UDC 519.6
Doi: 10.31772/2587-6066-2020-21-3-347-354

For citation: Shkaberina G. Sh., Kazakovtsev L. A., Li R. Models and algorithms for automatic grouping of objects based on the k-means model. *Siberian Journal of Science and Technology.* 2020, Vol. 21, No. 3, P. 347–354. Doi: 10.31772/2587-6066-2020-21-3-347-354

Для цитирования: Шкаберина Г. Ш., Казаковцев Л. А., Ли Ж. Модели и алгоритмы автоматической группировки объектов на основе модели k-средних // Сибирский журнал науки и технологий. 2020. Т. 21, № 3. С. 347–354. Doi: 10.31772/2587-6066-2020-21-3-347-354

MODELS AND ALGORITHMS FOR AUTOMATIC GROUPING OF OBJECTS BASED ON THE K-MEANS MODEL

G. Sh. Shkaberina*, L. A. Kazakovtsev, R. Li

Reshetnev Siberian State University of Science and Technology
31, Krasnoyarskii rabochii prospekt, Krasnoyarsk, 660037, Russian Federation
*E-mail: z_guzel@mail.ru

The paper is devoted to the study and development of new algorithms for automatic grouping of objects. The algorithms can improve the accuracy and stability of the result of solving practical problems, such as the problems of identifying homogeneous batches of industrial products. The paper examines the application of the k-means algorithm with the Euclidean, Manhattan, Mahalanobis distance measures for the problem of automatic grouping of objects with a large number of parameters. A new model is presented for solving problems of automatic grouping of industrial products based on the k-means model with the Mahalanobis distance measure. The model uses a training procedure by calculating the averaged estimate of the covariance matrix for the training sample (sample with pre-labeled data). A new algorithm for automatic grouping of objects based on an optimization model of k-means with the Mahalanobis distance measure and a weighted average covariance matrix calculated from a training sample is proposed. The algorithm allows reducing the proportion of errors (increasing the Rand index) when identifying homogeneous production batches of products based on the results of tests. A new approach to the development of genetic algorithms for the k-means problem with the use of a single greedy agglomerative heuristic procedure as the crossover operator and the mutation operator is presented. The computational experiment has shown that the new mutation procedure is fast and efficient in comparison with the original mutation of the genetic algorithm. The high rate of convergence of the objective function is shown. The use of this algorithm allows a statistically significant increase both in the accuracy of the result (improving the achieved value of the objective function within the framework of the chosen mathematical model for solving the problem of automatic grouping), and in its stability, in a fixed time, in comparison with the known algorithms of automatic grouping. The results show that the idea of including a new mutation operator in the genetic algorithm significantly improves the results of the simplest genetic algorithm for the k-means problem.

Keywords: automatic grouping, k-means, Mahalanobis distance, genetic algorithm.

МОДЕЛИ И АЛГОРИТМЫ АВТОМАТИЧЕСКОЙ ГРУППИРОВКИ ОБЪЕКТОВ НА ОСНОВЕ МОДЕЛИ K-СРЕДНИХ

Г. Ш. Шкаберина, Л. А. Казаковцев, Ж. Ли

Сибирский государственный университет науки и технологий имени академика М. Ф. Решетнева
Российская Федерация, 660037, г. Красноярск, просп. им. газ. «Красноярский рабочий», 31
*E-mail: z_guzel@mail.ru

Работа посвящена исследованию и разработке новых алгоритмов автоматической группировки объектов, которые позволяют повысить точность и стабильность результата решения практических задач, например, таких как задача выделения однородных партий промышленной продукции. В статье исследуется применение алгоритма k-средних с Евклидовым, Манхэттенским, Махаланобиса мерами расстояния для задачи автоматической группировки объектов с большим количеством параметров. Представлена новая модель для решения задач автоматической группировки промышленной продукции на основе модели k-средних с мерой расстояния Махаланобиса. Данная модель использует процедуру обучения путем вычисления усредненной оценки ковариационной матрицы для обучающей выборки (выборка с предварительно размеченными данными). Предложен новый алгоритм автоматической группировки объектов, основанный на оптимизационной модели k-средних *с мерой расстояния Махаланобиса и средневзвешенной ковариационной матрицей, рассчитанной по обучающей выборке. Алгоритм позволяет снизить долю ошибок (повысить индекс Рэнда) при выявлении однородных производственных партий продукции по результатам тестовых испытаний. Представлен новый подход к разработке генетических алгоритмов для задачи k-средних с применением единой жадной агломеративной эвристической процедуры в качестве оператора скрещивания и оператора мутации. Вычислительный эксперимент показал, что новая процедура мутации является быстрой и эффективной по сравнению с исходной мутацией генетического алгоритма, показана высокая скорость сходимости целевой функции. Применение данного алгоритма позволяет статистически значимо повысить точность результата (улучшить достигаемое значение целевой функции в рамках выбранной математической модели решения задачи автоматической группировки), а также его стабильность за фиксированное время по сравнению с известными алгоритмами автоматической группировки. Результаты показывают, что идея включения нового оператора мутации в генетическом алгоритме значительно улучшает результаты простейшего генетического алгоритма для задачи k-средних.*

Ключевые слова: автоматическая группировка, k-средних, расстояние Махаланобиса, генетический алгоритм.

