Efficiency of distance measures in the automatic grouping of electronic radio devices by k-means algorithm

To cite this article: G Sh Shkaberina et al 2020 IOP Conf. Ser.: Mater. Sci. Eng. 734 012136

View the article online for updates and enhancements.
Efficiency of distance measures in the automatic grouping of electronic radio devices by k-means algorithm

G Sh Shkaberina¹, E M Tovbis¹, L A Kazakovtsev¹,², A O Shemyakov³, A P Romanov⁴ and N V Lukonin⁴

¹ Reshetnev Siberian State University of Science and Technology, 31, Krasnoyarskiy Rabochiy av., Krasnoyarsk, 660037, Russia
² Siberian Federal University, prosp. Svobodny 79, Krasnoyarsk 660041, Russia
³ Moscow Aviation Institute (National Research University), 4 Volokolamskaoe shosse, Moscow, 125993, Russia
⁴ JSC Academician M.F. Reshetnev «ISS» Zheleznogorsk, Krasnoyarsk region, Russia

E-mail: levk@bk.ru

Abstract. This study investigates the problem of separation mixed lot of semiconductor devices for space applications into homogeneous production batches. The impact of distance measures in clustering algorithm k-means on the accuracy of the separating a mixed production batch is studied. This paper proposes the K-means algorithm with squared Mahalanobis distance as an alternative the squared Euclidean distance. With the various combined mixed lots of semiconductor devices it was shown, that the proposed squared Mahalanobis distance highly depends on the method for determining initial cluster centers.

1. Introduction

Electronic component base (ECB) designed for installation in spacecraft equipment, must meet extremely high requirements. Specialized testing centers conduct a variety of tests for each installed semiconductor device for space applications. Along with incoming control, testing centers subjects ECB to additional rejection tests, including a selective destructive physical analysis (DPA). DPA allows to confirm the good quality of the production batches of ECB, or to identify the batches, having defects due to manufacturing technology and are not detected during conventional rejection tests and additional non-destructive testing. In order to be able to transfer the results of DPA of several devices to the entire batch of semiconductor devices, all devices from the same batch must be made from the same raw materials. If the homogeneous production batches have not been partitioned, the results of the destructive tests of small number of selected devices cannot be extended to the whole shipped lot. Therefore the problem of automatic grouping of semiconductor devices by production batches is very relevant.

It was shown [1], that the problem of allocation of homogeneous batches can be further reduced to a problem of cluster analysis, where each cluster represents one homogeneous batch. To solve the problem of identifying homogeneous batches, in papers [1–4], the application of the clustering optimization algorithm k-means is proposed. In [5], authors consider the clustering method based on the EM algorithm which maximizes the log likelihood function. A model of separation of homogeneous production batches based on a mixture of Gaussian distributions was proposed in [6]. In
[7], authors propose using ensembles of optimization models (k-means, k-medoids, k-medians), EM, as well as their optimized versions. In [1], authors consider the application of genetic optimization algorithms with greedy heuristic procedures, in combination with the EM algorithm for the separation of homogeneous batches of electronic devices. In [8,9] suggested automated ECB grouping by using methods of factor analysis.

2. K-means algorithm with Mahalanobis distance

K-means is one of the most popular unsupervised clustering algorithms, due to its simplicity and efficiency[10]. For a vector of observations X, the k-means algorithm aims to predict k centroids and assign the data points to each centroid to form clusters c_j, j=1..k, while minimizing the average within-cluster dissimilarity. Basic k-means algorithm [10,11] consists of iterative repetition of following steps:

- Step 1. Select an initial partition with K clusters; repeat steps 2 and 3 until cluster membership stabilizes.
- Step 2. Generate a new partition by assigning each pattern to its closest cluster center (centroid).
- Step 3. Compute new cluster centers.

There are various distance measures for calculating the minimum distance between each data point and closest centroid [12]. The distance functions and its definition play an important role in clustering problem. In our study we discover squared Euclidean distance and squared Mahalanobis distance[13-15].

Let μ is the mean of data or the center of clusters (1)

\[ \mu_i = \frac{1}{m} \sum_{j=1}^{m} X_{ji} \]  

where m is number of devices, X_j is vector of values of measured parameter (j=1..m).

Squared Euclidean measure D defined as (2)

\[ D(X_j, \mu_i) = \sum_{i=1}^{n} (X_{ji} - \mu_i)^2 \]  

where n is number of parameters.

Square of Mahalanobis distance D_M defined as (3)

\[ D_M(X) = \sum_{i=1}^{n} (X - \mu)^T S^{-1} (X - \mu) \]  

where S is covariance matrix.

Algorithm k-means with Mahalanobis distance we used as follows:

- Step 1. Split devices randomly into k clusters and compute centroids \( \mu_i \) for each cluster. The centroid \( \mu_i \) calculates as the arithmetic average of all points in the cluster (1). Repeat steps 2 and 3 until cluster membership stabilizes.
- Step 2. Assign each data point to the closest cluster (centroid) by use Mahalanobis distance (3) and generate a new partition.
- Step 3. Compute new cluster centers by (1).

