Ensembles of criteria for determining the number of homogeneous groups in a combined batch of industrial production

I P Rozhnov1,2, L A Kazakovtsev1,2, M V Karaseva1,2, A A Stupina1,3 and E V Lapunova1

1 Reshetnev Siberian State University of Science and Technology, prosp. Krasnoyarskiy Rabochiy 31, Krasnoyarsk 660031, Russia
2 Siberian Federal University, prosp. Svobodny 79, Krasnoyarsk 660041, Russia
3 Krasnoyarsk State Agrarian University, 90 Mira Av., Krasnoyarsk 660049, Russia

E-mail: ris2005@mail.ru

Abstract. An article proposes a procedure for ensembles of criteria formation for determine automatically a number of clusters for dividing a combined batch of industrial production. It is shown that these ensembles of criteria can improve the accuracy of the separation of combined batches of industrial production into homogeneous ones.

1. Introduction

A current rapid development of technologies for automatic data collection, transmission and information storage, Data Mining, as well as technological growth in some industries and economies have led to the gigantic arrays of multidimensional data. The continuously increasing application of large-scale data arrays stimulates the increased interest in the development and application of methods and tools for processing these arrays and their analysis. One of the most promising areas is cluster analysis, that helps to organize (group) objects into homogeneous groups (clusters), and to solve the problem of automatic grouping (clustering) is reduced to developing an algorithm that can detect these groups without pre-tagged data [1, 2].

One of the most important components of the problem of improving the reliability of the system as a whole is to equip critical system components with a component base with the increased quality requirements (for example, electronic equipment). It is important that the same type of system elements have very close characteristics (to be homogeneous) to ensure the coordinated operation. The uniformity of characteristics of identical elements of the system is achieved if these elements were made from one batch of raw materials in one production batch. Therefore, it is necessary to use the corresponding components, which are subject to increased quality requirements in assembling critical system components with the increased quality and reliability requirements.

One must be sure that he is dealing with a batch of products made from a homogeneous (single) batch of raw materials in order to disseminate the test results of a production sample to the whole production batch. Therefore, the identification of homogeneous production batches from combined batches is one of the most important stages in testing. A series of tests are carried out, from tens to several thousand for each product, the results are tabulated and serve as data for analysis. The procedure for the parameters
separation should be regulated. The recalculation of data should give the same results or very close ones [2-4]. However, the most precise algorithms for the considered problems are randomized [2-5]. Taking into account the information given above, the problem is formed as the division into homogeneous production batches on the basis of test data (i.e., a clustering problem).

2. Estimation of the clusters number

Modern methods of cluster analysis to identify heterogeneous in the aggregate parameters of groups (clusters) offer a wide selection of tools. At the same time, a method of identifying product groups with different parameters (for example, in practice, military or space applications) should give reproducible results for a limited time for calculation. Thus, improving the accuracy of automatic grouping methods to identify groups of industrial production with different parameters is an important task. This article considers one of the methods for solving the criteria for determining a number of clusters in clustering algorithms applying an ensemble of criteria [3, 4].

One of the problems in data clustering is the automatic determination of a number of clusters (groups). In most cases, a task of a number of clusters determinations comes down to the problem of model selection. As a rule, automatic grouping algorithms are run within a certain acceptable limit of a number of possible groups, and the best value (a number of clusters) is selected based on the compactness criterion.

In cluster analysis, there are the following main criteria for determining a number of clusters: Kalinsky-Kharabash index [6], Davis-Bouldin index (DBI) [7], Krzhanowski-Lai index [8], Hartigan criterion [9], Bayes information criterion (BIC - Bayesian Information Criterion) [10], GAP criterion [11], Akaike information criterion (AIC) [12], silhouette criterion [13]. Consider some of them in detail.

Calinski-Harabasz criterion [6].

It is one of the first methods proposed for determining a number of clusters. It is effective for small amounts of data. A number of clusters is defined as the value of the argument maximizing the function $CH(k)$:

$$CH(k) = \frac{B(k)(k-1)}{W(k)(n-k)},$$

where $B(k)$ and $W(k)$ are external and internal (respectively) sums of squares of data elements with $k$ clusters.

Krzhanowski and Lai criterion [8].