Introduction. Automatic grouping (AG) involves dividing a set of objects into subsets (groups) so that objects from one subset are more similar to each other than to objects from other subsets according to some criterion. General characteristics of the object and the methods by which the division took place are taken into account in the process of grouping objects of a certain set into certain groups (subsets).

To exclude the emergence of unreliable electrical radio products intended for installation in the on-board equipment of a spacecraft with a long period of active existence, the entire electronic component base passes through specialized technical test centers [1; 2]. These centers perform operations of full incoming inspection of electrical radio products, additional verification tests, diagnostic non-destructive testing and selective destructive physical analysis. Detection of initial homogeneous production batches of electrical radio products from shipped batches is an important stage during testing [1].

The *k*-means model is one of the best known cluster analysis models. The goal is to find *k* points (centers) X_1, …, X_k in *d*-dimensional space, such that the sum of the squared distances from known points (data vectors) A_1, …, A_N to the nearest of the required points (centers) reaches a minimum [3]:

$$\arg\min F(X_1,...,X_k) = \sum\nolimits_{i=1}^{N} \min_{j\in\{1,k\}} \left\| X_j - A_i \right\|^2 . \qquad (1)$$

Initially it is necessary to predict the number of groups (subsets) in the *k*-means algorithm. In addition, the result obtained depends on the initial choice of centers. The distance function and its definition also play an important role in the problem of dividing the set under study into groups.

The first genetic algorithm for solving the discrete p-median problem was proposed by Hosage and Goodchild [4]. The algorithm [5] gives fairly accurate results. However, the rate of convergence of the objective function is very slow. In their work O. Alp, E. Erkut, Z. Drezner [6] presented a faster simple genetic algorithm with a special recombination procedure, which also gives accurate results. These algorithms solve discrete problems. The authors of the work “Genetic algorithm-based clustering technique” [7] encode solutions (chromosomes) in their GAs as sets of centroids, represented by their coordinates (vectors of real numbers) in a multidimensional space.

The analysis of the literature has shown that the existing solutions in the field of AG of multidimensional objects either have high accuracy, or ensure the stability of the result with multiple runs of the algorithm, or have high speed of operation, but do not combine all these qualities at the same time. To date, algorithms for *k*-means and *k*-medians have been developed only for the most common distance measures (Euclidean, Manhattan). However, taking into account the feature space peculiarities of a specific practical problem when choosing a distance measure can lead to increasing the accuracy of AG objects. In the presented work, we use the Rand Index (RI) [8] as a measure of the clustering accuracy.

It is extremely difficult to improve the AG result of multidimensional objects with increased requirements for the accuracy and stability of the result using known algorithms without a significant increase in time costs. When solving practical problems of the AG of multidimensional data, for example, the problems of identifying homogeneous batches of industrial products, the adequacy of the models and, as a result, the accuracy of the AG of industrial products are questionable. It is still possible to develop algorithms that further improve the result based on the chosen model, for example, the *k*-means model.

In a multidimensional feature space, there is often a correlation between individual features and groups of features. The use of correlation dependences can be used by moving from search in the space with the Euclidean or rectangular metric to search in the space with the Mahalanobis metric [9–11]. The square of the Mahalanobis distance D_M is defined as follows:

$$D_M(X) = (X-\mu)^T C^{-1}(X-\mu) \ , \qquad (2)$$

where *X* is the vector of values of the measured parameters, μ is the vector of mean values (for example, the center of the cluster), *C* is the covariance matrix.

The aim of the study in the presented work is to improve the accuracy and stability of the result of solving problems in automatic grouping of objects.

The idea of the work is to use the Mahalanobis distance measure with the averaged estimate of the covariance matrix in the *k*-means problem to reduce the proportion of the AG error in comparison with other known algorithms, and also to use the mutation operator as a part of the genetic algorithm to improve the accuracy and stability of the solution according to the achieved value of the objective function in a fixed execution time in comparison with known algorithms for separating objects.

