Sequential agglomerative procedure for sorting a production batch of electronic radio devices into homogeneous groups

IOP Conf. Series: Materials Science and Engineering **734** (2020) 012013 doi:10.1088/1757-899X/734/1/012013

Sequential agglomerative procedure for sorting a production batch of electronic radio devices into homogeneous groups

S M Golovanov[1,2], V I Orlov[1,2], L A Kazakovtsev[2] and V V Fedosov[1,2]

[1] Testing Technical Center - NPO PM, 20, Molodezhnaia st., Zheleznogorsk, 662970, Russia
[2] Reshetnev Siberian State University of Science and Technology 31, Krasnoyarskiy Rabochiy av., Krasnoyarsk, 660037, Russia

E-mail: itcnpopm@atomlink.ru

Abstract. Authors proposed a clustering algorithm for sorting a batch of electronic radio devices by homogeneous groups according to the results of non-destructive tests. Algorithm is based on searching for the best silhouette criterion value of clusters using the sequential agglomerative procedure (SAP). It merges groups with the smallest distance between their centers one by one. New procedure was tested on dataset of microcircuits for space applications in a specialized test center.

1. Introduction

A data interpretation problem arises in many branches of technical sciences where data are obtained by testing of homogenous elements. If according to the test results, a batch of elements can be divided into homogeneous groups, it allows us to conclude that the resulting division caused by some unknown to us reasons: different groups of raw materials, different processing technologies, different conditions of transportation and storage. Thus, obtained groups of elements probably have different characteristics [1, 2, 3] and it is necessary to identify and consider this fact.

Various clustering algorithms are used for solving the problem of dividing electronic radio products into homogenous groups, and k-means algorithm is one of the simplest and the most efficient [4], however, it also has some significant disadvantages: it requires constant number of clusters and it significantly depends on initial conditions. Today, a universal approach to solve the clustering problem does not exist due to its proved complexity and every existing method have some issues and disadvantages. Nowadays, approaches based on the use of specially introduced auxiliary criterion [5, 6, 7] are widespread. Also, the practice of solving the electronic radio products sorting problem shows that the Silhouette criterion introduced by Kaufman, Rousseeuw [8] is the most convenient and useful. An approach proposed in [9] is an example of using the Silhouette criterion. In that paper authors used genetic algorithms with various heuristics [10, 11] to find optimal by silhouette criterion solution.

In this paper, we propose an approach also based on the search for the best option that provides the maximum value of the silhouette criterion. However, unlike the approach [9, 10, 11], we use a new procedure based on merging of groups with closest centers instead of agglometative heuristics in genetic algorithms. The main advantage of this approach is its speed. This is especially important for solving problems of large dimensions.

Approaches using an ensemble of algorithms [12, 13] popularity grows at present, instead of algorithms which includes only one kind of algorithms. It is obvious that with the number of algorithms in ensemble effectiveness of whole method also grows.

In this paper, we propose the approach, created as a result of working on clustering and testing of ERP for space purposes problem in company "Testing Technical Center - NPO PM", city of Zheleznogorsk, Russia

2. Problem statement, homogeneity criterion

Let P be a batch of N_e elements E_i , $i = \overline{1, N_e}$, and P is tested on N_t tests T_j, $j = \overline{1, N_t}$. Thus, after all tests, every element E_i in batch P has vector of testing results $V_{Ei}$of dimension N_t. T_j test results can have completely different physical nature and units: voltage (V), amperage (A), amplification factors (dB or dimensionless values) etc.To make testing results suitable for use in a clustering model based on distances between elements, it is necessary to normalize vectors V_{Ei} and convert all values to dimensionless quantities. As a result, for each element $Ne(i = \overline{1, N_e})$ we will obtain a normalized vector of test results V_{Ei}^n.The selection of an appropriate standardization method for a every single case is a separate task.

In this paper, we use the boosted Silhouette criterion used (Kaufman, Rousseeuw) [8] to estimate the number of clusters (groups). Let us denote the Silhouette criterion as S.

