Ensembles of Clustering Algorithms for Problem of Detection of Homogeneous Production Batches of Semiconductor Devices

Ivan Rozhnov[1], Victor Orlov[1], and Lev Kazakovtsev[1,2]

[1] Reshetnev Siberian State University of Science and Technology,
prosp. Krasnoyarskiy Rabochiy 31, 660031, Krasnoyarsk, Russia
[2] Siberian Federal University,
prosp. Svobodny 79, 660041, Krasnoyarsk, Russia
levk@bk.ru

Abstract. To complete the on-board equipment of space systems with a highly reliable electronic component base (ECB), specialized test centers perform hundreds of tests to analyze each semiconductor device. One of the requirements is that the shipped lot of products must be made from a single batch of raw materials (wafers) which is not guaranteed if the devices are not manufactured for use in the space industry only. To solve the problem of detecting homogeneous production batches, various clustering algorithms are implemented on multidimensional data of test results. In practice, it is impossible to predict in advance which of the algorithms in each particular case will show the most adequate results and the use of the ensemble approach is promising. Most of the clustering algorithms for the problem of dividing the ECB mixed lot into two homogeneous production batches show rather high accuracy. With an increase in the number of homogeneous production batches in the mixed lot, the accuracy decreases.

Authors propose an approach to constructing an ensemble of clustering algorithms based on co-occurrence matrices with weight coefficients. Results of computational experiments on specially mixed lots of the ECB show that for the such large-scale problems, the use of the ensemble approach allows to achieve a higher adequacy of the results. Individual algorithms can show results that exceed the ensemble's accuracy, but the accuracy of the ensemble is still higher than the averaged accuracy of individual algorithms.

Keywords: Clustering algorithms · Electronic component base · Semiconductor devices · Ensembles of algorithms

In: S. Belim et al. (eds.): OPTA-SCL 2018, Omsk, Russia, published at http://ceur-ws.org

1 Introduction

Intensive use of big data in various areas leads to increased interest of researchers in methods and tools for processing and analysing datasets of huge volumes and diversity. One of the promising directions of big data analysis is the cluster analysis, which allows solving such problems as reducing the size of the initial data set, identifying patterns, etc [14]. The goal of the automatic grouping (clustering) is detection of a natural grouping of a number of samples, points or objects. The solution of the clustering problem is reduced to the development of an algorithm or an automated system capable of detecting these natural groupings in unmarked data.

Clustering [13] is segmentation through the allocation of certain associations of homogeneous elements which are considered as independent objects with certain properties [3]. As a result, the clustering procedure forms "clusters", i.e. groups of very similar objects [25].

A criterion for the clustering quality is some functional which depends on the scatter of objects within the group (cluster) and the distances between them [4]. Modern methods of cluster analysis offer a wide variety of methods for revealing heterogeneous groups of parameters. The most common of these methods is the k-means procedure [21, 2]. Algorithms implementing this method are local optimization algorithms which depend on a choice of initial parameters (centroids of clusters). At the same time, for many problems, the preferred methods of identifying groups in data must produce reproducible results.

The on-board units of spacecrafts must be equipped with a highly reliable electronic component base (ECB). First of all, it is necessary to prevent counterfeit products that do not meet the reliability requirements, ensure the purchase of ECB from authorized suppliers and passing through the 100 % input control, additional rejection tests and destructive physical analysis (DPA) of ECB. Individual rejection tests of components are essential [24]. The shipped ECB lots (batches) may be inhomogeneous, collected from several production batches [23]. Therefore, the test results of the DPA of several ECB samples cannot be extended to the entire lot (batch) of components unless we are sure that all components of this lot are manufactured as a single production batch from a single batch of wafers. Relatively small fluctuations in the manufacturing process can radically affect the sensitivity to radiation and other characteristics of the semiconductor devices.

ECB clustering is important in terms of ensuring reliability and, even more, radiation resistance. Ionizing radiation as a physical factor of the space environment determines the period of active existence of space systems.

At present, there is a tendency to use collective methods in cluster analysis [10, 27]. The algorithms of cluster analysis are not universal: each algorithm has its own special field of application. In case of different types of data sets, to select clusters, a researcher needs to apply a set of various algorithms to select the best one. The ensemble (collective) approach allows to reduce the dependence of the final solution on the parameters of the original algorithms and obtain a stable solution, even in case of noise and emissions in the dataset [4].

