Detection of homogeneous production batches of semiconductor devices by greedy heuristic clustering algorithms with special distance metrics

G Sh Shkaberina, I P Rozhnov, V P Popov, L A Kazakovtsev and E V Lapunova
Reshetnev Siberian State University of Science and Technology, prosp. Krasnoyarskiy Rabochiy 31, Krasnoyarsk 660031, Russia

E-mail: levk@bk.ru

Abstract. Authors present a comparative efficiency analysis of application of k-means and k-medoids clustering models for solving the problem of grouping of semiconductor devices into homogeneous production batches using three types of metrics: Euclidean distance, Mahalanobis distance, Manhattan distance.

1. Introduction
A modern spacecraft is a complicated electronic design, which, when in space, must be able to comply with the assigned tasks during one to two decades. The spacecraft contains from 100 to 200 thousand electronic parts. The significant cost of the apparatus and the lack of the ability to repair radio and electronic products during a space flight requires the exceptional quality to the electronic component base (ECB).

The shipped batches of the ECB can be heterogeneous and be assembled from several production batches of plates. For this reason, the results of the destructive tests on an ECB sample should be distributed to the entire delivered batch of devices carefully, at least with the conviction that the shipped lot is made of a single batch of raw materials (silicon wafers) or that the variation in the parameters of various crystal batches is small. This is due to the fact that relatively minor changes in the manufacturing process can dramatically affect the sensitivity characteristics. It should be understood from how many homogeneous groups the production batch of electric radio products (ERI) is assembled [1, 2]. If the set of components consists of several different groups, then in order to have a reasoned position about each of them, tests should be carried out for each group, having previously identified them [3].

The idea of this work is to study the effectiveness of the use of greedy heuristic algorithms with special distance measures for tasks of automatic grouping of products [4].

2. Data sets description
The input data represent a certain set of results of testing of the semiconductor devices (current and voltage characteristics of the input and output circuits of the microchips). As an example, 3 samples are considered (certain types of devices): 1526TL1 microcircuit (Sample I), 1526IE10 microcircuit (Sample II) and 3OT122A diodes (Sample III).

Each sample is composed of data on products with a specific set of input parameters, obviously belonging to different homogeneous batches. Sample I consists of three homogeneous batches (the
number of products is 625 pieces); Samples II and III consist of five and three homogeneous batches, respectively (the number of products in sample II is 870, 279 for sample III). Parameters with zero values are excluded from each sample. For analysis, consideration of sample I leaves 4 input parameters, and 6 parameters for sample II and sample III.

3. Greedy heuristic clustering algorithms

One of the most popular clustering algorithms is the EM algorithm used in mathematical statistics to find estimates of the maximum likelihood of the parameters of probabilistic models in the case when the model depends on some hidden variables [2, 5, 6]. Each iteration of the algorithm consists of two steps. At the E-step (expectation), the expected value of the likelihood function is calculated, while the hidden variables are considered observable. At the M-step (maximization), the maximum likelihood estimate is calculated, thus increasing the expected likelihood calculated at the E-step. This value is then used for the E-step in the next iteration. The algorithm runs until \( \varepsilon \)-convergence [6].

This algorithm is unstable according to the initial data, since it finds a local extremum, the value of which can be much lower than the global maximum [2, 5]. Depending on the choice of the initial approximation, the algorithm can converge to different points. The rate of convergence can also vary greatly. In addition to this, the algorithm does not allow determining the number of components of the mixture. This value is a structural parameter of the algorithm [5, 6].

In this paper, we used three methods of cluster analysis: the most commonly used in practice method of k-means (k-means) [7], a modified method of k-VNS [1, 8], and the k-medoid algorithm [9].

**K-means** algorithm. The idea of this simple method is that at each iteration, the center of mass for each cluster obtained in the previous step is calculated, then the vectors are again divided into clusters according to which of the new centers is closer in the selected metric. The algorithm ends when at some iteration there is no change in the intracluster distance.

**K-VNS method** (Variable Neighborhood Search). This method provides, as one of the options for organizing local searches, the use of search algorithms with alternating neighborhoods [10, 11]. It is a metaheuristic method for solving a set of combinatorial optimization and global optimization problems. This algorithm explores remote areas of the current existing solution and moves from there to a new one if improvement has been made [8, 10, 11].

**K-medoids method.** The PAM (Partitioning Around Medoids) algorithm for clustering based on the k-medoids model is one of the classical separation clustering algorithms in which only clustered objects (called medoids) can be selected as cluster centers [4, 12].

Using the classical k-means method and the modified k-VNS method as an example, we show the possibility of the effective use of greedy heuristic algorithms in the problems of automatically grouping the semiconductor devices into homogeneous batches.

The number of clusters was selected depending on the actual number of homogeneous batches in the mixture.

For each algorithm tested, 30 runs were performed for each sample. All the results achieved by each of the algorithms in each attempt were recorded, then the minimum and maximum values of the objective function for each data sample were determined among the results, as well as the average value of the objective function (table 1).

The minimum value of the standard deviation of the objective function shows that the presented algorithms of the greedy heuristic methods show high accuracy and stability of solutions. In this case, the classical k-means algorithm is inferior to the modified k-VNS algorithm, which in most startup cases returned the matching values of the objective function.

Detailed comparative efficiency analysis of k-VNS algorithm for the k-means problem was also given in [8].

| Algorithm         | Objective function value (total intercluster distance square) |
|-------------------|--------------------------------------------------------------|
|                   | M    | Max   | Avg  | Std. deviation |

Table 1. Results of the k-means and k-VNS algorithms.
**4. Various distance metrics**

The k-means and k-medoid methods were applied with a different set of informative features for each sample. For each method, 30 test runs of the program were performed using three types of metrics: Euclidean distance [13], Mahalanobis distance, Manhattan distance.

