Impact of data normalization methods and clustering model in the problem of automatic grouping of industrial products

F G Ahmatshin and L A Kazakotsev
Reshetnev Siberian State University of Science and Technology, 31 Krasnoyarsky Rabochy Av., Krasnoyarsk, 660037, Russian Federation

E-mail: levk@bk.ru

Abstract. We compare the results of splitting batches of industrial products (semiconductor devices) into several prospective homogeneous production batches using the standard k-means and p-median clustering models with various normalization methods. In the case of clustering problems, quadratic Euclidean distances are the most popular. We use them as well as the Mahalanobis distances to calculate the differences (distances) in the normalized space of the industrial product features, and compare the clustering results with the use of the Rand index, and empirically establish the advantage of the p-median model.

1. Introduction
The impact of data diversity and increase in their volume require new advances in the methodology of data grouping. Automatic data grouping (cluster analysis) [1] detects it using an algorithm and / or an automatic system based on a certain degree of similarity. So that objects belonging to the same group have similar characteristics (features), and objects belonging to different groups are not similar in their characteristics. We use the metric similarity determination - the distance in a certain attribute space between objects.

We propose various approaches for solving problems of automatic grouping of objects and data. Algorithms for finding related areas of high density [2-12] depend on the selected scale, in which the distances are measured, and on the maximum distance by which points in the group should be removed. The choice of the listed parameters determines the accuracy of the method and the adequacy of the automatic grouping model [13]. Other approaches operating with probabilistic density models have limitations on the dimension of the data, i.e. the number of characteristics of the object.

In all mathematical formulations of the problem of automatic grouping of objects (data points or data vectors), the distance \(L(A_i, A_j)\) between two objects plays a key role. A more precise formulation of the problem (1) is to minimize the sum of the distances from each of the objects to the nearest centroid or center \(X\):

\[
\arg \min_{C(i),j=1,n} \sum_{i=1}^{n} \min_{X \in C(i)} L(X, A_i)
\]

where \(n\) is the number of objects, \(C(i)\) is the set (group) of nodes to which the \(i\) node belongs (there are no more such groups in total than \(p\)), \(A_i\) is the \(i\)th object.
In the p-median clustering model, \( L(X,A) \) means the distance between two points, while in the k-means model, \( L(X,A) \) is the squared distance.

Since the distances or squared distances \( L(X,A) \) between points in a \( d \)-dimensional space obtained as a result of measurements, which are characterized by different parameters and have a different range of values, and different units of measurement, the data is normalized. This paper compares different normalization methods for solving the problem of automatic grouping of objects based on the k-means model and examines the results of clustering accuracy as well as different clustering models.

2. Methods

In k-means and p-median clustering models, various distance measures can be applied. Distance functions as in equation (1) and their definition play an important role. In our study, we use the squares Euclidean distance and the squares Mahalanobis distances. Let \( \mu' \) in equation (2) be the average value of normalized data or the center of clusters

\[
\mu'_j = \frac{1}{m} \sum_{j=1}^{m} X'_{ji}
\]  

The square Euclidean measure \( D \) is defined as in equation (3)

\[
D(X'_j, \mu'_j) = \sum_{i=1}^{n} (X'_{ji} - \mu'_j)^2
\]  

where \( n \) is the number of parameters. The square of the Mahalanobis distance \( D_M \) is defined as in equation (4)

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

where \( S \) is the covariance matrix.

The Mahalanobis distance is scale-invariant [14]. Due to this characteristic data normalization does not matter if this distance is applied. At the same time, a special normalization method along the acceptable parameter drift limits showed high efficiency. Interval normalization (0-1-normalization) or normalization by standard deviation equalizes the significance of all parameters, inevitably increasing the significance of non-informative parameters containing only noise. The binding of the limits of the parameter drift acceptable values are determined by their physical nature. This scale is proportional to the deviation limits of these parameters under operating conditions (permissible drift limits), without reference to the magnitude and variance of these values in a particular production batch.

When moving to the Mahalanobis distance, these advantages are lost. The solution to the problem can be the use of a new method of data normalization \( (X'_1,..,X'_k) \) using the range of admissible values \( A_j \) of the controlled parameters (parameter limits) of the corresponding test modes:

\[
X'_j = \frac{(X_j - X_{\text{min}})}{A_j},
\]  

where \( X'_j \) is the vector of normalized values of the measured parameter \( (j = 1..m) \).