Let us determine the value of S. Let batch P of elements E_i ($i = \overline{1, N_e}$) be divided into m groups G_j $j = \overline{1, m}$, with centers (or centroids/medoids depending on the statement of the problem) C_j, $j = \overline{1, m}$. Then the Silhouette criterion $S(P)$ is determined as

$$S(P) = (\sum_{i=1}^{N_e} S(i))/N_e \quad (1)$$

where $S(i)$ is a silhouette of the element E_i ($i = \overline{1, N_e}$) determined as

$$S(i) = 1 - \frac{a(i)}{b(i)} \quad (2)$$

Here, $a(i)$ is a distance from element E_i ($i = \overline{1, N_e}$) to the center of its group G_j, $j = \overline{1, m}$, in chosen metric, $b(i)$ is the distance from element E_i ($i = \overline{1, N_e}$) to the center of the other nearest group G_n in chosen metric.

We will understand the problem of clustering a batch of elements P as the task of sorting the elements E_i ($i=1,N_e$) into homogeneous groups G_j ($j = \overline{1, m}$), providing the maximum value of $S(P)$.

3. Sequential Agglomerative Procedure

Our Sequential agglomerative procedure (SAP) is based on the popular bottleneck clustering algorithm [14, 15, 16] and the ideas of hierarchical clustering [17].

Let P be a batch of elements E_i, $i = \overline{1, N_e}$, and P is tested on N_t tests T_j, $j = \overline{1, N_t}$, and for every element E_i there are normalized vector of results V_{Ei} of dimension N_t. Sequential agglomerative procedure (SAP) merges data vectors into N_g homogenous groups and works as follows:

Algorithm 1. Sequential Agglomerative Procedure (SAP)

1) On the first iteration, the number of homogenous groups of the batch P is equal to the number N_e of elements E_i in the batch P, so every element E_i is in its own group $G_{1i}(i = \overline{1, N_e})$. Thus, vector V_{Ei} is a centerC_{1i} of groupG_{1i}, $i = \overline{1, N_e}$.

2) On the k^{th} iteration, we choose two groups from obtained on (k-1) iteration N_e-k+2 groups: G_i^{k-1} (N_i^{k-1}elements: E_1^i,E_2^i ,…,$E_{N_i^{k-1}}^i$) and G_j^{k-1} (N_j^{k-1} elements: E_1^j, E_2^j , …, $E_{N_j^{k-1}}^i$) with minimal distance between their centers C_i^{k-1}иC_j^{k-1}. Then we merge them into $G_m^k = G_i^{k-1} \cup G_j^{k-1}$. New center $C_{G_m^k}^k$ of group G_m^k is defined as:

$$C_{G_m^k}^k = (\sum_{l=1}^{N_i^{k-1}} V_{E_l^i} + \sum_{l=1}^{N_j^{k-1}} V_{E_l^j})/(N_i^{k-1} + N_j^{k-1}) \quad (3)$$

Note 1*:* Distances between centers C_i^{k-1} and C_j^{k-1} and addition of vectors $V_{E_l^i}$ and $V_{E_l^j}$ defines according to chosen metric.

Note 2: Numbers of homogenous groups formed on k^{th} iteration is N_e-k-1.

3) Repeat steps 1 and 2 while the number of obtained groups is not equal to a given number N_g.

4. Practical application of the SAP

Let P be a production batch of N_e elements $E_i, i = \overline{1, N_e}$, tested on N_t tests T_j, $j = \overline{1, N_t}$. For each element E_i, we form normalized results vector V_{Ei}^n of dimensionality N_t. Let us also have two additional limitations: maximum of homogenous groups in batch P is N_G^{max} and minimal number of elements in one group N_{el}^{min}.

Note 3: Number N_G^{max} is selected considering a priori knowledge of the party in question.

Note 4: According to the test results, single outlier elements (OE) may appear as a one-element homogeneous groups. Outlier elements can be caused both by the physical feature of a given element, and by distortions introduced by measuring devices. When sorting the elements of a batch into homogeneous groups, it is necessary to identify the OE with deeper analysis further. To separate OE from other batch elements we introduced number N_{el}^{min}. In practice, it is often accepted: N_{el}^{min} = 3.

The algorithm for sorting batch elements into homogeneous groups using SAP works as follows:

1) The elements E_i, $i = \overline{1, N_e}$, of the batch P are sorted by a given number of homogeneous groups N_g using SAP.The number of "significant" formed groups is determined as N_g^s the number of elements of which is not less than N_{el}^{min}.

а) If N_g^s< 2 then sorting result is not taken into final result.

б) If N_g^s>= 2 then we perform a new sorting of the elements of the batch E_i by the k-means algorithm on N_g^s homogeneous groups with the centers of the "significant" groups as initial values. Result of sorting and its Silhouette criterion $S(P)$ are memorized.

