Scheme of optimal ensembles of clustering algorithms with a combined use of the Greedy Heuristics Method and a matched binary partitioning matrix

I P Rozhnov, L A Kazakovtsev and A M Popov

Reshetnev Siberian State University of Science and Technology, prosp. Krasnoyarskiy Rabochiy 31, Krasnoyarsk 660031, Russia

E-mail: levk@bk.ru

Abstract. We propose a procedure for building optimal ensembles of automatic grouping (clustering) algorithms with a combined use of the genetic algorithm of the Greedy Heuristics Method and a matched binary partitioning matrix for practical problems. It is shown that our procedure allows us to increase the accuracy of separation of a mixed batch of industrial products into homogeneous batches.

1. Introduction

The agro-industrial complex of Russia has risen to a certain level of maturity, as evidenced by the growing competition among producers and the stabilization of the level of investment in agriculture. The volume and quality of modern tools and solutions, including systems for collecting, storing and processing data, are increasing in the agro-industrial complex. The volume of data is also growing, and as a result, the need for their qualitative analysis and reliable conclusions for decision making [1, 2].

The increasing use of high-dimensional data in the agro-industrial complex stimulates an increased interest in the development and application of methods and tools for processing and analyzing massive data of huge volumes. In the analytics of intensive data, one of the promising areas is cluster analysis, which allows solving the problem of reducing the dimension of the initial data set, identifying patterns, etc. [3].

2. Problem of identifying of homogeneous production batches

One of the most important parts of the problem of increasing the complex technical system reliability is to supply its critical nodes with details of increased quality requirements (for example, in the case of electronic equipment). To ensure the coordinated work of the same type of system elements, it is important that they have very similar characteristics (are homogeneous). The uniformity of characteristics of identical elements of the system is achieved if these elements were made from the same batch of raw materials in the same production batch. Therefore, when assembling critical system components, it is necessary to use the corresponding components, which are subject to increased quality requirements.

In order to extend the results of selective destructive tests to the entire production batch, it is necessary to be sure that we are dealing with a batch of products made from a single (homogeneous) batch of raw materials. Therefore, the identification of homogeneous production batches of...
prefabricated batches is one of the most important activities during the quality testing [2, 4]. In practice, the testing centre performs a series of tests, from tens to several thousand for each product, the results are tabulated and serve as data for analysis. The procedure of separation of parameters should be regulated, the re-calculation of data should give the same or very close results. From the above, a task is formed: the division into homogeneous production batches on the basis of test data (clustering). All data collected during tests for deviations from the specified parameters are used. This means that clustering is performed in a multidimensional space [2].

Modern methods of cluster analysis offer a wide range of tools for identifying groups that are heterogeneous in terms of the aggregate parameters. At the same time, the method of identifying groups (clusters) of different parameters should produce reproducible results. The algorithms proposed in [2, 3, 5] allow increasing the accuracy of automatic grouping methods for identifying groups of products with different parameters.

3. Ensembles of clustering algorithms

A lot of statistical and other methods are suggested for data mining problems, however, the development of a technology (method), suitable for solving the widest variety of clustering problems remains an important unsolved problem. For example, after repeated studies, the use of ensembles of clustering algorithms allows us to make a conclusion about their comparative effectiveness for solving various problems [6-8]. The method of forming such an ensemble is a question. As practice shows, the formation of effective ensembles is a difficult problem; choice to create an ensemble of algorithms that demonstrate the best results does not always lead to the formation of an ensemble giving the best accuracy [4, 6].

In the ensemble approach, we compose a preliminary binary matrix of differences of size n x n (where n is the number of objects) for each of standalone clustering algorithms.

\[ H_i = \{ h_i(i, j) \} \]

where \( h_i(i, j) \) equals 0 if the \( i \)th and \( j \)th elements belong to the same cluster, 1 otherwise.

The next step in composing an ensemble of clustering algorithms is to compile a matched matrix of binary partitions.

\[ H^* = \{ h^*(i, j) \}, \quad h^*(i, j) = \sum w_i h_i(i, j), \]

where \( w_i \) is the weight of the algorithm. We assume that the weight is equal to the averaged accuracy of the standalone algorithm implemented on a set of test problems.

Genetic algorithms demonstrate high efficiency in building ensembles of neural networks which are used for solving problems of automatic grouping. We applied the genetic algorithm of the Greedy Heuristic Method [5, 9] to form an ensemble of arbitrary algorithms. The choice of this method is due to the fact that the algorithms of this method for practical problems produce results that are difficult to significantly improve by other methods in comparable time. In addition, computational experiments show good results (by the value of the objective function and the stability of these values) for the tasks of automatic grouping of a large number of objects (hundreds of thousands) and data vectors of high dimensionality.

