Intelligent Data Analysis 20 (2016) S5–S15
DOI 10.3233/IDA-160842
IOS Press

About estimation of quality of clustering results via its stability

Vladimir Ryazanov
Dorodnicyn Computing Centre, Federal Research Center "Computer Science and Control" of Russian Academy of Sciences, Vavilov st. 40, 119333 Moscow, Russia
Tel.: +7 499 135 4498; Fax: +7 499 135 6159; E-mail: rvvccas@mail.ru; http://www.ccas.ru/rvv

Abstract. We know that there are many clustering methods for the case of a known/unknown number of clusters. Clustering is a result of fulfillment of some stopping criterion. Usually, optimisation of some quality criterion is performed or iterative processes are accomplished. How to estimate the quality of clustering obtained by some method? Is the obtained clustering result corresponding to the objective reality or some stopping criterion of the algorithm is made and we have obtained only some partition? Here, a practical approach and the common general criteria based on an estimation of the stability of clustering are submitted. The criterion does not use any probabilistic assumptions or distances in feature space. For some well-known clustering algorithms, efficient methods for computing the introduced stability criteria according to the training set are obtained. Some illustrative real and artificial examples for various situations are shown.

Keywords: Clustering, stability, cluster, feature, a hierarchical grouping, variance, mean

1. Introduction and modern state

Currently, there are different approaches and algorithms for data clustering. Many applications require a partition of data into "similar" groups or data clustering. Data clustering is an important part of modern data analysis. In the books [2,12] given many approaches and methods of clustering (minimization of the variance criterion, minimizing the determinant of intra-scattering, hierarchical grouping methods, k-means algorithm, grid-based and spectral clustering, etc.). In book [3] are discussed many current issues of cluster analysis: clustering algorithms and their generalizations for clusters of arbitrary shape (representative-based, hierarchical, probabilistic, density-based, graph-based, matrix factorization-based methods, etc.), problems of feature selection for clustering, the presence of outliers, the choice of the number of clusters. Here was claimed that, after a determination of clustering of the data, it is important to evaluate its quality or to make cluster validation. It was proposed a lot of various internal clustering criteria. Here are some well-known criteria. Let us fixed some metric in feature space. Than Dunn index [14] is defined as the quotient of the minimal distance between points of different clusters and the largest within-cluster distance. Let us denote by d_k the mean distance of the points of the cluster number k and barycenter of this cluster. We denote by d_{kl} the distance between the barycenters of clusters number k and l. Later, the maximal value of $M_k = \frac{d_k+d_l}{d_{kl}}$ over all other l is calculated. The Davies-Bouldin index is defined as the mean value of M_k among all the clusters [10]. Let us fix the within-cluster mean distance, the mean distance between points of various clusters and the smallest of these mean distances. The mean through all clusters of some simple their expression is called the silhouette index [22]. The indices use mean distances and dispersions, logarithms of the traces of the variance-covariance matrix of

each cluster, distances between the pairs of points inside each cluster and belonging to various clusters, centroids of the observations, scatter matrices, variance-covariance matrices, determinants of the scatter matrix, traces of the matrices, within-group and between-group scatter matrices, etc. The external criteria allow to assess the similarity of two different clusterings of the same data. It uses clustering results of all possible pairs of points of training sample. In [11] provides 27 different internal and 11 external criteria. As was said in [3], "cluster validity criteria are far from perfect", and "such measures should be used with caution".

A large number of modern studies is based on the stability of the clustering, which is understood in different ways. As a general approach for comparing and evaluating clusterings, various formalizations of the idea of stability are discussed. The problem of clustering is usually considered in statistical statement and the idea of resampling is used. In [5] the theoretical analysis of clustering stability of k-means clustering algorithm was conducted. The algorithm A is regarded as stable on probability distribution P when expectation of Hamming distance between two clusterings tends to 0 if number of samples tends to infinity. Asymptotic result was obtained. The authors denote the extension to predicting the behavior of stability over finite number of samples and case of arbitrary probability distributions are open problems. The approach [26] accepts that stable cluster contains majority of instances that are mapped correctly to this cluster. Given a stable clustering result, it shows that matching pairs from the same cluster and re-cluster should get similar result. The stability measure introduced in [19] quantifies the reproducibility of clustering solutions on a second data set. The preferred number of clusters is determined by minimizing classification risk as a function of the number of clusters. It is alleged that were obtained good results on both the model and the real data. It is assumed in paper [25] that the clustering data is actually a set from some probability distribution. The stability of a clustering algorithm A is defined as expectation of some "dissimilarity" function between two clusterings. Distance oriented stability factors (ratio to inter and intra stability factor) are introduced in [9] to compute the quality of clusters. Inter stability factor of a cluster represents the proximity of samples within that cluster, while intra stability factor establishes relationship among the data samples of different clusters. The paper [13] proposed a resampling method to estimate the number of clusters by repeatedly and randomly dividing the original data set into two non-overlapping sets. In [6] the stability is characterized by the distribution of pairwise similarities between clusterings obtained from subsamples of the data. The information stability index is proposed in [21]. Note that is used here a probabilistic model in the classification. In [8] dendrograms as ultrametric spaces are represented and the Gromov-Hausdorff distance is used. The quantification of perturbations in the input metric space is made and its affect on the result of hierarchical methods was investigated. This paper uses a metric space, and has a theoretical significance. A method is proposed in [7] for estimation clustering stability under the removal of some objects from initial clustering set. The creation of the method gives us possibility to estimate the stability of a partition, that can be considered as a weighted mean of the stability measures of all clusters in the partition that are defined as Loevinger's measures of rule quality. In [27] the results of one empirical study on the stability of two clustering algorithms are reported (k-means and normalized spectral clustering).