2) Algorithm 1 is sequentially repeated with N_g=2, 3, … until one of two conditions is satisfied: $N_g^s > N_G^{max}$ or $K_{sg} > K_{sg}^{min}$.

The parameter K_{sg} is introduced to separate the initial phase of the SAP from the final phase: $K_{sg} = N_{el}^{sg}/N_e$. The parameter K_{sg}^{min} is selected when the algorithm is tuned to a specific type of electronic radio products. Usually K_{sg}^{min}= 0,5.

3) A solution with the maximum value of $S(P)$ is chosen as the optimal one for our problem. If the algorithm is not able to generate a single solution with N_g^s>= 2, then the whole batch of P of elements E_i, $i = \overline{1, N_e}$, will be considered as homogeneous.

5. Computational experiment

We illustrate the application of the proposed algorithm by the example of a combined batch P of microcircuits 140UD25AS1VK (183 products), composed of three obviously uniform batches: batch $P_C$of 53 microcircuits, batch P_D of 70 microcircuits and batch P_E of 60 microcircuits released in different time and tested for 16 various parameters in “Testing Technical Center - NPO PM”. As a result of testing of each microcircuit M_i, $i = \overline{1, 183}$, of combined batch P a normalized vector of test results $V_{M_i}^n$ of dimension 16 was formed. Thus, batch P corresponds to a set of points in 16-dimensional space. Applying the transformation to two-dimensional space to this set by multidimensional scaling we obtain the graph shown in Figure 1.

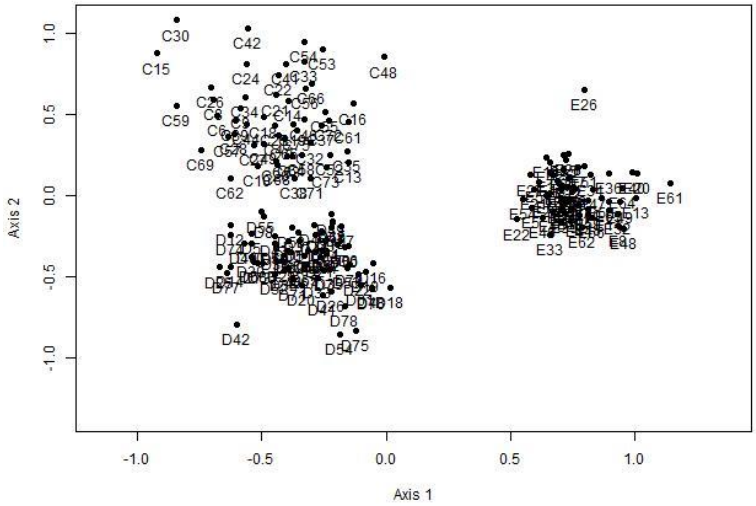

Figure 1. MDS visualization of the mixed production batch.

The results presented in Fig. 1 confirm that the combined batch P of microcircuits 140UD25AS1VK is composed of three uniform batches: P_C, P_D, и P_E.

Let us apply the proposed algorithm to the given batch with the following parameter values: N_e=183, N_t=16, N_{el}^{min}=3, N_G^{max}=10, K_{sg}^{min}=0,5. The result of our experiment is shown in Table 1 and Fig. 2 where the dependence of the Silhouette criterion $S(P)$ on the number N_g of homogeneous groups generated by the algorithm is presented: $G_1^{N_g}$, $G_2^{N_g}$, …, $G_{N_g}^{N_g}$.

Table 1. Values of the Silhouette criterion.

$P=P_C \cup P_D \cup P_E$	$\bigcup_{i=1}^{2} G_i^2$	$\bigcup_{i=1}^{3} G_i^3$	$\bigcup_{i=1}^{4} G_i^4$	$\bigcup_{i=1}^{5} G_i^5$	$\bigcup_{i=1}^{6} G_i^6$	$\bigcup_{i=1}^{7} G_i^7$	$\bigcup_{i=1}^{8} G_i^8$	$\bigcup_{i=1}^{9} G_i^9$	$\bigcup_{i=1}^{10} G_i^{10}$
N_g	2	3	4	5	6	7	8	9	10
$S(P)$	0,548	0,563	0,515	0,495	0,453	0,464	0,429	0,439	0,445

