Recursive clustering algorithm based on silhouette criterion maximization for sorting semiconductor devices by homogeneous batches

S M Golovanov¹, V I Orlov¹, L A Kazakovtsev¹² and A M Popov¹

¹Reshetnev Siberian State University of Science and Technology
31, Krasnoyarskiy Rabochyi av., Krasnoyarsk, 660037, Russian Federation
²Krasnoyarsk State Agrarian University, 90, Mira av., Krasnoyarsk, 660049, Russian Federation

E-mail: levk@bk.ru

Abstract. We propose a recursive algorithm for dividing a presumably mixed production batch of semiconductor devices into homogeneous groups (clusters) with the use of data of non-destructive testing, based on the maximization of the silhouette criterion in the sequential solving of the k-Means clustering problems. Based on the silhouette criterion, we formulate the concept of homogeneity of a production batch. The application of the developed algorithm to the problem of formation of homogeneous production batches of integrated circuits for space industry is illustrated with a computational example.

1. Introduction

The problem of interpreting data obtained from the results of testing a production batch of similar items (objects) arises in many areas of technology. As a rule, the purpose of testing is to determine whether the tested parameters belong to the specified ranges. However, besides this main purpose, the test results obviously contain useful information about the batch as a whole. If the test results can be used to divide a batch of items into several homogeneous groups, then we can conclude that the resulting division reflects some deep-seated phenomena: the use of different groups of raw materials, different manufacturing technologies, different conditions of transportation and storage. Thus, there is a possibility that the resulting groups of items will behave differently in the future when operating [1-4]. This fact must be identified and taken into account.

Authors [5] propose various clustering models and algorithms for solving the problem of partitioning a presumably mixed batch of semiconductor devices into homogeneous groups (clusters). The simplest algorithms such as k-means [6, 7] are quite efficient. Unfortunately, a universal approach to solving this problem, due to its complexity, does not exist yet. As a rule, all the proposed methods have various disadvantages. In particular, simple parametric methods such as k-means, k-medoid, EM-algorithm [8, 9] require specifying the number of groups (clusters). In this case, other methods based on special criteria must be used for determining the number of groups. Algorithms based on density estimation are limited by data dimensionality [10]. Algorithms like X-means [11], which are capable of simultaneously
determining the number of groups, have a lower accuracy. Moreover, they still use a special criterion to determine the number of groups. Finally, in [5], authors propose efficient algorithms for the simultaneous solving of a series of k-means or similar problems, differing only in the number of groups (clusters). However, these algorithms also require post-processing of results with the use of a certain criterion.

This article proposes an approach that was developed as a result of the implementation of the semiconductor device clustering in the process of their rejection testing for space applications in a specialized testing center.

2. Problem Statement, Homogeneity Criterion

Suppose that we have a batch \( P \) of \( N \) items \( E_i, i = 1, N_e \), which passed \( N \) tests \( T_j, j = 1, N_t \). Thus, according to the test results, we have a matrix of test results \( M_{ij} \) of dimension \( N \times N \). Test results \( T_j \) can have completely different physical nature and units, for example, voltage (V), current strength (A), gain factors (dB) etc.

In order for the test results to be suitable for applying a clustering model based on distances between objects, such as k-Means, we have to normalize the \( M_{ij} \) matrix to obtain the \( M_{ijN} \) matrix with all results converted into dimensionless values. The selection of the appropriate normalization method [12] for a particular case is a problem requiring a separate in-depth study.

To solve the clustering problem, researchers introduce various auxiliary criteria for determining the number of clusters in data [13-16]. In this paper, we use the fast silhouette criterion (Kaufman, Rousseeuw) [16, 17]. The choice of this criterion is determined by the results of a comparative analysis of many known criteria for estimating the number of clusters [18] in the problem of dividing the mixed batch of semiconductor devices into homogeneous batches. In the future, for convenience, the fast version of the silhouette criterion will be referred as silhouette criterion (\( S \)).