Initial data. The study was carried out on the data of testing the batches of integrated circuits [12], intended for installation in space vehicles. The tests were carried out in a specialized center for technical tests. The data is a set of parameters for electrical radio products (ERP). The original batch of ERI belongs to different homogeneous batches, in accordance with the manufacturer's marking. The total number of products is 3987. In each batch, the product is described by 205 measured parameters. Batch 1 contains 71 products, batch 2 – 116 products, batch 3 – 1867 products, batch 4 – 1250 products, batch 5 – 146 products, batch 6 – 113 products, and batch 7 – 424 products.

The algorithm of k-means with the Mahalanobis distance with an averaged estimate of the covariance matrix over the training sample. The computational results of experiments on automatic grouping of industrial products with *k*-medoid and *k*-means models, in which the Mahalanobis metric is applied, show an increase in clustering accuracy with automatic grouping into 2–6 clusters and a small number of objects and informative features [13].

Instead of the covariance matrix from (2), it was proposed to calculate the averaged estimate of the covariance matrix for homogeneous batches of products (according to pre-labeled data) using the training sample:

$$C = \frac{1}{n}\sum_{j=1}^{k} C_j n_j \,, \qquad (3)$$

where n_j is the number of objects (products) in the j-th batch, n is the total sample size, C_j are the covariance matrices of individual batches of products.

In this paper, we propose an algorithm for automatic grouping of objects based on the *k*-means model with the adjustment of the Mahalanobis distance measure parameter (covariance matrix) based on the training sample:

The algorithm of 1. *k*-means with the Mahalanobis distance with averaged estimate of the covariance matrix

Step 1. Using the *k*-means method with Euclidean distance, divide the sample into a certain number of *k* clusters (here *k* is some expert estimate of the possible number of homogeneous groups, not necessarily accurate);

Step 2. Calculate the center μ_i for each cluster. The center is defined as the arithmetic mean of all points in the cluster

$$\mu_i = \frac{1}{m}\sum_{j=1}^{m} X_{ji} \,, \qquad (4)$$

where m is the number of points, X_j is a vector of values of one measured parameter (j = 1...m);

Step 3. Calculate the averaged estimate of the covariance matrix (3). If the averaged estimate of the covariance matrix is degenerate, go to step 4, otherwise go to step 5;

Step 4. Increase the number of clusters by (k + 1) and repeat steps 1 and 2. Form new clusters with the squared Euclidean distance:

$$D(X_j, \mu_i) = \sum_{i=1}^{n} (X_{ji} - \mu_i)^2 \,, \qquad (5)$$

where n is the number of parameters. Return to step 3 with a new training example (set).

Step 5. Match each point to the nearest centroid using the square of the Mahalanobis distance (2) with the averaged estimate of the covariance matrix C (3) to form new clusters.

Step 6. Repeat the algorithm from step 2 until the clusters stop changing.

The paper presents the results of three groups of experiments on the data of industrial product samples.

The first group. The training sample corresponds to the working sample for which the clustering was carried out.

The second group. The training and working samples are not the same. In practice, a test center can use retrospective data on deliveries and testing of products of the same type as a training sample.

The third group. The training and working samples also do not match, but the results of the automatic grouping of products (*k*-means in the multistart mode with the Euclidean metric) were used as the training sample.

In each group of experiments, for each working sample, the *k*-means algorithm was run 30 times with each of the five clustering models studied.

DM1model – *k*-means with the Mahalanobis distance, the covariance matrix is calculated for the entire training sample.

DC model – *k*-means with a distance similar to the Mahalanobis distance, but using a correlation matrix instead of a covariance matrix.

DM2 model – *k*-means with Mahalanobis distance, with averaged estimate of the covariance matrix.

DR model – *k*-means with the Manhattan distance.

DE model – *k*-means with the Euclidean distance.

For each model, the minimum (Min), maximum (Max), average (Average) values, standard deviation (Std.Dev) of the Rand index (RI) and the objective function, as well as the values of the coefficients of variation (V) and the range (R) of the target functions (tab. 1) are calculated.

It was found that the new *DM2* model with an averaged estimate of the covariance matrix shows the best accuracy among the presented models in almost all series of experiments according to the Rand index (RI) and in all cases it exceeds the *DE* model, where the Euclidean distance is used. The experiments also showed that in most cases the coefficient of variation of the objective function values is higher for the *DE* model, where the Euclidean measure of distance is used, and also that the coefficient of the range of the objective function values has the highest values for the *DM2* model, where the Mahalanobis distance measure with an averaged estimate of the covariance matrix is used.