So, we have a number of questions connected with the validity of data clustering. The quality of clustering is measured in the individual indicators for each method. How to compare different clusterings? Which of the clusterings obtained by different methods is better and which method is the best for concrete data? Does the resulting data partition correspond to the objective existence of cluster structure in the data or just run the stop condition and we have some partition? The most important task is the choice of the best clustering algorithm for specific data and determining the number of clusters. The latter problem is a multi-extremal. Obviously, the natural number of clusters can accept multiple values. In this

Fig. 1. A simple example of a degenerate solution.

paper we propose a common nonstatistical approach to estimating the quality of clusterings that is also based on the idea of stability. Features and clustering algorithms can be various. We do not require the presence of metrics in the feature space. Let arbitrary finite data set and an algorithm for its clustering are given. We apply this clustering algorithm and obtain a solution. If clustering is a characteristic of the sample, it should not be changed (or changed much) for a small change of the data set. It is natural to consider as the smallest change in the data set removing from set one sample. If the resulting clustering does not change when deleting of different samples, it is natural to assume a high quality. Criteria for estimating the quality of the clustering, efficient algorithms to compute it, and results of numerical experiments on model and real data are proposed. The results illustrate that the proposed quality criteria can be used to estimate the quality of the solutions and, in particular, to calculate the true number of clusters. Note that the cases of big data with outliers were not considered in this paper.

The paper is organized as follows. The main approach and definitions are introduced in Section 2. The main idea of clustering results estimation is based on stability of data partitions that were previously obtained by the used clustering method. Sections 3–5 concretize algorithms for quality estimation of clusterings obtained by minimization of variance criterion, "k-means", and hierarchical grouping methods. Here we use the well-known algorithms and obtain for them a simple numerical methods of calculating the basic criterion in Section 2. Later, the experimental results including clusterings of various artificial and real data are given in Section 6. Section 7 contains the conclusions and future work. This article is an extension of the report [23]. The idea of this approach was proposed in [4].

2. Calculation of the quality of clustering based on evaluating of stability of the partition

At first, consider as an illustrative example one model unsupervised classification task. Shown in Fig. 1, the model problem similar to the problem shown in Fig. 10.8, page 629 [12]. Variance of the first feature of the second set of samples is 15 times greater than the corresponding variance for the first set of samples. Visual solution of samples partition into two clusters is obvious. However, finding the minimum-variance partition leads to degenerate clusterings. Proposed in the article approach allows to characterize the obtained solution as an unstable and therefore bad.

Suppose there is the method of clustering, given a set of the feature descriptions of samples $X = \{\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_m\}, \bar{x}_i = (x_{i1}, x_{i2}, \ldots, x_{in}), \bar{x}_i \in R^n$, and obtained a solution to the problem of clustering the data set at l clusters $\mathrm{K} = \{K_1, K_2, \ldots, K_l\}$, $K_i \subseteq X, i = 1, 2, \ldots, l, \bigcup_{i=1}^{l} K_i = X, K_i \bigcap K_j = \emptyset, i \neq j$.

We call the partition $\mathrm{K}^*(\bar{x}_t) = \{K_1^*, K_2^*, \ldots, K_l^*\}$ of the sample $X_t = \{\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_m\} \backslash \{\bar{x}_t\}, t = 1, 2, \ldots, m$, identical to the clustering K of the sample X, if $\mathrm{K}^*(\bar{x}_t)$ is the clustering and $K_j^* \subseteq K_j, j = 1, 2, \ldots, l$, is satisfied. For simplicity, we assume that the numbering of clusters coincides. Identical clusterings are denoted as $\mathrm{K}^*(\bar{x}_t) \approx \mathrm{K}$. This means that we have the situation $K_j^* = K_j, j = 1, 2, \ldots, l, j \neq i, K_i^* = K_i \backslash \{\bar{x}_t\}$, if $\bar{x}_t \in K_i$. Further we suppose that the partitions K are non-degenerate, i.e. $|K_i| > 1, i = 1, 2, \ldots, l$.

Definition 1. The function of the clustering quality is the quantity $F(\mathrm{K}) = |\{\bar{x}_t, t = 1, 2, \ldots, m$: $\mathrm{K}^*(\bar{x}_t) \approx \mathrm{K}\}|/m$.

Calculation of $F(\mathrm{K})$ is similar to performing a leave-one-out procedures in evaluating the quality of supervised classification algorithms, but it does not usually require the solution of m clustering tasks. To estimate the quality of clustering by $F(\mathrm{K})$ it is enough to be able to effectively determine the fact of clusterings identity $\mathrm{K}^*(\bar{x}_t) \approx \mathrm{K}$ for each $t = 1, 2, \ldots, m$. Let $\bar{x}_t \in K_i$. It is sufficient to determine whether or not the partition $\mathrm{K} = \{K_1, K_2, \ldots, K_{i-1}, K_i \backslash \{\bar{x}_t\}, K_{i+1}, \ldots, K_l\}$ is the result of clustering for each t. This requires to check whether a corresponding condition is "stopping" of the algorithm. Note that we consider the stability of a fixed partition with respect to the particular clustering method. For some well-known methods there may be "economical" methods for calculating the quantities $F(\mathrm{K})$.