In terms of semiconductor device clustering, the definition of the criterion \( S \) is as follows. Let the mixed batch \( P \) consisting of objects \( E_i, i = 1, N_e \), be divided into \( m \) clusters \( G_j, j = 1, m \) with centers (centroids, medoids etc. depending on the clustering model) \( C_j(j = 1, m) \). Then the silhouette criterion \( S(P) \) is defined as follows:

\[
S(P) = \frac{\sum_{i=1}^{N_e} S(i)}{N_e}
\]

where \( S(i) \) is the silhouette of object \( E_i, i = 1, N_e \), calculated as:

\[
S(i) = 1 - \frac{a(i)}{b(i)}, \quad i = 1, N_e.
\]

Here, \( a(i) \) is distance (depends on the selected distance metric) from object \( E_i \) to the center of cluster \( G_j \) \( (j = 1, m) \), which contains object \( E_i \), \( b(i) \) is distance from object \( E_i \) to the nearest center of other cluster \( G_k \).

The problem of clustering a batch \( P \) of objects will be understood as the problem of sorting objects \( E_i \) into homogeneous groups. Below, we introduce the concept of homogeneity. The homogeneity criterion is formulated as follows.

**Definition:**
A batch \( P \) of objects \( E_i, i = 1, N_e \) is considered *homogeneous* if it is impossible to divide \( P \) into \( m \) groups \( G_j, j = 1, m \), which meet the following conditions:

\[
\begin{cases}
S(P) > S_{\text{max}}, \\
R_{ij} > R_{\text{min}} \quad \forall i, j = 1, m, j \neq i.
\end{cases}
\]

Here, \( R_{ij} \) is the distance between cluster centers \( G_i \) and \( G_j \), \( S_{\text{max}} \) is the threshold of the silhouette criterion \( S(P) \), which allows us to consider batch \( P \) of objects \( E_i \) as homogeneous, \( R_{\text{min}} \) is the minimum distance between cluster centers \( G_i \) and \( G_j \), considered as significant.
Parameter $R_{\text{min}}$ is related to the physical properties of the batch $P$ and determines the minimum distinguishable distance between the objects of a homogeneous batch. In addition, it is obvious that the value of the parameter $R_{\text{min}}$ depends on the normalization method.

Parameter $S_{\text{max}}$ defines the boundaries separating homogeneous batches from heterogeneous ones. Its value is determined experimentally.

3. New Algorithm

Based on the introduced homogeneity definition, we propose a recursive algorithm for maximizing the silhouette criterion (RAmaxS) for solving the clustering problem.

**Algorithm 1.** Recursive algorithm for maximizing the silhouette criterion (RAmaxS)

**Required:** batch $P$.

The $k$th recursion level:

1. The algorithm splits the batch $P$ of elements $E_i$, $i = 1, N_e$, into clusters $G_j$, $j = 1, m$ with various numbers of clusters $m = 1, m_{\text{max}}$. In practice, the value of $m_{\text{max}} = 10$ is sufficient. For clustering, we use the k-means model with the sum of the squared Euclidean distances as the objective function. The algorithm then chooses a solution with such a number of clusters $m$ that $S(P)$ reaches its maximum.

2. If condition (3) is satisfied, then batch $P$ and the resulting clusters $G_j$, $j = 1, m$ are considered potentially heterogeneous. Each of the received clusters $G_j$ is considered as a separate data set (separate batch $P$), and this algorithm is started for each of such clusters at the next recursion level.

3. If condition (3) is not satisfied, then the batch $P$ of objects $E_i$, $i = 1, N_e$ is considered homogeneous, and the process stops.

The result of the algorithm is a set of homogeneous batches (final clusters).

4. Experimental Results

The application of the proposed recursive algorithm is illustrated with an example of a mixed production batch $P$ of the integrated circuits 140UD25AC1BK, 160 units, which is composed of three obviously homogeneous batches: batch $P_C$ of 45 microcircuits, $P_D$ of 57 microcircuits, and $P_E$ of 58 microcircuits with different release dates, tested (6 non-destructive tests) in a specialized testing centre.

Parameter $R_{\text{min}}$ is determined from the condition that during the rejection testing in accordance with the methods of the specialized testing centre, the minimum distinguishable distance for each test is the parameter drift range introduced to estimate the maximum allowable shift of the value of the tested parameter in the process of electrical thermal training (ETT). Taking into account the method of data normalization, we obtain $R_{\text{min}} = 1$.

Parameter $S_{\text{max}}$ was determined experimentally. After conducting a series of experiments, we have $S_{\text{max}} = 0.5$.