The calculation of the Euclidean metric is calculated as follows:

\[ d_{pq} = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}, \]

where \( p, q \) are \( n \)-dimensional data vectors, \( n \) is number of device parameters.

For Mahalanobis distances, the equation is

\[ d_{M}(x) = \sqrt{((x - \mu)^T S^{-1}(x - \mu)}, \]

where \( S \) is the covariance matrix of the sample. If \( S=E \) then this distance is equal to the Euclidean distance. If \( S \) is a diagonal matrix then such a distance is called normalized Euclidean distance.

Manhattan distance is defined as

\[ d(p, q) = \sum_{i=1}^{n} |p_i - q_i|. \]

If the Euclidean distance clustering is applied to the Sample I, the clustering accuracy for k-means is 82.4-86.4%, and 80.96-85.12% for k-medoids. For the Manhattan distance, k-means accuracy is 96.96-98.4%, and 97.12-97.6% for k-medoids. However, for Mahalanobis distances with k-means, the accuracy is much higher, 96.64-98.4%, and 97.28-97.76% for k-medoids (table 2). The results for Samples II and III are similar (tables 3 and 4).

| Sample I (3 clusters) | k-means in multistart mode | k-VNS | S | k-means in multistart mode | k-VNS | S |
|------------------------|----------------------------|-------|---|---------------------------|-------|---|
|                        | 581.484                    | 592.637 | 583.637 | 7.252         | 581.484 | 581.484 | 0 | 1135.702 | 1151.079 | 1143.391 |
|                        | 1101.165                   |        |     |                           |       |     |

**Table 2.** Clustering results for sample I (1526TL1, 625 pieces).

| Clustering model (greedy heuristic algorithm applied) | Distance metric | Number of the informative parameters |
|-------------------------------------------------------|-----------------|--------------------------------------|
|                                                       | 2               | 3         | 4               | 2   | 3   | 4   |
|                                                       | Correctly clustered, pieces. | Correctly clustered, % | |
| k-means                                               | Euclidean       | 540       | 527            | 515 | 86.4 | 84.32 | 82.4 |
|                                                       | Manhattan       | 615       | 613            | 606 | 98.4 | 98.08 | 96.96 |
|                                                       | Mahalanobis     | 615       | 612            | 604 | 98.4 | 97.92 | 96.64 |
| k-medoids                                             | Euclidean       | 532       | 513            | 506 | 85.12| 82.08 | 80.96 |
|                                                       | Manhattan       | 610       | 608            | 607 | 97.6 | 97.28 | 97.12 |
|                                                       | Mahalanobis     | 611       | 610            | 608 | 97.76| 97.6  | 97.28 |

**Table 3.** Clustering results for sample II (1526IE10, 870 pieces)

| Clustering model (greedy heuristic) | Distance metric | Number of the informative parameters |
|-------------------------------------|-----------------|--------------------------------------|
|                                     | 2               | 4         | 6               | 2   | 4   | 6   |
|                                     | Correctly clustered, pieces. | Correctly clustered, % | |
Analysis of the results of clustering showed that with an increase in the number of informative features, the accuracy of clustering does not increase. It was also found that regardless of clustering algorithms, using the Mahalanobis distance metric, the clustering accuracy is maximum. For our problems, application of various classification methods can result in much better accuracy. However, problems such as clustering the semiconductor devices into homogeneous production batches [1], in real production conditions, we have no training sample needed for such methods.

5. Conclusions
The article explores the possibilities of using greedy heuristic algorithms for the tasks of separating a mixture of distributions with special distance measures. The possibility of using special distance measures in the tasks of automatic grouping of industrial products with a small amount of input data was also considered. Compared different measures of distances. It was found that the Mahalanobis distance metric increases the accuracy of product separation regardless of clustering algorithms (k-means, k-medoid).

References
[1] 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
[2] Kazakovtsev L A et al. 2018 Development of algorithmic support for the analysis of the homogeneity of batches of electronic radio products for the integration of electronic equipment for spacecraft. Krasnoyarsk, Reshetnev University
[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 Conference Series: Materials Science and Engineering 537 022035
[4] Kazakovtsev L A, Stupina A A 2014 Fast genetic algorithm with greedy heuristic for p-median and k-means problems 2014 International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT’2014) pp 602-6
[5] Kazakovtsev L, Stashkov D, Gudyma M and Kazakovtsev L 2019 Algorithms with Greedy
Heuristic Procedures for Mixture Probability Distribution Separation. *Yugoslav Journal of Operations Research* **29**(1) 51-67

[6] Korolev V Yu 2007 EM algorithm, its modifications and their application to the probability distribution separation problem. Moscow, IPI RAN

[7] Cheung Y M 2003 k-Means: A new generalized k-means clustering algorithm. *Pattern Recognition Letters* **24**(15) 2883-93

[8] 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

[9] Zahang Q, Couloigner I 2005 A New and Efficient K-Medoid Algorithm for Spatial Clustering. *Lecture Notes in Computer Science* **3482** 181-9

[10] Hansen P and Mladenovic N 2001 Variable neighborhood search: principles and applications. *Eur. J. Oper. Res.* **130** 449–67

[11] Mladenovic N and Hansen P 1997 Variable neighborhood search. *Comput. Oper. Res.* **24** 1097-100

[12] Ovelgonne M 2011 *Scalable Algorithms for Community Detection in Very Large Graphs* Karlsruhe

[13] Deza M M and Deza E 2009 *Encyclopedia of Distances* (Springer)