In this study, we used data from test results performed at a test center for batches of integrated circuits (microchips). A separate verification of the belonging of each measured control parameter to a given range based on the parameters of tightened norms was carried out. Integrated circuit batches passed additional screening tests with mandatory burn-in period. It was carried out in dynamic mode for a long time with elevated temperatures.

The initial data is a combination of some parameters of the electronic semiconductor products measured during the mandatory tests. The sample (mixed batch) initially consisted of data on products...
belonging to different homogeneous batches of two integrated circuits 140UD25A and 140UD26A before and after burn-in period. The total number (n) of electronic devices is 1228 and 746 for the first and second type of integrated circuits, respectively. A batch of 140UD25A devices consists of 18 batches, and 9 batches for the 140UD26A devices. Each batch of the first type integrated circuits includes 134, 31, 360, 29, 3, 39, 155, 158, 58, 12, 19, 49, 21, 57, 32, 7, and 15 devices. Batches of the second type integrated circuits have 53, 33, 21, 182, 123, 32, 161, 75 and 66 devices. Products (devices) in each batch are described by 18 input measured parameters.

We use the Rand index as a measure of clustering accuracy [15], which determines the proportion of objects for which the reference sample and resulting cluster splitting are similar. A mixed batch is composed of a combination of batches containing devices belonging to different homogeneous batches. With this approach, a presumably heterogeneous batch can be divided into several presumably homogeneous batches. At the same time, a mixed batch can be divided into a larger number of homogeneous batches than it really contains: smaller clusters are more likely to contain data of the same class, i.e. the likelihood of false attribution of objects of different classes to one cluster is reduced. The share of objects of one class falsely assigned to different classes is not so important for assessing the statistical characteristics of homogeneous groups of objects.

To isolate presumably homogeneous batches with some accuracy, known cluster analysis models were used:

- DE: k-means clustering model with Euclidean distances;
- DM: k-means model using the Mahalanobis distance measure, the covariance matrix is calculated for the entire training sample;
- DM2: k-means model with Mahalanobis distances with a correlation matrix used as S in (4);
- DM3: Mahalanobis distance measures with an average weighted matrix used as S;
- DR: k-means with rectangular distances (Manhattan distances);
- DE2: p-median clustering model with the Euclidean distance.

The objective function is the sum of squares of distances.

### 3. Results and discussion

For our experiment, we used a working sample consisting of four batches. The training set corresponds to the working sample for which clustering was performed. Data sets were recalculated with normalized values \( X' \), as in equation (7).

Normalization by acceptable parameter value limits for which the corresponding norms were established was calculated for 3-16 parameters, the rest of the data was vanished. The difference in the change of 3-6 parameters before and after burn-in period was taken into account by adding 19-22 parameters.

The results were obtained by k-means and p-median clustering. We conducted 30 experiments for each model. The average clustering results are shown in Table 1, the maximum (Max), minimum (Min), average value (Mean) and standard deviation (StD) for the Rand index and the objective function (1) are presented.

| Table 1. Comparison of the results of clustering with various indicators of distance, the number of accurate hits before the burn-in period. |
|---|---|---|---|---|---|
| Batches 1, 3, 8, 9 (n=807) integrated circuit 140UD25A | DE | DM | DM2 | DM3 | DR | DE2 |
| Normalization by standard deviation, Rand index | | | | | |
| Max | 0.718 | 0.610 | 0.792 | 0.599 | 0.599 | 0.803 |
| Min | 0.531 | 0.569 | 0.647 | 0.594 | 0.567 | 0.733 |
| Mean | 0.633 | 0.601 | **0.754** | 0.597 | 0.592 | **0.777** |
| StD | 0.048 | 0.010 | 0.027 | 0.002 | 0.008 | 0.023 |
| Normalization by standard deviation, Objective function (1) | | | | | |

