parts. First, clustering criteria based on sustainability are introduced. Next, we propose an approach to processing the sets of obtained partitions of the same sample. As the initial clustering, it is better to use stable clustering. It is shown how a person can be used in the construction of the committee synthesis of ensemble clustering.

Acknowledgements

The reported study was funded by RFBR according to the research project No 17-01-00634 and No 18-01-00557.

Author details

Vladimir Vasilevich Ryazanov

Address all correspondence to: rvvccas@mail.ru

Dorodnicyn Computing Centre, Federal Research Center, "Computer Science and Control" of Russian Academy of Sciences, Moscow, Russia

References

[1] Halkidi M, Batistakis Y, Vazirgiannis M. Cluster validity methods: Part 1. SIGMOD Record. 2002;**31**(2):40-45. DOI: 10.1145/601858.601862

[2] Aggarwal C, Reddy C. Data Clustering: Algorithms and Applications. CRC Press; 2014

[3] Duda R, Hart P, Stork D. Pattern Classification. 2nd ed. New York: Wiley; 2000

[4] Lloyd S. Least squares quantization in PCM (PDF). IEEE Transactions on Information Theory. 1982;**28**(2):129-137. DOI: 10.1109/TIT.1982.1056489

[5] Kriegel H, Kröger P, Sander J, Zimek A. Density-based clustering. WIREs Data Mining and Knowledge Discovery. 2011;**1**(3):231-240. DOI: 10.1002/widm.30

[6] Dempster A, Laird N, Rubin D. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B. 1977;**39**(1):1-38 JSTOR 2984875. MR 0501537

[7] Jain A, Dubes R. Algorithms for Clustering Data. Englewood Cliffs: Prentice-Hall, Inc.; 1998

[8] Kaufman L, Rousseeuw P. Finding Groups in data: An Introduction to Cluster Analysis. New York: Wiley; 2009

[9] Aggarwal C. Data Mining: The Textbook. Yorktown Heights/New York: IBM T.J. Watson Research Center; 2015. 771 p. DOI: 10.1007/978-3-319-14142-8

[10] Kuncheva L. Combining Pattern Classifiers: Methods and Algorithms. Hoboken: Wiley; 2004. DOI: 10.1002/9781118914564

[11] Desgraupes B. Clustering indices. University Paris Ouest. Lab Modal'X; 2013

[12] Ryazanov V. Commitee synthesis of algorithms for recognition and classification. Journal of Computational Mathematics and Mathematical Physics. 1981;**21**(6):1533-1543. DOI: 10.1016/0041-5553(81)90161-0

[13] Ryazanov V. On the synthesis of classification algorithms on finite sets of classification algorithms (taxonomy). Journal of Computational Mathematics and Mathematical Physics. 1982;**22**(2):429-440. DOI: 10.1016/0041-5553(82)90049-0

[14] Ryazanov V. One approach for classification (taxonomy) problem solution by sets of heuristic algorithms. In: Proceedings of the 9-th Scandinavian Conference on Image Analysis; 6–9 June 1995; Uppsala; 1995(2). pp. 997-1002

[15] Biryukov A, Shmakov A, Ryazanov V. Solving the problems of cluster analysis by collectives of algorithms. Journal of Computational Mathematics and Mathematical Physics. 2008;**48**(1):176-192. DOI: 10.1134/S0965542508010132

[16] Kuhn H. The Hungarian method for the assignment problem. Naval Research Logistics Quarterly. 1955;**2**:83-97. DOI: 10.1002/nav.3800020109

[17] Ryazanov V. Estimations of clustering quality via evaluation of its stability. In: Bayro-Corrochano E, Hancock E, editors. CIARP 2014. LNCS. Vol. 8827; 2014. pp. 432-439. DOI: 10.1007/978-3-319-12568-8_53

[18] Ryazanov V. About estimation of quality of clustering results via its stability. Intelligent Data Analysis. 2016;**20**:S5-S15. DOI: 10.3233/IDA-160842

[19] Sigillito V, Wing S, Hutton L, Baker K. Classification of radar returns from the ionosphere using neural networks. Johns Hopkins APL Technical Digest. 1989;**10**:262-266

[20] Rastrigin L, Erenstein R. Collective decision-making in pattern recognition. Avtomatica i telemehanika. 1975;**9**:133-144

[21] Method of committees in pattern recognition. Sverdlovsk, IMM AN USSR; 1974

[22] Yu Z. On the algebraic approach to solving problems of recognition or classification. Problems of Cybernetics. 1978;**33**:5-68

[23] Zuev Y. The method of increasing the reliability of classification in the presence of several classifiers, based on the principle of monotony. Journal of Computational Mathematics and Mathematical Physics. 1981;**21**(1):157-167

[24] Krasnoproshin V. About the optimal corrector of the set of recognition algorithms. Journal of Computational Mathematics and Mathematical Physics. 1979;**19**(1):204-215

[25] Zhuravlev Y. Selected Scientific Works. Moscow: Publishing House Magister; 1998. p. 420