The accuracy of standalone clustering algorithms and can be estimated from the available marked sample. That is, we need a data set in which the belonging of objects to actual groups is known in advance.

The accuracy of algorithms and their ensembles will be evaluated as follows:

\[ Fit^1 = A / N \to \max, \]

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

Procedure for constructing the optimal ensembles of clustering algorithms
Required: a set of $m$ test problems with marked data (actual division of the data array into groups of data must be known), set of $n$ clustering algorithms $C$, population size $q$, number of algorithms in the resulting ensemble $p$.

The solutions ("individuals") in the algorithm are subsets $S$ of $p$ algorithms selected for the ensemble.

Step 1. Generate randomly $q$ initial solutions (individuals) of the genetic algorithm.

Step 2. Evaluate the criterion (1), having calculated its average value for $m$ problems, after implementing the ensemble encoded by each of the “individuals” to each of the problems. Store the averaged criterion to variable $Fit_i$ where $j$ is the index of the “individual”.

Step 3. Check the stop conditions (time limitation), STOP if needed.

Step 4. Select randomly two indexes of the “individuals” $i, j$. Compose the ensemble: $S = S_i \cup S_j$.

Step 5. While $|S| > p$ do:
   Step 5.1. For each $i$: $C_i \in S$ do:
      Step 5.1.1. Remove the $i$th algorithm from the ensemble $S$: $S' = S \setminus C_i$.
      Step 5.1.2. For $S'$, evaluate the criterion (1) averaged by $m$ problems, having implemented the ensemble $S'$ to each of the problems. Store the average criterion value to variable $Fit_i$.
      Step 5.1.3. Next iteration 5.1.
   Step 5.2. From $S$, remove the algorithm $C_i$ which has the least value of $Fit_i$. $S' = S \setminus C_i$.
   Step 5.3. Next iteration 5.

Step 6. For $S$, evaluate the criterion (1) averaged for $m$ problems, having implemented the ensemble $S$ for each of problems. Store the averaged criterion value to $Fit_{new}$.

Step 6. Select an index $k$ of the “individual” with the least value of $Fit$. If $Fit_{new} > Fit_k$, then replace the $k$th “individual” by $S, S_k = S$; $Fit_k = Fit_{new}$.

Go to Step 2.

A diagram of the procedure for compiling optimal ensembles of automatic grouping algorithms is shown in figure 1. First, calculations are performed by all algorithms for each data set, after which one algorithm of each model is selected that has shown the best indicators of the objective function and an automatic grouping algorithm is compiled from them. Note that with the help of a genetic algorithm, an ensemble of models is actually made up, and within each model, the choice occurs without the participation of the genetic algorithm.

The diagram shows four models of automatic grouping algorithms using the algorithms described in [2, 3, 5]. In fact, there can be any number of models (as well as algorithms in each model) (as indicated by the ellipses in the diagram). Their number depends on the particular problem being solved, computational resources, and the time available to the researcher (or a specialist at a particular enterprise).

We apply this procedure to the task described above of creating optimal ensembles for solving problems of automatic grouping of products into production batches. The genetic algorithms of the greedy heuristics method do not require a large population for their work. We used $q = 10$ to compose ensembles of 3 and 5 algorithms ($p = 3, p = 5$)

4. Experimental results
For our study, we used classical datasets from the repositories UCI (Machine Learning Repository) and Clustering basic benchmark.

We applied five basic clustering algorithms: k-Means, k-Means-fast, k-Means-kernel (k-means core method), k-Medoids, EM (Expectation Maximization).

At the output of this process, we evaluate clustering by the Accuracy clustering parameter. By accuracy, we mean the proportion of data objects assigned to the "right" cluster. This "correctness" can be assessed by having a sample of marked data, for which it is known in advance that they are assigned to a particular cluster. In this case, our samples are combined from data from separate homogeneous batches of the ECB. The results are summarized in table 1.
We realized two versions of all clustering algorithms: 1 – default and 2 – optimized. In the second versions, we try to improve the accuracy of the algorithms varying one of the algorithm parameters. For algorithms such as k-Means, k-Means(fast) and k-Medoids, we varied the type of the distance measure. For the k-Means(kernel), we varied the kernel type (dot/radial kernel).

Let us compose the ensembles of 3 and 5 algorithms (table 2) for each of datasets listed in table 1.