2 Results of Various Standalone Clustering Algorithms

As datasets for our experiments, we used the results of non-destructive tests of mixed production batches of the ECB performed in a specialized test center. The composition of the mixed production batches was known in advance. These mixed batches were completed from several obviously homogeneous batches of the ECB:

- 140UD25AVK: 2 production batches (ECB clusters) and comparatively small data volume (56 data vectors of dimensionality 18);
- 3OT122A: 2 batches (767 data vectors of dimensionality 10);
- 1526LE5: 6 batches (963 data vectors of dimensionality 41).

Our problem was to divide the mixed batch into homogeneous components and analyze the quality of this division.

We used 5 common clustering algorithms [6]: k-Means [21, 22, 7, 1], k-Means-fast [11], k-Means-kernel [9], k-Medoids [12], EM algorithm (Expectation Maximization) [8].

In addition to the actual form of the clustering algorithm, the result is significantly influenced by the parameters of the algorithms which can be optimized by their values. By optimization, we mean the selection of such values of some optimized parameter at which the maximum clustering accuracy is ensured, that is, the best match of the result of clustering to the true partition of the mixed batch into homogeneous batches of ECB is achieved. As an optimized parameter in the k-Means, k-Means (fast) and k-Medoids algorithms, we used the type of distance measure. For the k-Means (kernel) algorithm, tried to use various types of the kernel (dot / radial kernel). For the EM algorithm, we tried to find the optimal number of optimization steps in each iteration.

At the output of this process, we evaluate our results by the accuracy. By accuracy, we mean the proportion of data objects assigned to the "right" cluster. This "correctness" can be assessed by having a sample of marked data, for which it is known in advance that they are assigned to a particular cluster. In this case, our samples are combined from data from separate homogeneous batches of ECB. The results are summarized in Table 1.

As we can see from this table, the clustering algorithms with relatively small data volumes and small number of production batches (clusters) show rather high accuracy. With the increase in data volumes and the number of clusters, clustering accuracy decreases.

For clustering models, the most important parameter affecting the result is the distance measure used. The use of special measures sometimes allows us to adapt simple models like k-means to rather complex clustering problems. In case of using some complex and non-standard distance measures, a sufficient condition for the applicability of the measure of distance is the existence of an algorithm for solving the corresponding Weber problem, i.e. the problem of finding the center of the cluster [15, 26].

For comparative analysis, in addition to the problems of ECB batches clustering, we analyzed the features of clustering algorithms and their ensembles on the most common data sets from the UCI Machine Learning Repository [20]

with comparable data volumes and dimensionalities :
- Cryotherapy [18, 19] - 2 clusters (90 data vectors of dimensionality 6);
- pima-indians-diabete - 2 clusters (768 data vectors of dimensionality 8);
- ionosphere - 2 clusters (351 data vectors of dimensionality 34);
- Iris - 3 clusters (150 data vectors of dimensionality 4);
- Zoo - 7 clusters (101 data vectors of dimensionality 16).

Results of standalone algorithms were summarized in Table 2.

Due to problems concerning the behavior of the EM algorithm with comparatively small datasets in the multidimensional space (all objects of a small cluster in a multidimensional space belong to the same hyperplane and the corresponding probability distribution represented by its covariation matrix collapses into the hyperplane), our realizatio of the EM algorithm did not allow to obtain results for some particular cases ("no result" in Table 2). In particular cases, analogous problem arose in the procedure which optimized parameters of the other algorithms.

3 Ensembles of Clustering Algorithms

The ensemble approach is one of the most promising directions in cluster analysis [14]. The following basic techniques for constructing an ensemble of algorithms are commonly used [5]:

1. Finding a consensus partition, i.e. consistent partitioning with several available solutions, optimal for some criterion;

2. Calculation of a consistent matrix of similarity/differences (co-occurrence matrix).

When forming the final solution, an ensemble uses the results obtained by various algorithms.

Let us consider an example of an ensemble of algorithms [14]. It is a combination of separate algorithms, each of which offers its own partition, and a hierarchical agglomerative algorithm that combines the resulting solutions with a special mechanism.