Table 1

The results of a computational experiment on the data of the 1526IE10_002 microcircuit (3987 data vectors with a dimension of 68), the training sample consists of 10 batches, the third group, the working sample is made up of 7 batches of products)

V-series	Rand index (RI)					Objective function				
	Model					Model				
	DM1	DC	DM2	DR	DE	DM1	DC	DM2	DR	DE
Max	0.767	0.658	0.749	0.740	0.735	255886	379167	281265	18897	6494.62
Min	0.562	0.645	0.696	0.703	0.705	250839	36997	274506	17785	5009.42
Average	0.632	0.650	**0.725**	0.714	0.719	252877	37178	277892	18240	5249.95
Std .Dev	0.047	0.003	0.016	0.008	0.006	1164.5	152.8	2358.9	452.7	366.5
V						0.461	0.411	0.849	2.482	**6.981**
R						5047	920	**6759**	1112	1485

Therefore, multiple attempts to run the *k*-means algorithm or to use other algorithms based on the *k*-means model (for example, *j*-means [14] or greedy heuristic algorithms [15]) are required to obtain consistently good values of the objective function.

Genetic cross-mutation algorithm for the *k*-means problem. The new algorithm improves the accuracy of solving the *k*-means problem and the stability of the result in a fixed limited execution time. In this chapter, by the accuracy of the algorithm we mean exclusively the achieved value of the objective function, without taking into account the indicators of the model adequacy and the correspondence of the algorithm results to the actual (real) separation of objects, if known.

A very limited set of possible mutation operators is known for genetic algorithms for solving the *k*-means problem with real coding of solutions. For example, the authors of the work "Genetic algorithm-based clustering technique" [7] encode solutions (chromosomes) in their GAs as sets of centroids represented by their coordinates (vectors of real numbers) in a multidimensional space. Each chromosome undergoes mutation with a fixed probability μ. The procedure (operator) of mutation is as follows.

Algorithm 2 3.1 Initial GA mutation procedure for the *k*-means problem

Step 1. Generation of a random number $b \in (0,1]$ with uniform distribution;

Step 2. IF $b < \mu$, then the chromosome mutates. If the position of the current centroid is υ, then after mutation it becomes:

$$\upsilon \leftarrow \begin{cases} \upsilon \pm 2 \times b \times \upsilon, & \upsilon \neq 0, \\ \upsilon \pm 2 \times b, & \upsilon = 0. \end{cases}$$

The signs "+" and "–" have the same probability. The centroid coordinates are shifted randomly.

In our work we replaced this mutation procedure for the *k*-means problem with the following procedure.

Algorithm 3 3.2 GA cross mutation procedure for the k-means problem

Step 1. Generating a random initial solution $S = \{X_1 \ldots X_k\}$;

Step 2. Applying the k-means algorithm to S to obtain the local optimum $S^{'}$;

Step 3. Applying a simple crossover procedure for the mutated individual $S^{'}$ from the population and S to obtain a new solution $S^{''}$;

Step 4. Applying the k-means algorithm to $S^{''}$ to obtain local optimum $S^{''}$;

Step 5. IF $F(S^{''}) < F(S^{'})$, **THEN** $S^{'} \leftarrow S^{''}$.

The proposed procedure is used with a mutation probability of 1 after each crossover operator.

The results of running the original algorithm 2, described with a mutation probability of 0.01, and its version with algorithm 3 as a mutation operator are shown in the figure (population size $N_{POP} = 20$). The new mutation procedure is fast and efficient in comparison with the original mutation of the genetic algorithm; a high convergence rate of the objective function has been shown.

Greedy genetic algorithms and many other evolutionary algorithms for the *k*-means problem do without mutation. The idea of a greedy agglomerative heuristic procedure is to combine two known solutions into one unacceptable solution with an excessive number of centroids, and then the number of centroids is successively reduced. The centroid which shows the smallest increase in the objective function value (1) is removed at each iteration.

Algorithm 4. Basic greedy agglomerative heuristic procedure

It is given: the initial number of clusters K, the required number of clusters k, $k < K$, the initial solution $S = \{X_1, \ldots, X_K\}$, where $|S| = K$.

STEP 1. Improve the initial solution by the *k*-means algorithm

WHILE $K > k$

CYCLE for each $i^{'} \in \{\overline{1,\mathrm{K}}\}$ perform:

STEP 2. $S^{'} \leftarrow S\{X_i^{'}\}$**.** Improve the solution $S^{'}$ by the *k*-means algorithm and store the corresponding obtained values of the objective function (1) as variables $F_i^{'} \leftarrow F\{X^{'}\}$.

END OF CYCLE

STEP 3. Select the subset S_{elim} from the centers n_{elim}, $S_{elim} \in S$, $|S_{elim}| = n_{elim}$ with the minimum value of the corresponding variables $F_{i}^{'}$. $n_{elim} = \max\{1, 0.2 \cdot (|S| - k)\}$.

```
    STEP 4. Get a new solution $S \leftarrow S / S_{elim}$, $K = K - 1$.
    Improve the solution by the k-means algorithm.
END WHILE
```

The initial solution S can also be obtained by combining two known solutions. Algorithms 5 and 6 modify the initial solution by the second known solution. In fact, Greedy procedure 1 supplements the first set in turn with each element from the second set. Greedy procedure 2 combines both sets.