For optimal division of a mixed batch into clusters, which maximizes the silhouette criterion, we use the methods described in [18, 19].

Let us apply Algorithm 1 to the mixed batch $P$ of 140UD25AC1BK microcircuits.

1st recursion level:

After searching for the optimal division (in the sense of maximum value of $S(P)$), our mixed batch $P$ is divided into two groups: $G_1$ (58 units), which contains only objects of the $P_E$ batch, and $G_2$ (102 units), consisting of the items of the $P_D$ and $P_C$ batches. Hereafter: $S(P) = 0.754$ and $R(G_1, G_2) = 2.591$. To visualize the result, figure 1 shows the mapping of the objects of the mixed batch $P$ with the use of the multidimensional scaling method (MDS) [20]. Since the condition (3) for the groups $G_1$ and $G_2$ is satisfied, both of these groups (clusters) are potentially heterogeneous, and both clusters go to the 2nd level of recursion. Note that the result obtained at this stage (splitting into 2 groups) does not correspond either to the actual composition of the mixed batch or to the previously proposed [18] partitioning model based on a single application of the silhouette criterion to the clustering results.
Figure 1. Multidimensional scaling of the objects of the mixed batch $P$.

2nd recursion level:

a) Search for an optimal division of the group $G_1$ in the sense of maximum of $S(G_1)$.
We get the following result: $G_1$ (58 units) is divided into three subgroups: $G_{11}$ (32 units), $G_{12}$ (15 units), and $G_{13}$ (11 units). Herewith: $S(G_1) = 0.453$ and $0.493 \leq R(G_{1i}, G_{1j}) \leq 0.585$ ($i, j = 1, 3$). Since the condition (3) for $G_1$ is satisfied, the group of objects $G_1$ (58 units), which are items of the $P_E$ batch, is considered homogeneous (Fig. 2, left side).

b) Search for an optimal division of the cluster $G_2$ in the sense of maximum of $S(G_2)$.
We have the following result: $G_2$ (102 units) is divided into two subgroups, $G_{21}$ (57 units), which are items of the batch $P_D$, and $G_{22}$ (45 units), which are integrated circuits of the batch $P_C$. In this case: $S(G_2) = 0.582$, $R(G_{21}, G_{22}) = 1.201$ (Fig. 2, right side). Since the condition (3) for the groups $G_{21}$ and $G_{22}$ is satisfied, both of these groups (clusters) are potentially heterogeneous, and both of them go to the 3rd level of recursion.

3rd recursion level:

a) Search for an optimal division of the cluster $G_{21}$ in the sense of maximum of $S(G_{21})$.
We have the following result: $G_{21}$ (57 units) is divided into three subgroups, $G_{211}$ (20 units), $G_{212}$ (20 units), and $G_{213}$ (17 units), with $S(G_{21}) = 0.485$, and $0.553 \leq R(G_{21i}, G_{21j}) \leq 0.644$ ($i, j = 1, 3$). Since the condition (3) for $G_{21}$ is not satisfied, the cluster $G_{21}$ (57 units which are objects of the batch $P_D$) is homogeneous (Fig. 3, left side).

b) We have the following result: $G_{22}$ (45 units) is divided into four subgroups, $G_{221}$ (27 units), $G_{222}$ (5 units), $G_{223}$ (7 units), and $G_{224}$ (6 units), with $S(G_{22}) = 0.494$, and $0.978 \leq R(G_{22i}, G_{22j}) \leq 1.959$ ($i, j = 1, 3$). Since the condition (3) for $G_{22}$ is not satisfied, the cluster $G_{22}$ (45 units which are objects of the batch $P_C$) is homogeneous (Fig. 3, right side).

Thus, using the proposed RAmaxS algorithm, we divided the mixed batch $P$ of semiconductor devices (160 units) into three homogeneous groups. In our example, the resulting cluster $G_1$ (58 units) is equivalent to a homogeneous batch $P_E$, which was mixed into the batch $P$, cluster $G_{21}$ (57 units) is equivalent to the homogeneous batch $P_D$, and cluster $G_{22}$ (45 units) corresponds to the homogeneous batch $P_C$. 