In the first step, each algorithm splits the data into clusters using its objective function, based on the distance metric or on the likelihood function. Then, the accuracy and weight of the view of the algorithm in the ensemble are calculated by the following equation:

$$W_i = \frac{Acc_i}{\sum_{i=1}^{L} Acc_i} \tag{1}$$

where Acc_i is the accuracy of the ith algorithm, i.e. the ratio of the number of correctly clustered objects to the volume of the entire sample, and L is the number of the algorithms in our ensemble.

For each partition obtained, our algorithm compiles a preliminary binary matrix of differences of size $n \times n$ (where n is the number of objects) to determine whether the objects of the partition are included in the same clusters. After that, our algorithm calculates a matched matrix of differences, each element of which

Table 1. Results of computational experiments with standalone clustering algorithms on ECB production batches

Algorith	Accuracy / optimized parameter value		
	140UD25AVK 2 batches	3OT122A 2 batches	1526LE5 6 batches
k-Means	100.00 (Euclidean Distance) [1]	76.53 (Euclidean Distance)	50.57 (Euclidean Distance)
k-Means(fast)	100.00 (Euclidean Distance)	67.67 (Euclidean Distance)	50.57 (Euclidean Distance)
k-Means(kernel)	100.00 (radial kernel)	59.19 (radial kernel)	47.14 (radial kernel)
k-Medoids	100.00 (Euclidean Distance)	60.63 (Euclidean Distance)	48.60 (Euclidean Distance)
EM	96.43 (100 optimization steps in each iteration)	90.09 (100 optimization steps in each iteration)	No result
k-Means Optim.	100.00 (Euclidean Distance)	76.53 (Euclidean Distance)	63.03 (Overlap Similarity)
k-Means(fast) Optim.	100.00 (Euclidean Distance)	76.53 (Euclidean Distance)	50.99 (KernelEuclidean Distance)
k-Means(kernel) Optim.	53.57 (dot kernel)	67.67 (dot kernel)	30.22 (dot kernel)
k-Medoids Optim.	100.00 (Euclidean Distance)	91.79 (Euclidean Distance)	55.97 (Manhattan Distance)
EM Optim.	96.43 (40 optimization steps)	95.44 (95 optimization steps)	No result

[1] optimization parameter value

Table 2. Results of computational experiments with standalone clustering algorithms on datasets from the UCI repository

Algorith	Accuracy / optimized parameter value				
	Cryotherapy 2 batches	Pima-indians-diabetes 2 batches	Ionosphere 2 batches	Iris 3 batches	Zoo 7 batches
k-Means	56.67 (Euclidean Distance) [1]	66.02 (Euclidean Distance)	71.23 (Euclidean Distance)	89.33 (Euclidean Distance)	75.25 (Euclidean Distance)
k-Means(fast)	56.67 (Euclidean Distance)	66.02 (Euclidean Distance)	71.23 (Euclidean Distance)	89.33 (Euclidean Distance)	75.25 (Euclidean Distance)
k-Means(kernel)	55.56 (radial kernel)	51.17 (radial kernel)	55.56 (radial kernel)	93.33 (radial kernel)	54.46 (radial kernel)
k-Medoids	57.78 (Euclidean Distance)	54.43 (Euclidean Distance)	68.09 (Euclidean Distance)	76.67 (Euclidean Distance)	79.21 (Euclidean Distance)
EM	56.67 (100 steps)	65.62 (100 steps)	No result	96.67 (100 steps)	No result
k-Means Optim.	75.56 (Camberra Distance)	66.28 (Manhattan Distance)	No result	96.67 (Cosine Similarity Distance)	83.17 (Manhattan Distance)
k-Means(fast) Optim.	75.56 (Camberra Distance)	66.28 (Manhattan Distance)	No result	96.67 (Cosine Similarity)	83.17 (Manhattan Distance)
k-Means(kernel) Optim.	53.33 (dot kernel)	65.10 (dot kernel)	64.10 (dot kernel)	33.33 (dot kernel)	40.59 (dot kernel)
k-Medoids Optim.	73.33 (Camberra Distance)	66.02 (Dynamic Time Warping Distance)	72.36 (Jaccard Similarity Distance)	97.33 (Cosine Similarity Distance)	80.20 (Cosine Similatity Distance)
EM Optim.	56.67 (1 optimization step)	66.28 (1 optimization step)	No result	96.67 (101 optimization steps)	No result

[1] optimization parameter value

is a weighted sum (using the weight of equation (1)) of the elements of the preliminary matrices. The obtained matrix is used as input for the algorithm of hierarchical agglomerative clustering. Then, using common techniques, such as determining the jump in the agglomeration distance, we can choose the most suitable cluster solution.