Algorithm 5. Greedy procedure 1 with partial join

```
It is given: Two sets of cluster centers $S^{'} = \{X^{'}_{1},...,X^{'}_{K}\}$ и $S^{"} = \{X^{"}_{1},...,X^{"}_{K}\}$
Cycle: for each $i^{'} \in \{\overline{1,K}\}$
    Step 1. Combine $S^{'}$ and one element from the set $S^{"}: S \leftarrow S \cup \{X^{"}_{1},...,X^{"}_{K}\}$.
    Step 2. Run Algorithm 3.3 with the initial solution $S$ and save the result.
END OF CYCLE
Step 3. Revert the best solutions saved in Step 2.
```

Algorithm 6. Greedy procedure 2 with full set union

```
It is given: Two sets of cluster centers $S^{'} = \{X^{'}_{1},...,X^{'}_{K}\}$ and $S^{"} = \{X^{"}_{1},...,X^{"}_{K}\}$
Step 1. Combine two sets of cluster centers $S \leftarrow S^{'} \cup S^{"}$.
Step 2. Run Algorithm 3 3.3 with the initial solution $S$.
```

The basic genetic algorithm (GA) for k-means problems is described as follows:

Algorithm 7. GA with the alphabet of real numbers for the k-means problem

```
It is given: Initial population size $N_{POP}$
STEP 1. Choose $N_{POP}$ of initial solutions $S_1,...,S_{N_{POP}}$, where $|S_i|=k$, and $\{S_1,...,S_{N_{POP}}\}$ is a randomly selected subset of the set of data vectors. Improve each initial solution by the k-means algorithm and store the corresponding obtained values of the objective function (1) as variables $f_k \leftarrow F(S_k), k = \overline{1, N_{POP}}$.
CYCLE
    STEP 2. IF the stop condition is met, THEN STOP. Return the solution $S_{i^*}, i^* \in \{\overline{1, N_{POP}}\}$ with the minimum value $f_{i^*}$.
    STEP 3. Randomly select 2 indices $k_1, k_2 \in \{\overline{1, N_{POP}}\}$, $k_1 \neq k_2$.
    STEP 4. Start crossing procedure: $S_c \leftarrow Crossingover(S_{k_1}, S_{k_2})$.
    STEP 5. Start mutation procedure: $S_c \leftarrow Mutation(S_c)$.
    STEP 6. Start the chosen selection procedure to change the population set.
END OF CYCLE
```

The following algorithm is proposed At STEP 6.

Algorithm 8. Selection procedure

```
STEP 1. Randomly select 2 indices $k_4, k_5 \in \{\overline{1, N_{POP}}\}$, $k_4 \neq k_5$.
STEP 2. IF $f_{k_4} > f_{k_5}$ THEN $S_{k_4} \leftarrow S_c$, $f_{k_4} \leftarrow F(S_c)$, ELSE $S_{k_5} \leftarrow S_c$, $f_{k_5} \leftarrow F(S_c)$.
```

A GA with greedy heuristic for p-medians and k-means problems can be described as follows:

Algorithm 9. A GA with greedy heuristic for p-medians and k-means problems (modifications GA-FULL, GA-ONE и GA-MIX)

```
It is given: Population size $N_{POP}$.
Step 1. Set $N_{iter} \leftarrow 0$. Choose a set of initial solutions $\{S_1,...,S_{N_{POP}}\}$, where $|S_i| = k$. Improve each initial solution by the k-means algorithm and store the corresponding obtained values of the objective function (1) as variables $f_k \leftarrow F(S_k), k = \overline{1, N_{POP}}$. In this work the initial value of the population is $N_{POP} = 5$.
Cycle
    Step 2. IF the stop condition is met, THEN STOP. Return the solution $S_{i^*}, i^* \in \{\overline{1, N_{POP}}\}$ with the minimum value $f_{i^*}$, ELSE set the population size as follows: $N_{iter} \leftarrow N_{iter} + 1$; $N_{POP} \leftarrow \max\{N_{POP,}[\sqrt{1+N_{iter}}]\}$; IF $N_{POP}$ has changed, THEN generate a new one $S_{N_{POP}}$ as described in Step 1.
    Step 3. Randomly select 2 indices $k_1, k_2 \in \{\overline{1, N_{POP}}\}$, $k_1 \neq k_2$.
    Step 4. Run Algorithm 5 (for GA-ONE*) or Algorithm 6 (for GA-FULL*) with solutions $S_{k_1}$ and $S_{k_2}$. For GA-MIX* Algorithm 5 or Algorithm 6 are chosen at random with equal probability. Get a new solution $S_c$.
    Step 5. $S_c \leftarrow Mutation(S_c)$ By default the mutation procedure is not used.
    Step 6. Run Algorithm 5
END OF CYCLE
```

* GA-ONE is a genetic algorithm with greedy heuristic with partial union, GA-FULL is a genetic algorithm with greedy heuristic with full union; GA-MIX is a random choice of algorithms 5 or 6

This algorithm uses a dynamically growing population. In our new version of Step 5, the cross mutation operator looks like this.