Figure 1. Composition of optimal ensembles of clustering algorithms.
### Table 1. Results of computational experiments with standalone clustering algorithms on datasets from the UCI repository.

| Algorithm   | Accuracy / optimized parameter value |
|-------------|--------------------------------------|
| Cryotherapy 2 clusters | 56.67 (Euclidean Distance) |
| pima-indians-diabetes 2 clusters | 66.02 (Euclidean Distance) |
| ionosphere 2 clusters | 71.23 (Euclidean Distance) |
| Iris 3 clusters | 89.33 (Euclidean Distance) |
| Zoo 7 clusters | 75.25 (Euclidean Distance) |
| k-Means (fast) | 56.67 (Euclidean Distance) |
| k-Means (kernel) | 55.56 (radial kernel) |
| k-Medoids | 57.78 (Euclidean Distance) |
| EM | 56.67 (100 steps) |
| k-Means-Optim.² | 75.56 (Camberra Distance) |
| k-Means (fast)-Optim. | 75.56 (Camberra Distance) |
| k-Means (kernel)-Optim.² | 53.33 (dot kernel) |
| k-Medoids-Optim.² | 73.33 (Camberra Distance) |
| EM-Optim. | 56.67 (1 optimization step) |

### Table 2. Results of computational experiments with ensembles production batches (accuracy).

| Dataset / algorithm | Cryotherapy 2 clusters | pima-indians-diabetes 2 clusters | ionosphere 2 clusters | Iris 3 clusters | Zoo 7 clusters |
|---------------------|------------------------|----------------------------------|----------------------|----------------|----------------|
| The best standalone algorithm | 75.56 | 66.28 | 72.36 | 97.33 | 83.17 |
| Ensemble of 3 algorithms | 75.56 | 66.28 | 71.23 | 96.71 | 83.17 |
| Ensemble of 5 algorithms | 75.56 | 65.89 | 68.66 | 96.67 | 81.15 |

5. Conclusions

Using the ensemble approach can be more efficient in comparison with individual clustering algorithms. In this case, individual algorithms are able to show results that exceed the results of the ensemble in accuracy, however, the accuracy of the ensemble is still higher than the average accuracy of individual algorithms [4, 6]. It is also necessary for a specific problem to take into account the number of algorithms.
used in an ensemble, due to the fact that the accuracy of an ensemble of automatic grouping algorithms for different data sets changes as the number of algorithms in an ensemble changes. Since in practice clustering accuracy cannot be determined due to the lack of information on the actual composition of the data array, and it is impossible to a priori predict which of the algorithms will show the most adequate results in a particular case, the use of an ensemble approach to solving such problems is promising and relevant. In particular, the use of the ensemble approach in combination with the algorithms for the automatic grouping of the greedy GH-VNS heuristics method, which provide the best result within the framework of a given model, will allow you to get results not only more adequate, but also reproducible with repeated launches of the algorithm.

Acknowledgements
Results were obtained in the framework of the state task No. 2.5527.2017/8.9 of the Ministry of Education and Science of the Russian Federation.

References
[1] Moskalev S.M. and Klimenok-Kudinova N.V. 2018 Artificial Intelligence and the Internet of Things as the Innovative Methods for Improving the Agroindustrial Sector Izvestiya Sankt-Peterburgskogo gosudarstvennego agrarnogo universiteta 3 (52) 121-30
[2] Orlov V, Stashkov D, Kazakovtsev L, Rozhnov I, Nasyrov I and Kazakovtseva O 2018 Improved method of forming production batches of electronic components with special quality requirements Modern high technology 1 37-42
[3] Orlov V, Kazakovtsev L, Rozhnov I, Popov N and Fedosov V 2018 Variable neighbourhood search algorithm for k-means clustering IOP Conf. Series: Materials Science and Engineering 450 022035 DOI:10.1088/1757-899X/450/2/022035
[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 Scientific-Technical Conference on Actual Problems of Electronic Instrument Engineering, APEIE 2018 363-67 DOI: 10.1109/APEIE.2018.8546294.
[5] Kazakovtsev L, Stashkov D, Gudyma M and Kazakovtsev V 2019 Algorithms with Greedy Heuristic Procedures for Mixture Probability Distribution Separation Yugoslav Journal of Operations Research 29 51-67
[6] 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 338-48
[7] Kausar N, Abdullah A, Samir B B, Palaniappan S, AlGhamdi B S and Dey N 2016 Ensemble Clustering Algorithm with Supervised Classification of Clinical Data for Early Diagnosis of Coronary Artery Disease. Journal of Medical Imaging and Health Informatics 6(1) 78-87
[8] Berikov V 2014 Weighted ensemble of algorithms for complex data clustering Pattern Recognition Letters 38 99-106
[9] Kazakovtsev L A 2016 The greedy heuristics method for systems of automatic grouping of objects (Krasnoyarsk)