As mentioned above, to obtain the best partitioning into clusters, a binary matrix of similarity/differences for each partition in the ensemble is constructed:

$$H_i = \langle h_i(i,j) \rangle$$

where $h(i,j) = 0$ if both ith and jth elements belong to the same cluster, and 1, otherwise.

The next step in composing an ensemble of clustering algorithms is to compile a matched matrix of binary partitions.

$$H^* = \langle h^*(i,j) \rangle, \quad h^*(i,j) = \sum_{i=1}^{L} w_i h_i(i,j)$$

where w_i is the weight of the ith algorithm.

The most popular clustering algorithms often fail for certain datasets that do not match well with the modeling assumptions [10]. Ensembles which include approaches such as k-means that are better suited to low-dimensional spaces in combination with other approaches designed for high-dimensional sparse spaces (spherical k-means, Jaccard-based clustering, EM-clustering with spherical Gaussian distributions [24] etc.) perform well across a wide range of data dimensionality [27]. At the same time, in high-dimensional cases, the choice of the best clustering models in not evident: sometimes, algorithms designed for high-dimensional data fail to improve the results of the simplest models such as k-means [24].

For constructing an ensemble (Table 3), we take three or five best algorithms showing the highest accuracy for each specific dataset (Table 1).

For the 140UD25AVK dataset, we used k-Means, k-Means(kernel) and k-Medoids to construct an ensemble of three best algorithms; for 3OT122A dataset, we used EM-Optim., k-Medoids-Optim. and EM; for 1526LE5, we used k-Means-Optim., k-Medoids-Optim. and k-Means(fast)-Optim.

Analogous results for various datasets from the UCI Repository are shown in Table 4 and Table 5.

A fragment of calculation of ensemble results for dataset 3OT122A is given in Table 6. In most rows, some of standalone algorithms demonstrate wrong result and the ensemble improves this situation.

4 Conclusions

Our computational experiments show that any clustering algorithms for the problem of dividing a batch of ECB batch into two homogeneous batches can be used with rather high accuracy. With increase in the number of homogeneous

Table 3. Results of computational experiments with ensembles on ECB production batches (accuracy)

ECB mixed production batch / algorithm	140UD25AVK 2 batches	3OT122A 2 batches	1526LE5 6 batches
The best standalone algorithm	100	95.44	63.03
Averaged accuracy of 3 best algorithms	100	92.44	56.66
Averaged accuracy of 5 best algorithms	100	86.08	54.23
Ensemble of 3 algorithms	100	95.04	57.01
Ensemble of 5 algorithms	100	95.44	52.54

Table 4. Algorithms for each dataset sorted by their accuracy

Dataset /range	Cryotherapy 2 clusters	Pima-indians-diabetes 2 clusters	ionosphere 2 clusters	Iris 3 clusters	Zoo 7 clusters
1	k-Means -Optim	k-Means -Optim	k-Medoids -Optim	k-Medoids -Optim	k-Means -Optim
2	k-Means(fast) -Optim	k-Means(fast) -Optim	k-Means	EM	k-Means(fast) -Optim
3	k-Medoids -Optim	EM	k-Means(fast)	k-Means-Optim	k-Medoids -Optim
4	k-Medoids	k-Means	k-Medoids	k-Means(fast) -Optim	k-Medoids
5	EM	k-Means(fast)	k-Means (kernel)-Optim	EM	k-Means

Table 5. Results of computational experiments with ensembles on datasets from the repository (accuracy)