The idea of the approach is to measure the order of internal variances variability. The criterion maximizes the function $KL(k)$:

$$KL(k) = \frac{\text{DIFF}(k)}{\text{DIFF}(k+1)}$$

where $\text{DIFF}(k) = (k-1)^{2/d}W_{k-1} - k^{2/d}W_k$, $d$ is dimension of space.

Hartigan criterion [9].

$$H_K = (W_K / W_{K+1} - 1)(N - K - 1),$$

where $N$ is a number of objects, and $W_K$ and $W_{K+1}$ are minimum values of the criterion for the total dispersion of parameters divided into $K$ and $K+1$ clusters. The Hartigan’s rule says that until $H_K$ is less than 10, it is necessary to split sequentially a set into an increasing number of clusters $K$; this $K$ will be taken as an estimation of a clusters number.

Silhouette criterion [3, 13].

Suppose the data has been ungrouped into $k$ clusters (groups). Define $a(i)$, as an average distance from a sample $i$ to other samples of the same cluster for each instance $i$. The value $a(i)$ shows how
correctly the object $i$ is assigned to the given cluster (a smaller value indicates a more reasonable classification of this object as a group).

An average distance from object $i$ to cluster $c$ will be defined as an average distance from $i$ to objects $c$. Define $b(i)$ as the smallest average distance from object $i$ to any other cluster (to which this object does not apply). The cluster with the smallest average distance will be called the “neighboring cluster” to $i$. Define the silhouette:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

Put it down as follows:

$$s(i) = \begin{cases} 
1 - \frac{a(i)}{b(i)}, & \text{if } a(i) < b(i), \\
0, & \text{if } a(i) = b(i), \\
\frac{a(i)}{b(i)} - 1, & \text{if } a(i) > b(i).
\end{cases}$$

So it follows that: $-1 \leq s(i) \leq 1$

A value of $s(i)$ is close to one means that an object is classified correctly. If a value of $s(i)$ is close to (-1), then this means that it would be more correct to attribute this object to a neighboring cluster. If $s(i)$ is near zero, then it means that an object is on the border of two clusters. Then $a(i) < b(i)$ is required for $s(i)$ be close to 1.

An average $s(i)$ for all cluster objects shows how dense the corresponding group is and how correctly the grouping is carried out as a whole. If while grouping, too great or too small clusters were selected, then this criterion will take on a low value. Thus, this criterion can be applied to choose a number of clusters.

The disadvantage of using the criterion is the requirement of significant computing resources. Therefore it sometimes makes sense to use an accelerated criterion, which is calculated much faster, instead of the “classical” silhouette criterion. But it is practically inferior to the original criterion in quality [3]. In calculating the accelerated criterion, we define $a(i)$ as a distance from the object $i$ to the center of the cluster to which it belongs. In turn, $b(i)$ is defined as a distance that is the smallest to the centers of the clusters to which the given object does not belong, that is, a distance to the center of the neighboring cluster.

Experiments on the application of each of the above criteria were carried out for industrial production applying the results of non-destructive tests of prefabricated production batches of electrical radio products as an example [14]. Since initially a number of groups (parties) is unknown, a series of clustering problems is solved immediately. Then, the likelihood of such a group is estimated using criteria (methods for estimating the value of a clusters number). It was presented in [2, 4, 15] that there are no any universal methods (algorithms) for clustering for all data sets; therefore, there is no universal criterion for estimating a number of clusters. Although the application of the silhouette criterion gave the smallest number of errors in determining a number of production batches [14], a collective (ensemble) approach was applied to increase the accuracy of determining a number of clusters in production batches of space radio-electronic products [15].

3. Ensembles of clustering algorithms and experimental results

As it was already presented in [4, 15–17], the ensemble approach is one of the most promising directions in cluster analysis and helps to increase the stability of solutions in problems of automatic grouping. The main idea of the ensemble approach is to form a consistent clustering result based on several data grouping alternatives. The combination of various cluster solutions allows improving the quality of the results and their resistance to parameters changing [4, 16, 17].
According to the practical implementation, the formation of effective ensembles is difficult, since the choice for the formation of an ensemble, for example, clustering algorithms that show the best results, does not always lead to the formation of an ensemble that gives the best accuracy [2, 4, 14, 15]. Therefore, in compiling an ensemble of criteria for determining a number of clusters of automatic grouping algorithms, the results obtained in various ways to estimate the number of clusters will be applied.