Algorithm 10. A cross mutation operator for Step 5 of Algorithm 9 (modifications GA-FULL-MUT, GA-ONE-MUT and GA-MIX-MUT)

```
Step 1. Run the k-means algorithm for a randomly selected initial solution to obtain solution $S^{'}$.
```

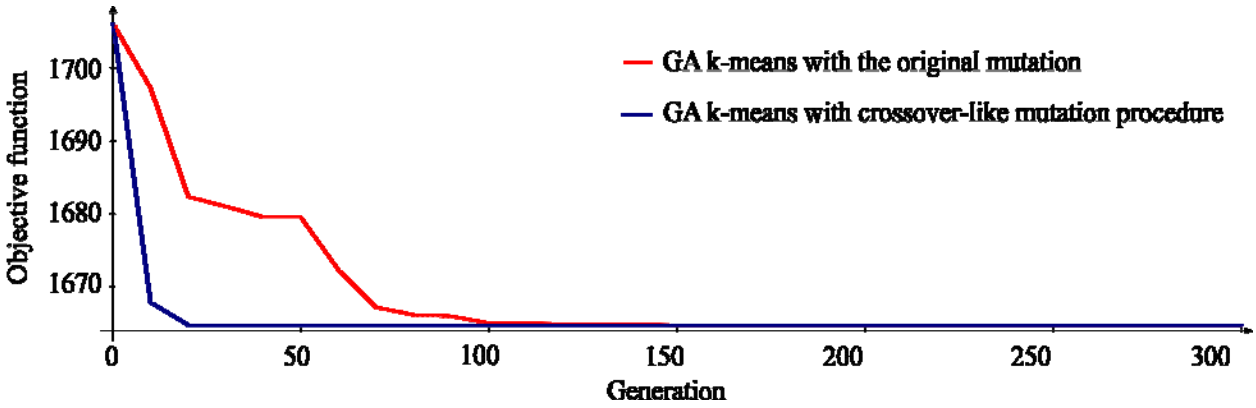
Step 2. Run Algorithm 5 (for GA-ONE) or Algorithm 6 (for GA-FULL) with solutions S_c and S''. Get a new solution $S_c^{'}$.

Step 3. If $F(S_c') < F(S_c)$, **THEN** $S_c \leftarrow S_c^{'}$.

Computational experiments with datasets from the Machine Learning Repository, Basic Benchmark repositories, as well as with data from industrial product samples (tab. 2) were carried out. New modifications of three GAs (GA-FULL-MUT, GA-ONE-MUT and GA-MIX-MUT) were compared with the well-known *j*-means and *k*-means algorithms (in multistart mode), GA without mutation (GA-FULL, GA-ONE and GA-MIX), automatic grouping algorithms for the k-means problem with combined application of search algorithms with alternating randomized neighborhoods formed by applying greedy agglomerative heuristics (*k*-GH-VNS1, *k*-GH-VNS2, *k*-GH-VNS3) and also for the *j*-means (*j*-means GH-VNS1, *j*-means GH-VNS2) problem. 30 attempts were made to run each algorithm for all datasets. For each algorithm, the minimum (Min), maximum (Max), average (Average) values and the standard deviation (Std.Dev.) of the objective function were calculated.

The best values of the new algorithms (*) are highlighted in bold, the best values of the known algorithms are indicated in italics, the most achieved values of the objective function are underlined (tab. 2). The Mann-Whitney U-test (↑↓↕) and Student's t-test (⇑⇓⇕) were used to confirm the statistical significance of the advantages (↑⇑) and disadvantages (↓⇓) of the new algorithms over the known algorithms.

The performed computational experiments show that GA with a greedy agglomerative crossover operator with a new idea of the mutation procedure is superior to GA without mutation in terms of the obtained value of the objective function.