Dataset/algorithm	Cryotherapy 2 clusters	pima-indians-diabetes 2 clusters	ionosphere 2 clusters	Iris 3 clusters	Zoo 7 clusters
The best standalone algorithm	75.56	66.28	72.36	97.33	83.17
Averaged accuracy of 3 best algorithms	74.82	66.28	71.61	96.89	82.18
Averaged accuracy of 5 best algorithms	67.78	66.18	69.40	96.80	80.20
Ensemble of 3 algorithms	75.56	66.28	71.23	96.71	83.17
Ensemble of 5 algorithms	75.56	65.89	68.66	96.67	81.15

Table 6. Comparison of results of an ensemble of 5 algorithms and standalone algorithms (incorrect results of the ensemble are marked by "*")

No. of semiconductor device in the mixed batch	True batch number	EM-Optim	k-Medoids -Optim	EM	k-Means	k-Means (fast) -Optim	Ensemble
1	1	1	2	1	2	1	1
2	1	1	2	1	1	1	1
3	1	1	1	1	1	1	1
4	1	1	1	1	1	1	1
5	1	1	2	1	2	2	2*
6	1	1	2	1	2	1	1
7	1	1	2	1	1	1	1
8	1	1	2	1	1	1	1
9	1	1	2	1	1	1	1
10	1	1	1	1	1	1	1
...	...	...	...	...	...	...	...
71	2	2	2	2	1	1	2
72	2	2	2	2	2	2	2
73	2	2	2	2	2	2	2
74	2	2	2	2	1	1	2
...	...	...	...	...	...	...	...

production batches in the mixed batch, the accuracy decreases. For different data sets, the best results are demonstrated by different algorithms.

Using ensemble approach allows achieving higher accuracy in comparison with standalone clustering algorithms. In this case, individual algorithms are able to show results that exceed the ensemble's accuracy, however, the accuracy of the ensemble is still higher than the averaged accuracy of the individual algorithms. It is also necessary for a particular problem to take into account the number of algorithms used in the ensemble, in connection with the fact that the accuracy of the ensemble of clustering algorithms for various data depends on the number of algorithms in the ensemble.

In practice, the accuracy of clustering cannot be determined due to the lack of information on the actual classes in the sample and it is impossible to predict a priori which of the algorithms in the particular case shows the most adequate results. Thus, usage of an ensemble approach to our problem is a promising research direction. In particular, the application of the ensemble approach in combination with the clustering algorithms that provide the best result within the framework of the given clustering model [17, 16] will make it possible to obtain results which are both more adequate and reproducible under repeated runs of the algorithm and hence verifiable.

The last table shows that for three of five data sets, the ensembles of algorithms show results that are worse than the averaged value of the individual algorithms from which they are composed. This is typical for ensembles of both three and five best algorithms.

Though for our problem of mixed production batch separation, the ensemble approach does not show an advantage over individual algorithms in all cases, in general, the ensemble approach allows reducing the dependence of the obtained results on the features of using separate algorithms to a specific data set. Taking into account that the best results for different data sets are achieved by different algorithms, selection of some set of the best algorithms that show good results for many problems of such class increases the reliability of the process of homogeneous ECB production batch separation.

Acknowledgement. Results were obtained in the framework of the state task No. 2.5527.2017/8.9 of the Ministry of Education and Science of the Russian Federation.

References

1. Arthur, D., Vassilvitskii, S.: How slow is the k-means method? In: Proceedings of the twenty-second annual symposium on Computational geometry. pp. 144–153. ACM (2006)
2. Arthur, D., Vassilvitskii, S.: k-Means++: The advantages of careful seeding. In: Proc. of the Eighteenth Annual ACM-SIAM Symp. on Discrete algorithms, ser. SODA '07. pp. 1027–1035 (2007)
3. Baturkin, S.A., Baturkina, E.Yu., Zaimenko, V.A., Sihinov, Y.V.: Statistical data clustering algorithms in adaptive learning systems. Vestnyk RHRTU 1(31), 82–85 (2010)
4. Berikov, V.B.: Construction of the Ensemble of Logical Models in Cluster Analysis. Lect. Notes Artif. Intel. 5755, 581–590 (2009)
5. Berikov, V.: Weighted ensemble of algorithms for complex data clustering. Pattern Recognition Letters 38, 99–106 (2014)
6. Berkhin, P.: A survey of clustering data mining techniques. Grouping multidimensional data. Springer, 25–71 (2006)
7. Bhattacharya, A., Jaiswal, R, Ailon, N.: A tight lower bound instance for k-means++ in constant dimension. Theory and Applications of Models of Computation. Springer, 7–22 (2014)
8. Dempster, A., Laird, N., Rubin, D.: Maximum likelihood estimation from incomplete data. Journal of the Royal Statistical Society, Series B 39, 1–38 (1977)
9. Dhillon, I.S., Guan, Y., Kulis, B.: Kernel k-means: spectral clustering and normalized cuts. In: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '04). pp. 551–556. ACM, New York, USA (2004). https://doi.org/10.1145/1014052.1014118
10. Ghosh, J., Acharya, A.: Cluster ensembles. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 1(4), 305–315 (2011)
11. Hamerly, G., Drake, J.: Accelerating Lloyd's algorithm for k-means clustering. Partitional Clustering Algorithms. Springer, 41–78 (2014)
12. Kaufman, L., Rousseeuw, P.J.: Clustering by means of Medoids. Statistical data analysis based on the L1-norm and related methods. pp. 405–416. Springer, US (1987)
13. Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley and Sons (1990)