3. Data and processing

Data sample contains seven different homogeneous batches. The total number of devices is 3987; data sample structure is given in table 1. Each batch contains information about 67 non-zero input measured parameters of the product. The sample is deliberately composed of batches, some of which are extremely difficult to separate by known methods of cluster analysis.
Table 1. Structure of data sample (number of devices).

| Batch | Batch 1 | Batch 2 | Batch 3 | Batch 4 | Batch 5 | Batch 6 | Batch 7 |
|-------|---------|---------|---------|---------|---------|---------|---------|
|       | 71      | 116     | 1867    | 1250    | 146     | 113     | 424     |

Nature of parameter’s distributions in different batches is identical, mean standard deviations are commensurable. Data sample in two dimensions is shown in figure 1. Note, that batch 4 visually divides in two groups. It may be related to situation, when two different homogeneous batches marked as one.

![Input data in two-dimensional plot (parameters In21, In39).](image)

Figure 1. Input data in two-dimensional plot (parameters In21, In39).

Various variants of the mixed lot consisting of 2, 3, 4 and 7 homogeneous batches were considered. Clustering was performed by k-means algorithm. The clustering accuracy for considered mixed lots is presented in table 2. Clustering accuracy calculated as a quantity and percentage of exact hits of the algorithm among all clusters. Three experiments were conducted. Experiment 1 and experiment 2 performed with set of random initial cluster centers, experiment 3 performed with pre-clustering, where initial cluster centers chosen from known clusters.

Table 2. Clustering accuracy score.

| Batches     | Experiment 1 | Experiment 2 | Experiment 3 |
|-------------|--------------|--------------|--------------|
|             | Squared      | Squared      | Squared      |
|             | Euclidean    | Euclidean    | Euclidean    |
|             | Mahalanobis  | Mahalanobis  | Mahalanobis  |
|             | distance     | distance     | distance     |
| Batch 1     | -15          | -14          | -13          |
| Batch 2     | -14          | -13          | -12          |
| Batch 3     | -13          | -12          | -11          |
| Batch 4     | -12          | -11          | -10          |
| Batch 5     | -11          | -10          | -9           |
| Batch 6     | -10          | -9           | -8           |
| Batch 7     | -9           | -8           | -7           |

Various advantages of the mixed lot consisting of 2, 3, 4 and 7 homogeneous batches were considered. Clustering was performed by k-means algorithm. The clustering accuracy for considered mixed lots is presented in table 2. Clustering accuracy calculated as a quantity and percentage of exact hits of the algorithm among all clusters. Three experiments were conducted. Experiment 1 and experiment 2 performed with set of random initial cluster centers, experiment 3 performed with pre-clustering, where initial cluster centers chosen from known clusters.
4. Conclusions
As was shown earlier in [5-8] the clustering accuracy decreases with the increase of the number of homogeneous batches in the mixed lot, and this process is independent of the distance calculation method. The best average quality of clustering was achieved in two-batch mixed lot (85% in average). Three-batch and four-batch mixed lots had similar quality of clustering (60% and 67% in average respectively). In seven-batch mixed lot clustering accuracy was lowest (35% in average).

In case of experiments 1 and 2 with random initial cluster centers squared Euclidean distance was stable in all experiments with maximum accuracy deviation of 2%. Mahalanobis distance, on the contrary, highly depends on the method for determining initial cluster centers.

References
[1] Orlov V I, Kazakovtsev LA, Masich IS and Stashkov DV 2017: Algorithmic Support of Decision-making on Selection of Microelectronics Products for Space Industry (Krasnoyarsk: Siberian state aerospace university)
[2] Ackermann M R et al 2012: StreamKM: A clustering algorithm for data streams. J. Exp. Algorithmics. 1 2.4:2.1-2.30
[3] Kanungo T, Mount D M, Netanyahu N S, Piatko C D, Silverman R and Wu A Y 1999: Computing nearest neighbors for moving points and applications to clustering Proc. of the tenth annual ACM-SIAM symp. on Discrete algorithms (Society for Industrial and Applied Mathematics) 931-2
[4] Kazakovtsev L A, Stupina A A and Orlov V I 2014: Modification of the genetic algorithm with greedy heuristic for continuous location and classification problems. Sistemy...
Orlov VI, Stashkov DV, Kazakovtsev LA and Stupina AA 2016: Fuzzy clustering of EEE components for space industry IOP Conf. Series: Materials Science and Engineering 155 012026

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. Ser.: Mater. Sci. Eng. 255 012004

Rozhnov I, Orlov V and Kazakovtsev L 2018 Ensembles of clustering algorithms for problem of detection of homogeneous production batches of semiconductor devices School-Seminar on Optimization Problems and their Applications, OPTA-SCL 2018 2098 338–48

Shkaberina G S, Orlov V I, Tovbis E M and Kazakovtsev LA 2019 Identification of the optimal set of informative features for the problem of separating of mixed production batch of semiconductor devices for the space industry Communications in Computer and Information Science 1090 33394

Shkaberina G, Orlov V, Tovbis E and Kazakovtsev L 2019 Factor analysis with use of the Spearman matrix in the problem of automatic grouping of electronic radio components on production butches Sistemy upravleniya I informatsionnye tehnologii. 2(56) 31-4

Jain A K 2010 Data clustering: 50 years beyond K-means Pattern Recogn Lett. 31 651–66

MacQueen J 1967 Some methods for classification and analysis of multivariate observations Proc. Fifth Berkeley Symp. Math. Stat. Probab. 1 281–97

Farahani R Z, Hekmatfar M2009:Facility Location. Concepts, Models, Algorithms and Case Studies (Springer)

De Maesschalck R, Jouan-Rimbaud D and Massart D L 2000 The Mahalanobis distance Chem Intell Lab Syst 50(1) 1–18

McLachlan G J 1999 Reson. 4(20) BF02834632

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