Results for data set Mopsi-Joensuu (6014 data vectors of dimension 2), 300 clusters, time limit 3 minutes

Результаты для набора данных Mopsi-Joensuu (6014 векторов данных размерностью 2), 300 кластеров, 3 минуты

Table 2

Results of computational experiments for the Europe dataset (169309 data vectors of dimension 2), 30 clusters, 4 hours

Algorithm	Objective function value			
	Min	Max	Average	Std.Dev.
j-means	7.51477E+12	7.60536E+12	7.56092E+12	29.764E+9
k-means	7.54811E+12	7.57894E+12	7.56331E+12	13.560E+9
k-GH-VNS1	*7.49180E+12*	7.49201E+12	*7.49185E+12*	*0.073E+9*
k-GH-VNS2	7.49488E+12	7.52282E+12	7.50082E+12	9.989E+9
k-GH-VNS3	7.49180E+12	7.51326E+12	7.49976E+12	9.459E+9
j-means-GH-VNS1	7.49180E+12	7.49211E+12	7.49185E+12	0.112E+9
j-means-GH-VNS2	7.49187E+12	7.51455E+12	7.4962E+12	8.213E+9
GA-FULL-MUT*	7.49293E+12	7.49528E+12	7.49417E+12	0.934E+9
GA-MIX-MUT*	7.49177E+12	7.49211E+12	7.49186E+12	0.117E+9
GA-ONE-MUT*↑⇑	**7.49177E+12**	7.49188E+12	**7.49182E+12**	**0.042E+9**

Conclusion. The proposed new model of automatic grouping of industrial products and a new algorithm based on an optimization model of k-means with Mahalanobis distances and a trained covariance matrix can reduce the proportion of errors (increase the Rand index) when identifying homogeneous production batches of products. The presented new genetic algorithm for the k-means problem with the original idea of using one procedure as the crossover operator and the mutation operator demonstrates a more accurate and stable result of the objective function value in a fixed execution time.

Acknowledgments. This work was supported by the Ministry of Science and Higher Education of the Russian Federation within the framework of state assignment № FEFE-2020-0013.

Благодарности. Работа выполнена при поддержке Министерства науки и высшего образования Российской Федерации в рамках государственного задания № FEFE-2020-0013.

References

1. Orlov V. I., Kazakovtsev L. A., Masich I. S., Stashkov D. V. *Algoritmicheskoe obespechenie podderzhki prinyatiya reshenii po otboru izdelii mikroelektroniki dlya kosmicheskogo priborostroeniya* [Algorithmic support of decision-making on selection of microelectronics products for space industry]. Krasnoyarsk, 2017, 225 p.

2. Kazakovtsev L. A., Antamoshkin A. N. [Greedy heuristic method for location problems]. *Vestnik SibGAU*. 2015, Vol. 16, No. 2, P. 317–325 (In Russ.).

3. MacQueen J. Some methods for classification and analysis of multivariate observations. *Proc. Fifth Berkeley Symp. Math. Stat. Probab*. 1967, Vol. 1, P. 281–297.

4. Hosage C. M., Goodchild M. F. Discrete Space Location-Allocation Solutions from Genetic Algorithms. *Annals of Operations Research*. 1986, Vol 6, P. 35–46. http://doi.org/10.1007/bf02027381

5. Bozkaya B., Zhang J., Erkut E. A Genetic Algorithm for the p-Median Problem, Drezner Z., Hamacher H. (eds,), Facality Location: Applications and Theory, *Springer, Berlin*, 2002.

6. Alp O., Erkut E., Drezner Z. An Efficient Genetic Algorithm for the p-Median Problem. *Annals of Operations Research*. 2003, Vol. 122, P. 21–42. doi: 10.1023/A:1026130003508

7. Maulik U., Bandyopadhyay S. Genetic Algorithm-Based Clustering Technique. *Pattern Recognition*. 2000, Vol. 33, No. 9, P. 1455–1465. doi: 10.1016/S0031-3203(99)00137-5

8. William M. Rand. Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association*. 1971, Vol. 66, No. 336, P. 846–850.

9. De Maesschalck R., Jouan-Rimbaud D., Massart D. L. The Mahalanobis distance. *Chem Intell Lab Syst*. 2000, Vol. 50, No. 1, P. 1–18. doi: 10.1016/S0169-7439(99)00047-7

10. McLachlan G. J. Mahalanobis Distance. *Resonance*. 1999, Vol. 4, No. 20. doi: 10.1007/BF02834632.

11. Xing E. P., Ng A. Y., Jordan M. I., Russel S. Distance metric learning with application to clustering with side-information. *Advances in Neural Information Processing Systems*. 2003, Vol. 15, P. 505–12.

12. Orlov V. I., Fedosov V. V. ERC clustering dataset, 2016. http://levk.info/data1526.zip

13. Orlov V. I., Shkaberina G. Sh., Rozhnov I. P., Stupina A. A., Kazakovtsev L. A. [Application pf clustering algorithm wirh special distance measures for the problem of automatic grouping of electronic and radio devices]. *Control systems and information technologies*. 2019, Vol. 3, No. 3, P. 42–46 (In Russ.).

14. Hansen P., Mladenovic N. J-means: a new local search heuristic for minimum sum of squares clustering. *Pattern recognition*. 2001, Vol. 34, No. 2, P. 405–413.