**Figure 2.** Multidimensional scaling of the objects of cluster $G_1$ (left) and $G_2$ (right).

**Figure 3.** Multidimensional scaling of the objects of $G_{21}$ (left) and $G_{22}$ (right).

**Conclusion**

The proposed definition of a homogeneous batch of semiconductor devices, as well as the new recursive algorithm for selecting homogeneous batches based on the silhouette criterion, allows us to split a mixed batch of semiconductor devices into homogeneous batches with high accuracy (in the example given, 100% accurate) and determine the number of these homogeneous batches. The results are confirmed by a computational experiment.

**References**

[1] Ooi M P-L et al 2011 Getting more from the semiconductor test: Data mining with defect-cluster extraction *IEEE Trans. Instrum. Meas* **60**(10) 3300-3317

[2] Ooi M P-L et al 2013 Identifying Systematic Failures on Semiconductor Wafers Using ADCAS Design &Test, *IEEE 30*(5) 44-53

[3] Khamidullina N M et al 1999 Predictions of integrated circuit serviceability in space radiation fields *RadiationMeasurements* **30** 633-8

[4] Fedosov V V and Orlov V I 2011 Minimal necessary extent of examination of microelectronic products at inspection test stage *Izvestiya Vuzov Priborostroenie* **54**(4) 62-8 (in Russian)

[5] Kazakovtsev L A, Antamoshkin A N and Fedosov V V 2016 Greedy heuristic algorithm for
solving series of EEE components classification problems IOP Conference Series: Materials Science and Engineering 122 Article ID 012011. DOI: 10.1088/1757-899X/122/1/012011

[6] Wu X et al 2008 Top 10 algorithms in data mining Knowl Inf Syst 14 1–37 DOI 10.1007/s10115-007-0114-2

[7] Kazakovtsev L A, Antamoshkin A N and Fedosov V V 2016 Greedy heuristic algorithm for solving series of EEE components classification problems IOP Conference Series: Materials Science and Engineering 122 Article ID 012011. DOI: 10.1088/1757-899X/122/1/012011

[8] Hastie T and Tibshirani F 2001 The EM algorithm The Elements of Statistical Learning New York: Springer 236-43

[9] Orlov V I, Stashkov D V, Kazakovtsev L A and Stupina A A 2016 Fuzzy clustering of EEE components for space industry IOP Conference Series: Materials Science and Engineering 155 012026. DOI: 10.1088/1757-899X/155/1/012026

[10] Ester M, Kriegel H-P, Sander J and Xu X 1996 A density-based algorithm for discovering clusters in large spatial databases with noise Proceedings of the Second International Conference on Knowledge Discovery and Data Mining KDD-96 226-31

[11] Pelleg D and Moore A 2000 X-Means: Extending K-Means with Efficient Estimation of the Number of Clusters Proceedings of 17th International Conference on Machine Learning, Stanford, CA, 29 June-2 July 2000 727-734

[12] Fedosov V V, Kazakovtsev L A and Gudyma M N 2016 Problem of Normalization of Input Test Data of Space Industry EEE Components for Automatic Grouping Algorithm Informatsionnye tekhnologii modelirovaniya i upravleniya 4 263-8 (in Russian)

[13] Schwarz G E 1978 Estimating the dimension of a model Annals of Statistics 6(2) 461-4 DOI:10.1214/aos/1176344136

[14] Bhat H S and Kumar N 2010 On the derivation of the Bayesian Information Criterion School of Natural Sciences, University of California: Oakland, USA

[15] Akaike H 1974 A new look at the statistical model identification IEEE Transactions on Automatic Control 19(6) 716-3 DOI:10.1109/TAC.1974.1100705

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

[17] Kaufman L and Rousseeuw P J 1990 Finding groups in data: an introduction to cluster analysis (New York: Wiley)

[18] Orlov V I, Kazakovtsev L A and Masich I S 2016 Silhouette Criterion for Automatic Grouping Algorithm of Spaceship Electronic Components. Vestnik SibGAU 17(4) 883-90 (in Russian)

[19] Kazakovtsev L A, Orlov V I, Stashkov D V, Antamoshkin A N and Masich I S 2017 Improved model for detection of homogeneous production batches of electronic components IOP Conf. Series: Materials Science and Engineering 255 012004 DOI:10.1088/1757-899X/255/1/012004

[20] Borg J F P 2005 Modern Multidimensional Scaling: Theory and Application (Springer)