Along with the criterion of $F(\mathrm{K})$ other quality criteria are of interest. Let a clustering of the set X_t was obtained under using of initial partition $\mathrm{K}^*(\bar{x}_t)$. If the condition $\mathrm{K}^*(\bar{x}_t) \approx \mathrm{K}$ is not satisfied, it will differ from partition $\mathrm{K}^*(\bar{x}_t)$. For simplicity, denote the obtained clustering also as initial partition $\mathrm{K}^*(\bar{x}_t)$.

Definition 2. $F^*(\mathrm{K}) = \sum_{t=1}^{m} d(\mathrm{K}^*\ (\bar{x}_t), \mathrm{K})/m$, where $d(\mathrm{K}^*(\bar{x}_t), \mathrm{K})$ is a similarity between the clusterings.

As a function of similarity between the clusterings, $d(\mathrm{K}^*(\bar{x}_t), \mathrm{K}) = \max_\alpha \sum_{i=1}^{l} |K_i^* \cap K_{\alpha_i}|/(m-1)$ can be taken, where $\alpha = \langle \alpha_1, \alpha_2, \ldots, \alpha_l \rangle$ is a permutation of $\langle 1, 2, \ldots, l \rangle$. This criterion reflects the "average" closeness of $\mathrm{K}^*(\bar{x}_t)$ and K. Calculation of $d(\mathrm{K}^*(\bar{x}_t), \mathrm{K})$ is reduced to find maximal matching in a bipartite graph for which there exists a polynomial algorithm [18].

Definition 3. $F^{\max}(\mathrm{K}) = \min_t d(\mathrm{K}^*(\bar{x}_t), \mathrm{K})$ where there is a similarity $d(\mathrm{K}^*(\bar{x}_t), \mathrm{K})$ between clusterings.

The criterion indicates the presence of point $\bar{x}_t$ significantly distorting the clustering K. Its removal leads to a substantial change in the clustering.

3. The stability of based on the minimization of variance criterion clustering

It is known [12] that for minimization of variance criterion $f(\mathrm{K}) = \sum_{j=1}^{l} \sum_{\bar{x}_i \in K_j} \rho^2(\bar{x}_i, \bar{m}_j))$ the condition of local optimality of the partition K is fulfillment of the inequalities

$$\frac{n_i}{(n_i - 1)} \|\widehat{x} - \bar{m}_i\|^2 - \frac{n_j}{(n_j + 1)} \|\widehat{x} - \bar{m}_j\|^2 \leqslant 0 \tag{1}$$

for K_i, K_j, and $\widehat{x} \in K_i$ (here $\rho(\bar{x}, \bar{y}) = (\|\bar{x} - \bar{y}\|, n_i = |K_i|, \bar{m}_i = \frac{1}{n_i} \sum_{\bar{x}_\alpha \in K_i} \bar{x}_\alpha)$.

We obtain for any $t = 1, 2, \ldots, m$ similar to Eq. (1) the conditions for partition $\mathrm{K}^*(\bar{x}_t)$ of the set X_t to be the result of clustering. Suppose that $\bar{x}_t \in K_i$ is a removed sample. Moving $\widehat{x}$ from K_ν^* to $K_\mu^*, i \neq \nu, i \neq \mu$ does not reduce the optimality criterion $f(\mathrm{K})$ while K is optimal. Therefore, it is sufficient to consider the cases of $\widehat{x}$ transfer from K_i^* to K_j^*, and vice versa.

A. Transfer of $\widehat{x}$ from K_i^* to K_j^*, $j = 1, 2, \ldots, l, j \neq i$.
In accordance with Eq. (1) it must be satisfied

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

where

$$n_i^* = |K_i^*|, \bar{m}_i^* = \frac{1}{n_i^*} \sum_{\bar{x}_\alpha \in K_i^*} \bar{x}_\alpha = \frac{n_i}{(n_i - 1)} \bar{m}_i - \frac{1}{(n_i - 1)} \bar{x}_t.$$

Then, Eq. (2) takes the form

$$\frac{(n_i-1)}{(n_i-2)}\left\|\widehat{x}-\bar{m}_i\right\|^2+\frac{2}{(n_i-2)}(\widehat{x}-\bar{m}_i,\bar{x}_t-\bar{m}_i)+$$
$$\frac{1}{(n_i-1)(n_i-2)}\left\|\bar{x}_t-\bar{m}_i\right\|^2-\frac{n_j}{n_j+1}\left\|\widehat{x}-\bar{m}_j\right\|^2\leqslant 0, \tag{3}$$

B. Transfer of $\widehat{x}$ from K_j^* to $K_i^*, j=1,2,\ldots,l, j\neq i$.
If $\mathrm{K}^*(\bar{x}_t)$ is the clustering result then

$$\frac{n_j}{(n_j-1)}\left\|\widehat{x}-\bar{m}_j\right\|^2-\frac{n_i^*}{(n_i^*+1)}\left\|\widehat{x}-\bar{m}_i^*\right\|^2\leqslant 0. \tag{4}$$

From Eq. (4) be

$$\frac{n_j}{(n_j-1)}\left\|\widehat{x}-\bar{m}_j\right\|^2-\frac{(n_i-1)}{n_i}\left\|\widehat{x}-\bar{m}_i\right\|^2-$$
$$\frac{2}{n_i}(\widehat{x}-\bar{m}_i,\bar{x}_t-\bar{m}_i)-\frac{1}{(n_i)(n_i-1)}\left\|\bar{x}_t-\bar{m}_i\right\|^2\leqslant 0 \tag{5}$$

After verifying $\mathrm{K}^*(\bar{x}_t)\approx\mathrm{K}$ for all $t=1,2,\ldots,m$ (i.e. the execution of the systems Eqs (3) and (5)), one gets the value of the criterion $F(\mathrm{K})$. Note, that the complexity of the calculation of the functional $F(\mathrm{K})$ is a polynomial.