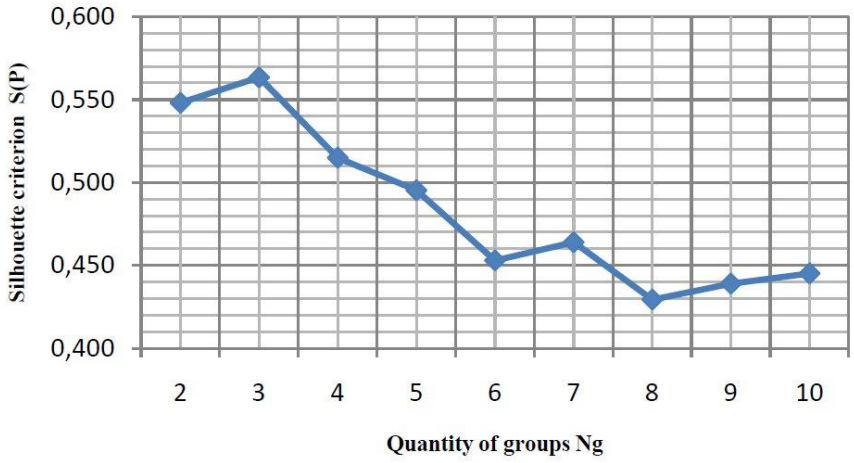

Figure 2. Dependence of Silhouette criterion on the number of groups.

Thus, we obtain:
G_1^2=P_C∪P_D, G_2^2=P_E;
G_1^3=P_C , G_2^3= P_D , G_3^3= P_E;
G_1^4=P_C^{41},G_2^{42}= P_C^2 (P_C^{41}∪P_C^{42}=P_C),G_3^4= P_D , G_4^3= P_E;
G_1^5=P_C^{51}, G_2^5=P_C^{52}, G_3^5=P_C^{53} (P_C^{51}∪P_C^{52}∪G_3^5=P_C), G_4^5= P_D , G_5^5= P_E etc.

According to Table 1 and Fig.2, the optimal solution (i.e. solution with maximum of the Silhouette criterion *Smax*(*P*)= 0,563) of the microcircuits sorting problem is achieved by setting number of groups N_g=3. According to this solution, G_1^3=P_C , G_2^3= P_D , G_3^3= P_E and it corresponds to a given distribution.

6. Conclusions

The proposed approach to sorting an ERI batch into homogeneous groups using the sequential agglomerative procedure (SAP) allows us to split an ERP batch into homogeneous production groups with determination of the number of these groups with high accuracy (100% in the given example). Our computational experiment proved the results.

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)** 3300-17
[2] 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
[3] 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
[4] 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** 012011
[5] Bhat H S and Kumar N 2010 *On the derivation of the Bayesian Information Criterion* (School of Natural Sciences, University of California: Oakland, USA)
[6] Akaike H 1974 A new look at the statistical model identification *IEEE Transactions on Automatic Control* **19(6)** 716-723
[7] Kaufman L and Rousseeuw P J 1990 *Finding groups in data: an introduction to cluster analysis.* (New York: Wiley)
[8] Rousseeuw P 1987 Silhouettes: a graphical aid to the interpretation and validation of cluster analysis *Journal of Computational and Applied Mathematics* **20** 53-65
[9] 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
[10] Kazakovtsev L A and Stupina A A 2015Fast genetic algorithm with greedy heuristic for p-median and k-means problems *2014 International Congress on Ultra Modern Telecommunications and Control Systems and Workshops* 602-6
[11] 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
[12] 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** 78-87
[13] Rozhnov I, Orlov V and Kazakovtsev L 2018 Ensembles of clustering algorithms for problem of detection of homogeneous production batches of semiconductor devices *CEUR-WS* **2098** 338-48

[14] Sun Zh. *et al* 2014 A parallel clustering method combined information bottleneck theory and centroid-based clustering *The Journal of Supercomputing* **69(1)** 452-67
[15] Tishby N, Pereira F C and Bialek W 1999 The Information Bottleneck Method *The 37th annual Allert on Conference on Communication, Control, and Computing* 368–77
[16] Harremoes P and Tishby N 2007 The Information Bottleneck Revisited or How to Choose a Good Distortion Measure2007 IEEE International Symposium on Information Theory *(Nice, 2007)* 566-70
[17] Hastie T, Tibshirani R and Friedman J 2009 Hierarchical clustering *The Elements of Statistical Learning* (New York: Springer) pp 520–8