There exist two main techniques for obtaining an ensemble [15]:

- Calculation of a co-occurrence matrix.
- Finding a consensus split, i.e., a consistent split for several existing solutions, optimal according to some criterion.

To select the clustering algorithms one should apply a procedure for compiling optimal ensembles of automatic grouping algorithms with the combined application of the genetic algorithm of the greedy heuristic method and the consistent binary partition matrix for practical problems [15]. In this procedure, there is a unit for determining a number of groups (number of clusters) at the decision-making level. In this block, an ensemble of criteria for automatically determining a number of clusters for automatic grouping algorithms is applied.

The accuracy of the separate criteria for automatically determining a number of clusters and their ensembles can be estimated from the available marked-up sample, i.e., a sample is required where the belonging the objects to the actual groups is known in advance.

The accuracy of the criteria and their ensembles will be estimated by the formula:

\[ \text{Fit}^1 = A / N \rightarrow \text{max}, \]

where \( A \) is a number of correctly clustered objects; \( N \) is a total number of objects.

The accuracy of the ensemble solution (in percent) is calculated by the formula:

\[ \text{Accuracy} = \text{Fit}^1 \cdot 100\% \]

The system was as follows: of Intel Core 2 Duo E8400CPU, 4GBRAM, NVIDIA GeForce 9600 GT graphics processor, with 2048 MB of RAM. The task was to divide the combined batch into homogeneous components followed by the analysis of the quality of this separation. The results of non-destructive test tests of prefabricated production batches of electric radio production (the composition of which is known in advance) were taken as test data sets. The production was carried out in a specialized test center of ITC-NPO PM JSC (Zheleznogorsk) for the assembly of spacecraft onboard equipment [4, 14, 17]. The combined batches were specially equipped from several obviously homogeneous batches of electric radio production:

- 140UD25AVK - 2 production batches (clusters) and a relatively small amount of data (56 vectors each with a dimension of 18);
- 3OT122A - 2 batches (767 vectors each of dimension 10);
- 1526LE5 - 3 batches (620 vectors each of dimension 41).

The multidimensional data collected while non-destructive tests for deviation from the specified parameters were applied. It means that clustering is performed on production batches of industrial production with a number of clusters up to 100, data volume up to tens of thousands of vectors, dimension up to hundreds of measurements. The procedure for dividing batches should be regulated; recalculation based on the same data should yield the same or very close results. It requires large computational and time resources. Therefore, in practice, samples from the electric and radio products of the supplied industrial batch are applied.
For each of the batches, automatic grouping algorithms were run 10 times with different values of the parameter $k$ (a number of cluster-batches). At the output of a computational experiment, clustering was estimated by the accuracy parameter. Consider the accuracy as the percentage of data objects assigned to the “right” cluster. We evaluate this “correctness” by a sample of labeled data for which their assignment to a particular cluster is known in advance. In this case, our samples were combined from the data of separate homogeneous batches of electric radio production.

Then, we calculate three “winners” of each criterion (the three best values of the parameter $k$) for each clustering algorithm. We assign weight to the “winners”; the first one has the highest weight, the second one has less weight and the third one has the smallest weight. After that, we calculate a total sum of the weights of all the criteria for each parameter $k$ (a number of clusters). Parameter $k$ with the largest total sum of weights wins (it is selected). The accuracy of the ensemble of criteria for estimating a number of clusters is calculated as the coincidence percentage of the resulting ensemble clusters number with a number of clusters of the actual batch of radio and radio products (with a predetermined number $k$).

Compose ensembles of three and five the most accurate criteria for evaluating a number of clusters (Tables 1) for the product data set 1526LE5. In fact, there can be any number of criteria. The number depends on the problem being solved, computational resources and the time available to the researcher (or a specialist in a particular enterprise) [4, 14, 17]. First, calculations are made by all criteria for each data set over the entire given range of the possible number of clusters. After that, three (five) criteria are selected. They showed the best indicators on the training set of electric radio production of this type and an ensemble of criteria is already compiled for determining a number of homogeneous groups in a combined batch of industrial production.

| A number of clusters | 1     | 2     | 3     | 4     |
|----------------------|-------|-------|-------|-------|
| Ensemble of three    | 87,6  | 91,3  | 97,5  | 89,8  |
| Ensemble of five     | 87,8  | 91,9  | 98,1  | 90,5  |

4. Conclusions
The application of the ensemble approach can be more effective in comparison with individual criteria for estimation a number of clusters. Moreover, individual criteria can show a result that is superior in accuracy to the ensemble result, but the accuracy of the ensemble is still higher than the average accuracy of separate criteria [4, 17].