4. Stability clusterizations obtained by k-means method

Assume that clustering K has been obtained by k-means method. It denotes that

$$\left\|\widehat{x}-\bar{m}_i\right\|\leqslant\left\|\widehat{x}-\bar{m}_j\right\|, \forall j\neq i, \tag{6}$$

for $\forall\widehat{x}\in K_i$. In the case of equalities, the sample is considered to belong to the cluster with the minimum number. Let $\bar{x}_t\in K_i$ be a deleted sample. We obtain the conditions of the type Eq. (6) to partition $\mathrm{K}^*(\bar{x}_t)$.

A. Case $\widehat{x}\in K_i^*$. $\left\|\widehat{x}-\bar{m}_i^*\right\|^2\leqslant\left\|\widehat{x}-\bar{m}_j\right\|^2$, or

$$\left\|\widehat{x}-\bar{m}_i\right\|^2+\frac{2}{(n_i-1)}(\widehat{x}-\bar{m}_i,\bar{x}_t-\bar{m}_i)+\frac{1}{(n_i-1)^2}\left\|\bar{x}_t-\bar{m}_i\right\|^2\leqslant\left\|\widehat{x}-\bar{m}_j\right\|^2, \tag{7}$$

and the case of equality can occur only when $i<j$.

B. Case $\widehat{x}\in K_j$. It must be satisfied $\left\|\widehat{x}-\bar{m}_j\right\|^2\leqslant\left\|\widehat{x}-\bar{m}_i^*\right\|^2, \forall j\neq i$,

$$\left\|\widehat{x}-\bar{m}_j\right\|^2\leqslant\left\|\widehat{x}-\bar{m}_i\right\|^2+\frac{2}{(n_i-1)}(\widehat{x}-\bar{m}_i,\bar{x}_t-\bar{m}_i)+\frac{1}{(n_i-1)^2}\left\|\bar{x}_t-\bar{m}_i\right\|^2. \tag{8}$$

Equality is permissible only in cases $j<i$. By removing one by one $\bar{x}_t, t=1,2,\ldots,m$, and checking the conditions Eqs (7) and (8), we find $F(\mathrm{K})$.

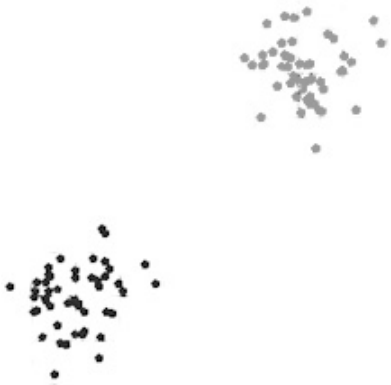

Fig. 2. Well separable clusters.

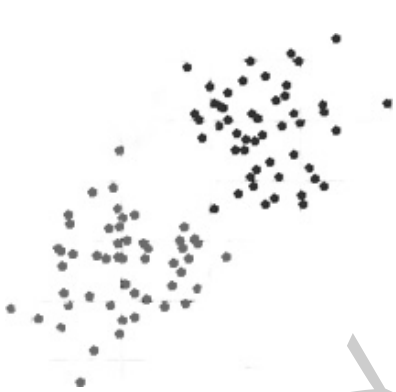

Fig. 3. Separable clusters.

Fig. 4. Satisfactory separable clusters.

5. The stability of clusterizations obtained by the methods of hierarchical grouping

To calculate the criterion $F(\mathrm{K})$, m clustering tasks are solved for sets X_t, and the identities of the obtained clusterings and the initial clustering are checked. Here, we can "economize" when checking the identity of K and some $\mathrm{K}^*(\bar{x}_t)$ because the fact of their non-identity may be detected before the final calculation of $\mathrm{K}^*(\bar{x}_t)$. We restrict ourselves to the case of the agglomerative groupings [12].

Let $\mathrm{K}^i(\bar{x}_t) = \{K_1^i, K_2^i, \ldots, K_{m-i}^i\}$ be a clustering of set X_t for $m-i$ clusters, $i \leqslant m-l, \mathrm{K}$ be a clustering of set X. The main property of the hierarchical grouping algorithms consists of the fact that $\forall k = 1, 2, \ldots, m-i, \exists j = 1, 2, \ldots, m-i-1$, will be $K_k^i \subseteq K_j^{i+1}$. Thus, if for some step $i, i \leqslant m-l$, for some k a condition $K_k^i \subseteq K_j$ is not satisfied at all $j = 1, 2, \ldots, l$, then there will be a non-identity between $\mathrm{K}^*(\bar{x}_t)$ and K. To calculate the criteria $F^*(\mathrm{K})$ and $F^{\max}(\mathrm{K})$ the hierarchical groupings $\mathrm{K}^*(\bar{x}_t)$ of the sets X_t, $t = 1, 2, \ldots, m$, are calculated.