14. Kausar, N., Abdullah, A., Samir, B.B., Palaniappan, S., AlGhamdi, B.S., Dey, N.: Ensemble clustering algorithm with supervised classification of clinical data for early diagnosis of coronary artery disease. Journal of Medical Imaging and Health Informatics 6(1), 78–87 (2016)
15. Kazakovtsev, L.A., Stanimirović, P.S., Osinuga, I.A., Gudyma, M.N., Antamoshkin, A.N.: Algorithms for location problems based on angular distances. Advances in Operations Research 2014 Article ID 701267 (2014)
16. Kazakovtsev, L.A., Antamoshkin, A.N.: Greedy heuristic method for location problems. Vestnik SibGAU 16(2), 317–325 (2015)
17. Kazakovtsev, L.A., Stashkov, D.V., Rozhnov, I.P., Kazakovtseva, O.B.: Further development of the greedy heuristic method for clustering problems. Control Systems and Information Technology 4(70), 34–40 (2017)
18. Khozeimeh, F., Alizadehsani, R., Roshanzamir, M., Khosravi, A., Layegh, P., Nahavandi, S.: An expert system for selecting wart treatment method. Computers in Biology and Medicine 81, 167–175 (2017)
19. Khozeimeh, F., Jabbari Azad, F., Mahboubi Oskouei, Y., Jafari, M., Tehranian, S. Alizadehsani, R. et al.: Intralesional immunotherapy compared to cryotherapy in the treatment of warts. International Journal of Dermatology (2017). https://doi.org/10.1111/ijd.13535
20. Lichman, M.. UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science (2013). http://archive.ics.uci.edu/ml
21. Lloyd, S.P.: Least squares quantization in PCM. IEEE Transactions on Information Theory 28(2), 129–137 (1982). https://doi.org/10.1109/TIT.1982.1056489
22. MacQueen, J.: Some methods for classification and analysis of multivariate observations. In: Proc. 5th Berkeley Symp. on Math. Statistics and Probability. pp. 281–297 (1967)
23. MIL-PRF-38535 Performance Specification: Integrated Circuits (Micricircuit) Manufacturing, General Specifications for. Department of Defence, United States of America (2007)
24. Orlov, V. I., Stashkov, D. V., Kazakovtsev, L. A., Stupina, A. A.: Fuzzy clustering of EEE components for space industry. In: IOP Conference Series: Materials Science and Engineering 155, Article ID 012026 (2016)
25. Sehgal, G., Garg, K.: Comparison of various clustering algorithms. International Journal of Computer Science and Information Technologies 5(3), 3074–3076 (2014)
26. Stojanovic, I., Brajevic, I., Stanimirović, P.S., Kazakovtsev, L.A., Zdravev, Z.: Application of heuristic and metaheuristic algorithms in solving constrained weber problem with feasible region bounded by arcs. Mathematical Problems in Engineering 2017 Article ID 8306732 (2017). https://doi.org/10.1155/2017/8306732
27. Strehl, A., Ghosh, J. Cluster ensembles - a knowledge reuse framework for combining ultiple partitions. J. Mach. Learn. Res. 3, 583–617 (2002)