It is also necessary to take into account a number of criteria used in the ensemble for a specific task (specific data set), due to the fact that the accuracy of the ensemble of criteria for different data sets changes when a number of criteria in the ensemble changes. Also, it is obligatory to take into account that an increase in a number of criteria in the ensemble needs the additional time and resource volumes.

Since in practice it is impossible to determine the accuracy of automatic grouping of objects due to the lack of information about the actual structure of the sample, and moreover, it is impossible to predict what algorithms in a particular case will show the most adequate results, the application of the ensemble approach to solving such problems is promising and relevant.

References
[1] Jain A K 2010 Data clustering: 50 years beyond K-means Pattern Recognition Letters 31 651-66
[2] Rozhnov I P, Orlov V I and Kazakovtsev L A 2019 VNS-based algorithms for the centroid-based clustering problem Facta Universitatis. Ser. Math. Inform. 34(5) 957–72 DOI.org/10.22190/FUMI1905957R
[3] Golovanov S M, Orlov V I, Kazakovtsev L A and Popov A M 2019 Recursive clustering algorithm based on silhouette criterion maximization for sorting semiconductor devices by
homogeneous batches IOP Conf. Series: Materials Science and Engineering 537 022035 DOI:10.1088/1757-899X/537/2/022035

[4] Rozhnov I P, Orlov V I and Kazakovtsev L A 2018 Increase in Accuracy of the Solution of the Problem of Identification of Production Batches of Semiconductor Devices 14th International Scientifical-Technical Conference on Actual Problems of Electronic Instrument Engineering, APEIE 2018 363-7 DOI: 10.1109/APEIE.2018.8546294

[5] Kazakovtsev L 2012 Algorithm for approximate solution of the generalized Weber problem with an arbitrary metric. Proceedings - UKSim-AMSS 6th European Modelling Symposium, EMS pp 109-14

[6] Calinski T and Harabasz J 1974 Adendrite method for cluster analysis Communications in Statistics 3 1-27 DOI: 10.1080/03610927408827101

[7] Davies D L and Bouldin D W 1979 A Cluster Separation Measure IEEE Transactions on Pattern Analysis and Machine Intelligence 1(2) 224–7

[8] Krzanowski W and Lai Y 1985 A criterion for determining the number of groups in a dataset using sum of squares clustering Biometrics 44 23–34

[9] Hartigan J A 1975 Clustering Algorithms (New York: Wiley) p 369

[10] Schwarz G 1978 Estimating the Dimension of a Model Annals of Statistics 2 461–4 DOI:10.1214/aos/1176344136

[11] Tibshirani R, Walther G and Hastie T 2001 Estimating the number of clusters in a data set via the gap statistic Journal of the Royal Statistical Society 63 411–23

[12] Akaike H 1974 A new look at the statistical model identification IEEE Transactions on Automatic Control 19(6) 716–23. DOI:10.1109/TAC.1974.1100705

[13] Roussseeuw P 1987 Silhouettes: a graphical aid to the interpretation and validation of cluster analysis Journal of Computational and Applied Mathematics 20 53–65

[14] Kazakovtsev L 2016 The greedy heuristics method for systems of automatic grouping of objects Diss ... Dr. tech. of science Krasnoyarsk

[15] Rozhnov I P, Kazakovtsev L A and Popov A M 2019 Scheme of optimal ensembles of clustering algorithms with a combined use of the Greedy Heuristics Method and a matched binary partitioning matrix IOP Conf. Series: Earth Environ. Sci. 315 032031

[16] Berikov V 2014 Weighted ensemble of algorithms for complex data clustering. Pattern Recognition Letters 38 99-106

[17] Rozhnov I, Orlov V and Kazakovtsev L 2018 Ensembles of clustering algorithms for problem of detection of homogeneous production batches of semiconductor devices. In the collection: CEUR Workshop Proceedings "OPTA-SCL 2018. Proceedings of the School-Seminar on Optimization Problems and their Applications". CEUR-WS. 2098 pp 338-48