6. Experimental results on simulated and real data

This section provides illustrative examples of calculation of criteria $F(\mathrm{K})$, $F^*(\mathrm{K})$, and $F^{\max}(\mathrm{K})$. Figures 2–5 show the visualization (see [12]) of four sets for which clusterings were carried out using dispersion method with the calculation values of all proposed criteria. Experiments on model problems with clusters of different "separability" level were performed. By choosing different expectations, the situations have been formed where the clusters are "well separable", "'separable", "satisfactory separable", and "indivisible". In each experiment as a training sample was a mixture of two-dimensional samples of a normal distribution with mean values (0, 0) and (5, 5), respectively. In four experiments the sample had standard deviations σ for each feature that were equal to 1, 2, 3 and 4. Clustering was performed at $l = 2$. To denote the objects of different clusters we used various figures, black and gray colors. At $\sigma = 1, 2$ and even at $\sigma = 3$ values of all criteria were equal to 1. Clustering provided $\sigma = 4$ proved to be unstable. Criteria values given in Table 1.

Stability criteria can be used to determine the true number of clusters. Figures 6–9 show the results of clustering the sample of 150 two-dimensional objects from a mixture of 3 normal distributions under the hypothesis $l = 2, 3, 4, 5$. It is clear that there are two possible solutions of expert. There are three cluster or two clusters, in which one cluster consists of two subclusters. At hypotheses $l = 2$ and $l = 3$ we have, respectively, partitions into two clusters and three clusters with unit values of the stability criteria.

Table 1
Values of criteria for $\sigma = 4$

$F(\mathrm{K})$	$F^*(\mathrm{K})$	$F^{\max}(\mathrm{K})$
0.88	0.99	0.96

Table 2
The stability criteria clustering for different numbers of clusters

l	$F(\mathrm{K})$	$F^*(\mathrm{K})$	$F^{\max}(\mathrm{K})$
2	1	1	1
3	1	1	1
4	0.98	0.99	0.99
5	0.98	0.99	0.97

Table 3
The results of the comparison of methods at a model problems

N	*MVC*	k-means	$d_{\min}$	$d_{\max}$	d_{avg}	d_{mean}	d_l
2	1	1	1	1	1	1	1
3	1	1	1	0.99	1	1	1
4	1	1	1	0.88	0.96	0.99	0.99
5	0.98	0.98	0.99	0.88	0.38	0.92	0.74
6	1	1	0.98	0.99	1	0.99	1
7	1	1	0.98	0.90	1	0.99	1
8	0.99	0.99	0.99	0.86	0.96	0.89	0.83
9	0.96	0.96	0.98	0.80	0.96	0.84	0.70

Fig. 5. Indivisible samples of objects.

Fig. 6. Partition at $l = 2$.

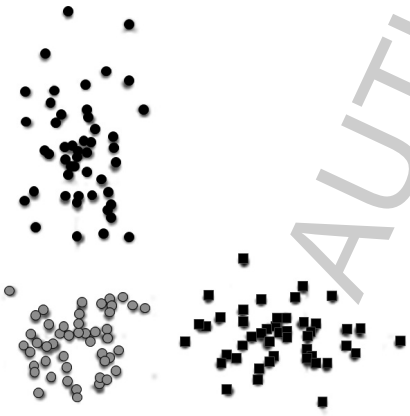

Fig. 7. Partition at $l = 3$.

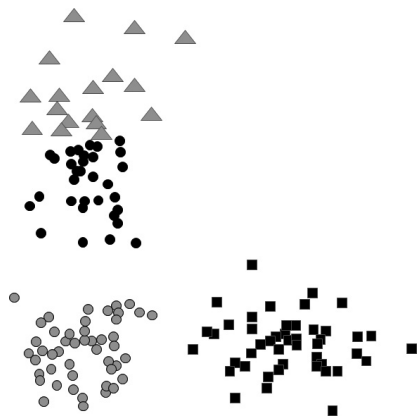

Fig. 8. Partition at $l = 4$.

The results are shown in Table 2. Thus, it may be offered the following scheme of determining the true number of clusters.

a) Clustering is performed under the hypothesis $l = 2, 3, \ldots, N$ where N is the maximum possible value of the number of clusters for the considered sample.

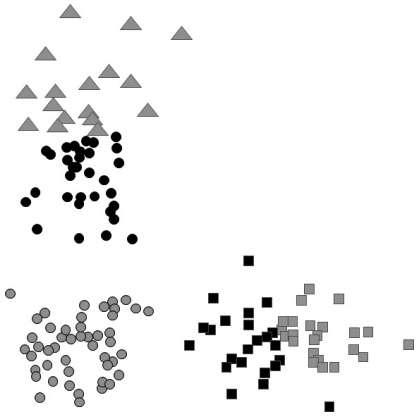

Fig. 9. Partition at $l = 5$.

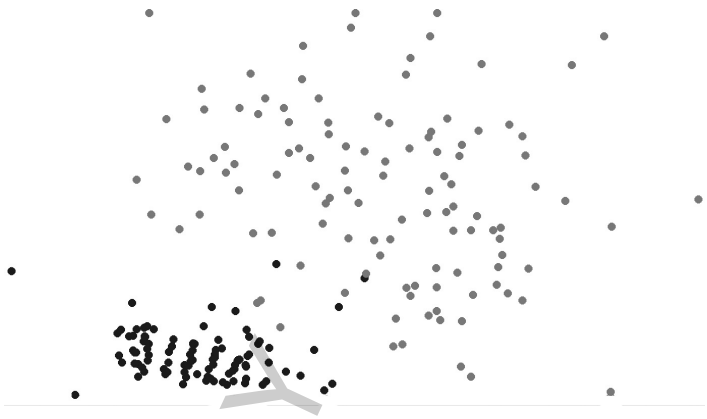

Fig. 10. Visualization of breast canser data.

b) As the admissible values of numbers of clusters we take the values l for which $F(\mathrm{K}) = 1$.