...
|       | Max  | Min  | 11280 | Mean  | 11468 | StD  |
|-------|------|------|-------|-------|-------|------|
|       | 11730 | 352  | 12621 | 809   | 155   | 2956 |
|       |      |      |       | 12121 | 799   | 103  |
|       |      |      |       | 793   | 112   | 2891 |
|       |      |      |       | 793   | 112   | 2891 |
|       |      |      |       | 793   | 112   | 2891 |
|       |      |      |       | 793   | 112   | 2891 |
| Max   | 0.643 | 0.720 | 0.680 | 0.712 | 0.711 | 0.672 |
| Min   | 0.477 | 0.563 | 0.563 | 0.414 | 0.574 | 0.602 |
| Mean  | 0.603 | 0.657 | 0.634 | 0.636 | 0.669 | 0.640 |
| StD   | 0.042 | 0.052 | 0.021 | 0.055 | 0.035 | 0.017 |
|       | Max   | 2245 | 3405  | 2422  | 1567  | 4122 |
| Min   | 1492  | 1589 | 1842  | 1047  | 1501  | 931  |
| Mean  | 1836  | 2283 | 2152  | 1291  | 2305  | 994  |
| StD   | 224   | 707  | 195   | 120   | 739   | 54   |
| Max   | 0.613 | 0.615 | 0.628 | 0.599 | 0.621 | 0.624 |
| Min   | 0.536 | 0.505 | 0.537 | 0.570 | 0.567 | 0.555 |
| Mean  | 0.586 | 0.583 | 0.596 | 0.592 | 0.608 | 0.604 |
| StD   | 0.023 | 0.028 | 0.025 | 0.007 | 0.013 | 0.017 |
| Max   | 10051 | 26388| 10219 | 7376  | 13358 | 2301 |
| Min   | 9621  | 22358| 9832  | 6659  | 11450 | 2233 |
| Mean  | 9866  | 23404| 10070 | 6813  | 11743 | 2255 |
| StD   | 115   | 667  | 101   | 160   | 518   | 17   |
| Max   | 0.884 | 0.909 | 0.976 | 0.619 | 0.654 | 0.973 |
| Min   | 0.578 | 0.574 | 0.751 | 0.583 | 0.556 | 0.751 |
| Mean  | 0.769 | 0.758 | 0.908 | 0.607 | 0.612 | 0.911 |
| StD   | 0.086 | 0.067 | 0.053 | 0.008 | 0.017 | 0.047 |
| Max   | 7663  | 29287| 10802 | 8095  | 17923 | 2212 |
| Min   | 7333  | 27884| 8366  | 7314  | 13847 | 1844 |
| Mean  | 7464  | 28540| 8777  | 7505  | 14402 | 1884 |
| StD   | 69    | 306  | 519   | 207   | 862   | 65   |
| Max   | 0.885 | 0.917 | 0.960 | 0.926 | 0.934 | 0.945 |
| Min   | 0.632 | 0.685 | 0.733 | 0.697 | 0.717 | 0.875 |
| Mean  | 0.821 | 0.867 | 0.902 | 0.871 | 0.892 | 0.910 |
| StD   | 0.059 | 0.050 | 0.043 | 0.049 | 0.062 | 0.022 |
| Max   | 1527  | 1706 | 4666  | 1013  | 1185  | 894  |
| Min   | 872   | 744  | 1939  | 550   | 717   | 687  |
| Mean  | 1148  | 928  | 2323  | 644   | 876   | 711  |
| StD   | 131   | 208  | 638   | 105   | 149   | 39   |
| Max   | 0.697 | 0.774 | 0.779 | 0.652 | 0.663 | 0.782 |
| Min   | 0.523 | 0.630 | 0.616 | 0.569 | 0.578 | 0.597 |
| Mean  | 0.629 | 0.715 | 0.700 | 0.610 | 0.644 | 0.689 |
| StD   | 0.046 | 0.040 | 0.029 | 0.013 | 0.017 | 0.044 |
| Max   | 6584  | 16515| 6742  | 5891  | 10386 | 1619 |
| Min   | 6326  | 15190| 6470  | 5035  | 8914  | 1561 |
| Mean  | 6474  | 15643| 6599  | 5181  | 9208  | 1578 |
| StD   | 57    | 361  | 64    | 207   | 293   | 19   |

Batches 1, 3, 8, 9 (n=201) integrated circuit 140UD25A

Normalization by standard deviation. Rand index
|       | Max   | Min   | Mean  | StD   |
|-------|-------|-------|-------|-------|
|       | 0.706 | 0.559 | 0.616 | 0.033 |
| Max   | 0.618 | 0.518 | 0.588 | 0.022 |
| Mean  | 0.819 | 0.659 | 0.748 | 0.035 |
| StD   | 0.598 | 0.594 | 0.596 | 0.002 |
|       | 0.609 | 0.580 | 0.767 | 0.006 |
|       | 0.838 | 0.628 |       | 0.045 |