15. Kazakovtsev L. A., Antamoshkin A. N. Genetic Algorithm with Fast Greedy Heuristic for Clustering and Location Problems. *Informatica*. 2014, Vol. 38, No. 3, P. 229–240.

Библиографические ссылки

1. Алгоритмическое обеспечение поддержки принятия решений по отбору изделий микроэлектроники для космического приборостроения / В. И. Орлов, Л.А. Казаковцев, И.С. Масич и др. ; Сиб. гос. аэрокосмич. ун-т. Красноярск, 2017. 225 с.

2. Казаковцев Л. А., Антамошкин А. Н. Метод жадных эвристик для задач размещения // Вестник СибГАУ. 2015. Т. 16, № 2. С. 317–325.

3. MacQueen J. Some methods for classification and analysis of multivariate observations // Proc. Fifth Berkeley Symp. Math. Stat. Probab. 1967. Vol. 1. P. 281–297.

4. Hosage C. M., Goodchild M. F. Discrete Space Location-Allocation Solutions from Genetic Algorithms // Annals of Operations Research. 1986. Vol 6. P. 35–46. http://doi.org/10.1007/bf02027381

5. Bozkaya B., Zhang J., Erkut E. A Genetic Algorithm for the p-Median Problem // Drezner Z., Hamacher H. (eds,), Facality Location. Applications and Theory, Springer, Berlin, 2002.

6. Alp O., Erkut E., Drezner Z. An Efficient Genetic Algorithm for the p-Median Problem // Annals of Operations Research. 2003. Vol. 122. P. 21–42. http://doi.org/10.1023/A:1026130003508

7. Maulik U., Bandyopadhyay S. Genetic Algorithm-Based Clustering Technique // Pattern Recognition. 2000. Vol. 33, No. 9. P. 1455–1465. https://doi.org/10.1016/S0031-3203(99)00137-5

8. William M. Rand. Objective Criteria for the Evaluation of Clustering Methods // Journal of the American Statistical Association. 1971. Vol. 66, No. 336. P. 846–850.

9. De Maesschalck R., Jouan-Rimbaud D., Massart D. L. The Mahalanobis distance // Chem Intell Lab Syst. 2000. Vol 50, No. 1. P 1–18. doi: 10.1016/S0169-7439(99)00047-7

10. McLachlan G J. Mahalanobis Distance // Resonance. 1999. Vol 4, No. 20. doi: 10.1007/BF02834632.

11. Distance metric learning with application to clustering with side-information / E. P. Xing, A. Y. Ng,

M. I. Jordan and et al. // Advances in Neural Information Processing Systems. 2003. Vol. 15. P. 505–512.

12. Орлов В. И., Федосов В. В. Набор данных электрорадиоизделий. 2016 [Электронный ресурс]. URL: http://levk.info/data1526.zip.

13. Применение алгоритмов кластеризации с особыми мерами расстояния для задачи автоматической группировки электрорадиоизделий / В. И. Орлов, Г. Ш. Шкаберина, И. П. Рожнов и др. // Системы управления и информационные технологии. 2019. № 3 (77). С. 42–46.

14. Hansen P., Mladenovic N. J-means: a new local search heuristic for minimum sum of squares clustering // Pattern recognition. 2001. Vol. 34, No. 2. P. 405–413.

15. Kazakovtsev L. A., Antamoshkin A. N. Genetic Algorithm with Fast Greedy Heuristic for Clustering and Location Problems // Informatica. 2014. Vol. 38, No. 3. P. 229–240.

Shkaberina Guzel Shariphanovna – Associate Professor of the Department of Information and Control System; Reshetnev Siberian State University of Science and Technology. E-mail: z_guzel@mail.ru.

Kazakovtsev Lev Aleksandrovich – Dr. Sc., the Head of the Department of Systems Analysis and Operations Research; Reshetnev Siberian State University of Science and Technology. E-mail: levk@bk.ru.

Li Rui – graduate student of the Department of Systems Analysis and Operations Research; Reshetnev Siberian State University of Science and Technology. E-mail: 646601833@qq.com.

Шкаберина Гузель Шарипжановна – доцент кафедры информационно-управляющих систем; Сибирский государственный университет науки и технологий имени академика М. Ф. Решетнева. E-mail: z_guzel@mail.ru.

Казаковцев Лев Александрович – доктор технических наук, доцент, заведующий кафедрой системного анализа и исследования операций; Сибирский государственный университет науки и технологий имени академика М. Ф. Решетнева. E-mail: levk@bk.ru.

Ли Жуя – аспирант кафедры системного анализа и исследования операций; Сибирский государственный университет науки и технологий имени академика М. Ф. Решетнева. E-mail: 646601833@qq.com.