If $F(\mathrm{K}) < 1$ for some l then this number of clusters is not natural. Similar results were obtained on other model sets.

Table 3 shows the results of the calculation of criterion $F(\mathrm{K})$ for model problems presented in Figs 2–9 using various methods: minimization of variance criterion (MVC), k-means, methods of hierarchical grouping. Columns $d_{\min}(K_i, K_j)$, $d_{\max}(K_i, K_j)$, $d_{avg}(K_i, K_j)$, $d_{mean}(K_i, K_j)$, $d_l(K_i, K_j)$ correspond to agglomerative grouping with different criteria of hierarchical combining groups. After selecting one of the criteria ($d_{\min}(K_i, K_j) = \min_{\bar{x}\in K_i, \bar{y}\in K_j} \|\bar{x} - \bar{y}\|$, $d_{\max}(K_i, K_j) = \max_{\bar{x}\in K_i, \bar{y}\in K_j} \|\bar{x} - \bar{y}\|$, $d_{avg}(K_i, K_j) = \frac{1}{n_i n_j} \sum_{\bar{x}\in K_i} \sum_{\bar{y}\in K_j} \|\bar{x} - \bar{y}\|$, $d_{mean}(K_i, K_j) = \|\bar{m}_i - \bar{m}_j\|$, or $d_l(K_i, K_j) = \sqrt{\frac{n_i n_j}{n_i + n_j}} \|\bar{m}_i - \bar{m}_j\|$), the groups of the previous step are combined together in one group at each step of grouping, for which the criterion value is minimal.

From experiments it is clear that in cases where the clusters are "compact" sets all criteria give the same stable results (Fig. 2 and, in fact, Fig. 3). The results presented in Figs 6 and 7 confirm widespread opinion that the criteria $d_{avg}(K_i, K_j)$ and $d_l(K_i, K_j)$ give more accurate solutions. The results of Table 3 confirm also known conclusion about the approximate meaning of methods of hierarchical grouping compared, for example, with the dispersion criterion.

We give examples of practical tasks of supervised classification with and without cluster structure. Here were considered clusterings using minimisation of variance criterion. Illustrative comparison of methods of supervised classification, unsupervised classification and estimation of the stability of the clusters was carried out on 4 practical problems. The problem of supervised classification of breast cancer [20] was defined by sample of 346 objects that have a tumor (task "Breast"). Thus, 224 objects had benign tumors. Each object is represented in the form of discrete values of 9 features. Figure 10 is a data visualization. We see here the result of non-parametric mapping of vectors from R^9 to vectors in R^2 at which are saved maximally distances between the objects ("distant" objects are mapped as "distant" objects, and "similar" objects are mapped as "similar" objects). Figure 10 shows that a first class of patients (having a benign tumor) forms a compact set (black dots). Descriptions of the second class are more diverse. The next task ("Wine") of supervised classification of wine [1] grades in a province of Italy on various features (alcohol content, malic acid, magnesium, hue, etc.) was presented by the following parameters: $n = 13, n_1 = 59, n_2 = 71, n_3 = 48, l = 3$.

The third task ("Melanoma") was supervised classification of melanoma disease by complex of radiological and geometrical features. Data on the problem were obtained from [16], the task had op-

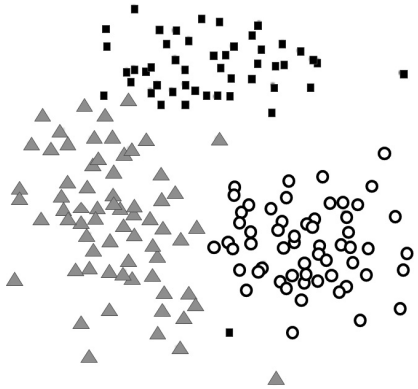

Fig. 11. Visualisation of wine data.

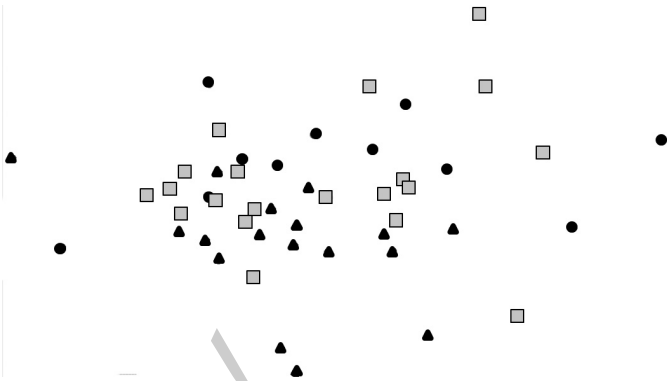

Fig. 12. Visualisation of melanoma disease data.

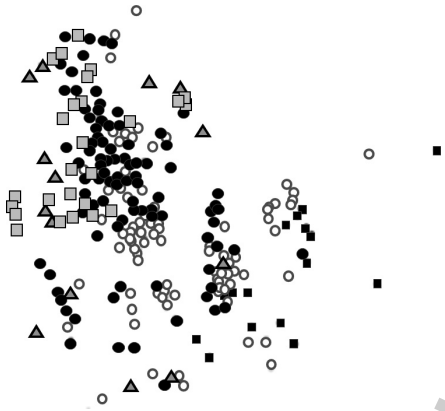

Fig. 13. Visualisation of home data.