Normalization by standard deviation. Objective function (1)

|       | Max   | Min   | Mean  | StD   |
|-------|-------|-------|-------|-------|
|       | 2871  | 89    | 3298  | 198   |
| Max   | 89    | 3298  | 198   | 33    |
| Min   | 2804  | 85    | 3029  | 196   |
| Mean  | 2840  | 86    | 3082  | 197   |
| StD   | 18    | 1     | 61    | 1     |
|       | 6     | 6     | 2     | 6     |

Normalization by acceptable drift value limits. Rand index

| Max   | Min   | Mean  | StD   |
|-------|-------|-------|-------|
| Mean  | 0.661 | 0.731 | 0.680 |
| StD   | 0.494 | 0.585 | 0.504 |
|       | 0.718 | 0.718 | 0.718 |
|       | 0.587 | 0.587 | 0.587 |
|       | 0.583 | 0.583 | 0.583 |
| Mean  | 0.688 | 0.642 | 0.638 |
| StD   | 0.032 | 0.039 | 0.031 |
|       | 0.037 | 0.037 | 0.037 |
|       | 0.027 | 0.027 | 0.027 |

Normalization by acceptable drift value limits. Objective function (1)

| Max   | Min   | Mean  | StD   |
|-------|-------|-------|-------|
| Mean  | 539   | 941   | 741   |
| StD   | 326   | 372   | 435   |
|       | 372   | 435   | 270   |
|       | 349   | 244   | 275   |
| Mean  | 435   | 616   | 563   |
| StD   | 57    | 198   | 63    |
|       | 18    | 13    | 28    |

Normalization by parameter acceptable value limits. Rand index

| Max   | Min   | Mean  | StD   |
|-------|-------|-------|-------|
| Mean  | 0.605 | 0.642 | 0.637 |
| StD   | 0.376 | 0.556 | 0.550 |
|       | 0.610 | 0.626 | 0.610 |
|       | 0.580 | 0.571 | 0.580 |
|       | 0.486 | 0.486 | 0.486 |
| Mean  | 0.605 | 0.642 | 0.637 |
| StD   | 0.058 | 0.021 | 0.025 |
|       | 0.014 | 0.014 | 0.014 |
|       | 0.027 | 0.027 | 0.027 |

Normalization by parameter acceptable value limits. Objective function (1)

| Max   | Min   | Mean  | StD   |
|-------|-------|-------|-------|
| Mean  | 2518  | 5764  | 2568  |
| StD   | 326   | 372   | 435   |
|       | 349   | 244   | 275   |
|       | 277   | 612   | 623   |
| Mean  | 435   | 616   | 563   |
| StD   | 57    | 198   | 63    |
|       | 18    | 13    | 28    |

Normalization by acceptable value limits. Objective function (1)

| Max   | Min   | Mean  | StD   |
|-------|-------|-------|-------|
| Mean  | 0.717 | 0.715 | 0.990 |
| StD   | 0.476 | 0.516 | 0.650 |
|       | 0.460 | 0.572 | 0.572 |
|       | 0.572 | 0.572 | 0.572 |
| Mean  | 0.902 | 0.601 | 0.601 |
| StD   | 0.055 | 0.052 | 0.071 |
|       | 0.015 | 0.015 | 0.015 |
|       | 0.095 | 0.095 | 0.095 |

Normalization by acceptable value limits. Objective function (1)

| Max   | Min   | Mean  | StD   |
|-------|-------|-------|-------|
| Mean  | 1912  | 7057  | 2980  |
| StD   | 1784  | 6664  | 2044  |
|       | 1743  | 3208  | 1600  |
|       | 1826  | 3384  | 1867  |
| Mean  | 1846  | 6855  | 2215  |
| StD   | 29    | 110   | 38    |
|       | 169   | 511   | 169   |
|       | 511   | 511   | 511   |