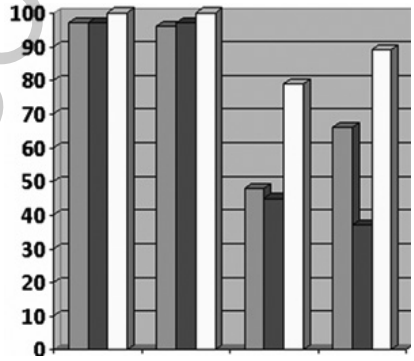

Fig. 14. Comparison of supervised classification. unsupervised classification, and estimtion of the stability.

tions $n = 33, n_1 = 17, n_2 = 20, n_3 = 11, l = 3$. The last supervised task ("Home") was to determine the cost of housing using complex of features [17]. This task is being in fact the task of finding a regression was reduced approximately to the supervised classification problem with parameters $n = 13, n_1 = 16, n_2 = 91, n_3 = 93, n_4 = 27, n_5 = 15, l = 5$. Classes were set by intervals of the cost of housing. Visualizations of tasks "Wine", "Melanoma", and "Home" are given, respectively, in Figs 11–13.

As supervised classification algorithm was used method Logical Regularities [24] of software system RECOGNITION [28]. The quality of this supervised classification method was estimated in a leave-one-out mode (percentage of correct answers). Quality of obtained by dispersion method unsupervised classification was estimated by the formula $d(\mathrm{K}, \mathrm{H}) = 100 \max_{\bar{\alpha}} \sum_{i=1}^{l} |K_i \cap H_{\alpha_i}| / m$, where $\mathrm{K} = \{K_1, K_2, \ldots, K_l\}$ is a unsupervised classification result, $\bar{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_l)$ is a permutation of $(1, 2, \ldots, l)$, and $\mathrm{H} = \{H_1, H_2, \ldots, H_l\}$ is a priori given classification result. Here K_i, H_i are the sets of objects of corresponding i-th cluster (class). Clustering was performed by dispersion method, and its quality for the specified number of clusters in the problem being assessed by criterion $F(\mathrm{K})$. The results are shown in Fig. 14. It is seen that for the "good" classification problems (the first two tasks) values of stability cluster solutions are equal to 1, and evaluations of the quality the solutions of classification

Table 4
Values stability criteria clustering for iris task

l	2	3	4	5
$F(\mathrm{K})$	1	0.98	0.98	0.98
$F^{*}(\mathrm{K})$	1	0.99	0.99	0.98
$F^{\max}(\mathrm{K})$	1	0.91	0.62	0.70

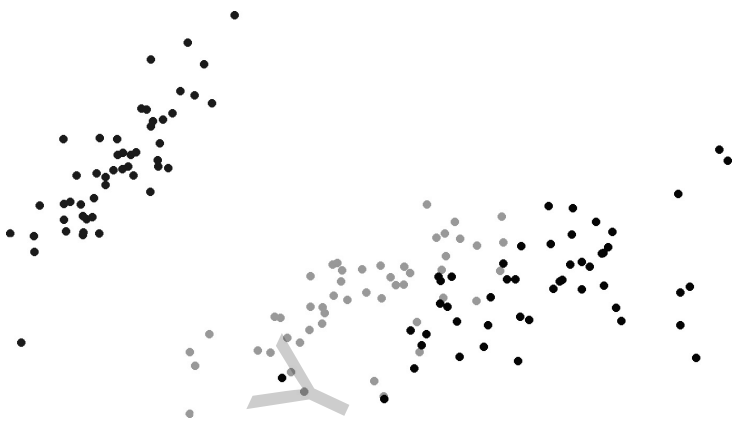

Fig. 15. Visualisation of iris data.

problems were high (above 95%). For "bad" classification problems stability of the solutions and the classification results were low.

Finally remark, visualization of well-known iris data [15] ($n = 4, m = 150, l = 3$) shown on Fig. 15, and their values of quality criteria are presented in Table 4. The iris set has a cluster structure for $l = 2$. It was obtained for degenerate clustering (see beginning of paper) $F(\mathrm{K}) = 0.99$ under $l = 2$.

7. Conclusion

The theoretical part of the paper is that it presents a general approach for the evaluation of clustering. We do not set the task of selecting a clustering model or search of universal clustering method. Just to give examples when "good" method leads to bad solutions. We do not know, good or bad clustering method for specific data. We just use this method and then look. Good or bad solution we got. The proposed criteria make it possible to say, fits (may be) a particular clustering method for specific data or not. It were proposed the methods for efficient calculation of main criterion $F(\mathrm{K})$ for three well-known clustering algorithms. In mathematics, stability theory addresses the stability of solutions of something under small perturbations of initial conditions. This approach may be considered as leave-one-out approach in the supervised classification. The theoretical analysis is usually associated with artificial probabilistic models. The behavior of criteria $F(\mathrm{K})$, $F^{*}(\mathrm{K})$, and $F^{\max}(\mathrm{K})$ for different methods of clustering, the length of the sample, the number and structure of the clusters requires large experimental calculations. The interesting study of the stability of clusters is in removing pairs, and more samples from clusters. In any case, any new criterion provides one additional informative estimate of the solution, which is useful for the practical user.