Normalization by standard deviation. Rand index

| Max   | Min   | Mean  | StD   |
|-------|-------|-------|-------|
| Mean  | 0.925 | 0.937 | 0.966 |
| StD   | 0.704 | 0.685 | 0.694 |
|       | 0.554 | 0.554 | 0.554 |
|       | 0.564 | 0.564 | 0.564 |
| Mean  | 0.891 | 0.856 | 0.856 |
| StD   | 0.062 | 0.076 | 0.065 |
|       | 0.099 | 0.099 | 0.099 |
|       | 0.086 | 0.086 | 0.086 |
|       | 0.017 | 0.017 | 0.017 |

Normalization by acceptable drift value limits. Objective function (1)

| Max   | Min   | Mean  | StD   |
|-------|-------|-------|-------|
| Mean  | 349   | 453   | 1273  |
| StD   | 206   | 200   | 462   |
|       | 219   | 159   | 256   |
| Mean  | 284   | 273   | 653   |
| StD   | 31    | 69    | 226   |
|       | 72    | 9     | 168   |
|       | 168   | 168   | 168   |

Normalization by parameter acceptable value limits. Rand index

| Max   | Min   | Mean  | StD   |
|-------|-------|-------|-------|
| Mean  | 0.640 | 0.737 | 0.700 |
| StD   | 0.501 | 0.521 | 0.372 |
|       | 0.555 | 0.555 | 0.555 |
|       | 0.580 | 0.580 | 0.580 |
| Mean  | 0.675 | 0.635 | 0.602 |
| StD   | 0.036 | 0.041 | 0.072 |
|       | 0.022 | 0.022 | 0.022 |
|       | 0.018 | 0.018 | 0.018 |
|       | 0.043 | 0.043 | 0.043 |

Normalization by parameter acceptable value limits. Objective function (1)

| Max   | Min   | Mean  | StD   |
|-------|-------|-------|-------|
| Mean  | 1615  | 4366  | 1769  |
| StD   | 200   | 159   | 256   |
|       | 168   | 168   | 168   |
| Mean  | 1401  | 1401  | 1401  |
| StD   | 2442  | 2442  | 2442  |
|       | 445   | 445   | 445   |
4. Conclusion

Our experiments with the labeled data of the industrial product testing show that the data normalization method as well as the selected clustering model play an important role. In according to the Rand index, the preferable (or at least competitive) model for the industrial product clustering is the $p$-median model, and the normalization by the acceptable drift value limits. Thus, the further development of efficient clustering algorithms based on the $p$-median model is a promising research area.

Acknowledgement

This work was supported by the Ministry of Science and Higher Education of the Russian Federation (Project FEFE-2020-0013).

References

[1] Tabachnick B G and Fidell L S 2007 Using Multivariate statistics (Boston, MA: Allynand Bacon)
[2] Frank I and Todeschini R 1994 Data analysis handbook Amsterdam: Elsevier Science Inc 14 227-8
[3] Ester M et al. 1996 Density-based algorithm for discovering clusters in large spatial databases with noise KDD-96 Proceedings 1 226-31
[4] Ruban A I 1982 Identification and sensitivity of complex (Tomsk State University) p 302
[5] Lloyd E 1989 Handbook of applied statistics (Moscow: Finance and statistics) 1 p 510
[6] Parzen E 1962 On the estimation of probability density function and the mode Ann. Math.Statist. 33 1065
[7] Parzen E 1982 On estimation of a probability density, function and Mode II IEEE Transactions on Information Theory 4 663-6
[8] Epanechnikov V A 1969 Non-parametric assessment of multidimensional probability density Theory of probability and its applications 14 156-61
[9] Medvedev A V 1973 Non-parametric estimates of probability density and its derivatives Automation of industrial experiment, Frunze: Ilim 22-31
[10] Nadaraya E A 1964 On regression assessment Probability theory and its application 9 157-9
[11] Sage A P and Melsa J L 1971 System Identification Mathematics in Science and Engineering 80 221
[12] Ivanova N V and Protasov K T 1982 Determination of smoothing parameters in non-parametric estimates of the density function by sample Mathematical statistics and its applications 8 50-65
[13] Bellman R E 1957 Dynamic programming (Princeton University Press.) p 392
[14] Kazakovtsev L A, Orlov V I and Stupina A A 2015 On distance metric for the system of automatic classification of the EEE devices by production batches Programmnye produkty i sistemy 2 124-9
[15] Rand W M 1971 Objective criteria for the evaluation of clustering methods Journal of the American Statistical Association 66 846-50