Acknowledgments

This work was supported by RAS Presidium programs number 8, Program number 2 of Department of Mathematical Sciences of RAS, RFBR 15-01-05776, 14-01-90413 Ukr, 14-01-00824, 13-01-12033.

References

[1] S. Aeberhard, D. Coomans and O. de Vel, Comparison of classifiers in high dimensional settings, Tech. Rep. no. 92-02, Dept. of Computer Science and Dept. of Mathematics and Statistics, James Cook University of North Queensland, 1992.

[2] C.C. Aggarwal and C.K. Reddy, Data clustering: Algorithms and applications, CRC Press, 2013.

[3] C.C. Aggarwal, Data Mining: The Textbook, IBM T.J. Watson Research Center Yorktown Heights, New York, 2015, 701.

[4] A.S. Arseev, K.L. Kotochigov and V.V. Ryazanov, Universal criteria for clustering and stability problems, in: *13th All-Russian Conference "Mathematical Methods for Pattern Recognition"*, S.-Peterburg, (2007), 63–64 (in Russian).

[5] S. Ben-David, D. Pal and H.U. Simon, Stability of k-means clustering, in: *Proceedings of the 20th Annual Conference on Learning Theory (COLT'07)*, Springer-Verlag Berlin, Heidelberg (2007), 20–34.

[6] A. Ben-Hur, A. Elisseeff and I. Guyon, A stability based method for discovering structure in clustered data, in: *Pacific Symposium on Biocomputing* **7** (2002), 6–17.

[7] P. Bertrand and G.B. Mufti, Loevinger's measures of rule quality for assessing cluster stability, *Computational Statistics & Data Analysis* **50**(4) (2006), 992–1015.

[8] G. Carlsson and F. Memoli, Characterization, stability and convergence of hierarchical clustering methods, *The Journal of Machine Learning Research* **11**(31) (2010), 1425–1470.

[9] A.K. Das and J. Sil, Cluster validation method for stable cluster formation, *Canadian Journal on Artificial Intelligence, Machine Learning and Pattern Recognition* **1**(3) (2010), 26–41.

[10] D.L. Davies and D.W. Bouldin, A cluster separation measure, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **PAMI-1**(2) (1979), 224–227.

[11] B. Desgraupes, Clustering Indices, University Paris Ouest Lab Modal'X, http://www2.uaem.mx/r-mirror/web/packages /clusterCrit/vignettes/clusterCrit.pdf.

[12] R.O. Duda, P.E. Hart and D.G. Stork, Pattern classification, John Wiley and Sons, 2nd edition, 2000.

[13] S. Dudoit and J. Fridlyand, A prediction-based re-sampling method for estimating the number of clusters in a data set, *Genome Biology* **3**(7) (2002), research 00361–0036.21.

[14] J. Dunn, Well separated clusters and optimal fuzzy partitions, *Journal of Cybernetics* **4** (1974), 95–104.

[15] R.A. Fisher, Iris Data Set, UCI Machine Learning Repository, http://archive.ics.uci.edu/ml.

[16] H. Ganster, M. Gelautz, A. Pinz, M. Binder, H. Pehamberger, M. Bammer and J. Krocza, Initial results of automated melanoma recognition, *Proceedings of the 9th SCIA*, Uppsala, Sweden **1** (1995), 209–218.

[17] D. Harrison and D.L. Rubinfeld, Hedonic prices and the demand for clean air, *J Environ Economics & Management* **5** (1978), 81–102.

[18] H.W. Kuhn, The hungarian method for the assignment problem, *Naval Research Logistics Quarterly* **2** (1955), 83–97.

[19] T. Lange, V. Roth, M.L. Braun and J.M. Buhmann, Stability-based validation of clustering solutions, *Neural Computation* **16**(6) (2004), 1299–1323.

[20] O.L. Mangasarian and W.H. Wolberg, Cancer diagnosis via linear programming, *SIAM News* **23**(5) (1990), 1–18.

[21] D. Pascual, F. Pla and J.S. Sanchez, Cluster validation using information stability measures, *Pattern Recognition Letters* **31** (2010), 454–461.

[22] P.J. Rousseeuw, Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, *Journal of Computational and Applied Mathematics* **20** (1987), 53–65.

[23] V. Ryazanov, Estimations of clustering quality via evaluation of its stability, *CIARP 2014*, E. Bayro-Corrochano and E. Hancock, eds, LNCS 8827, 2014, pp. 432–439.

[24] V.V. Ryazanov, Logical regularities in pattern recognition problems (parametric approach), *Computational Mathematics and Mathematical Physics* **47**(10) (October 2007), 1720–1735.

[25] O. Shamir and N. Tishby, Cluster stability for finite samples, *Advances in Neural Information Processing Systems* **21** (2007), 1–8.

[26] Z. Volkovich, R. Avros and M. Golani, A comparative approach to cluster validation, *Journal of Pattern Recognition Research* **2** (2011), 230–243.

[27] R. Zafarani, M. Makki and A.A. Ghorbani, An empirical analysis on the stability of clustering algorithms, *IEEE Computer Society Conference: 20th IEEE International Conference on Tools with Artificial Intelligence*, Ubicacion: Dayton, OH, (2008).

[28] Y.I. Zhuravlev, V.V. Ryazanov and O.V. Senco, Recognition mathematical methods, The software system, *Practical Applications*, Phasis, Moscow, (2006), 168.