European Proceedings of
Computers and Technology
EpCT

www.europeanproceedings.com

e-ISSN: 2672-8834

DOI: 10.15405/epct.23021.45

**HMMOCS 2022**
**International Workshop "Hybrid methods of modeling and optimization in complex systems"**

# GREEDY HEURISTICS FOR THE CHOICE OF THE RADIUS OF LOCAL CONCENTRATIONS IN FOREL-2 ALGORITHM

F. G. Ahmatshin (a)*, L. A. Kazakovtsev (b)
*Corresponding author

(a) Reshetnev Siberian State University of Science and Technology, Krasnoyarsk, Russia, ahmatshin_fg@sibsau.ru
(b) Siberian Federal University, Krasnoyarsk, Russia, Reshetnev Siberian State University of Science and Technology, Krasnoyarsk, Russia, levk@bk.ru

## Abstract

The authors examine the problem of choosing the search radius for local concentrations in the FOREL-2 clustering algorithm with an initial number of clusters. Our approach was aimed at improving the accuracy and stability of the result, such as identifying homogeneous batches of industrial products. We examined the k-means and FOREL-2 algorithms by using normalized standard deviation test values and by valid parameter values for the problem of automatic classification of objects in a multi-dimensional space of measured parameters. For such problems, with the use of the FOREL-2 algorithm, we apply greedy heuristic procedures to select the radius of local concentrations. According to the obtained Rand index, the approach which uses the FOREL-2 algorithm demonstrated the best accuracy with a larger value of the objective function in comparison with the k-means algorithm. The accuracy and speed of the software implementation of the algorithm are quite acceptable for solving the problem of clustering electronic radio products based on test data. The use of greedy heuristics for choosing the radius of the search for local concentrations in the FOREL-2 clustering algorithm with a specified number of clusters has an advantage in the speed of exact clustering compared to the k-means algorithm that uses greedy heuristics for choosing centroids.



*Keywords:* FOREL-2, k-means, clustering, greedy heuristics, electronic radio products

https://doi.org/10.15405/epct.23021.45
Corresponding Author: F. G. Ahmatshin
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2672-8834

## 1. Introduction

MacQueen (1967) proposed the classical and most common problem-oriented clustering algorithms. In the field of data clustering of electronic radio products, an algorithm of this kind is the k-means algorithm proposed primarily by Lloyd (1982). Kazakovtsev et al. (2015) continued research in this area related to the choice of data normalization method, the choice of the distance metric and the choice of the centroid initialization method. The algorithm of Milligan and Cooper (1985) depends on giving a priori the number of clusters and finds all groups to cluster at the same time. In the work of Klosgen and Zytkow (1996), the clustering process consists of grouping n test case observations into k groups. In the algorithm of Anderberg (1973), Bobrowski and Bezdek (1991), a representative observation is specified by a centroid using a d-dimensional feature vector. Optimal clustering is achieved at the minimum value of the objective function. The objective function takes into account the set of observations ($x_1,x_2,\ldots,x_n$) that are divided into k clusters $C=\{C_1,C_2,\ldots,C_n\}$ by the following formula:

$$\arg\min \sum_{i=1}^{k} \sum_{x \in C_i} \left\| x - \mu_i \right\|^2 , \quad (1)$$

The algorithm of Li and Wu (2012) raises the speed of convergence effectively by improving the way of selecting initial cluster focal point. Hossain et. al. (2019) improved the k-means algorithm also related to the Euclidian distance between two points which is less than or equal to the threshold value, then these two data points will be in the same group. The criterion of Pérez-Ortega et al. (2018) enables us to balance the use of benefits at the convergence stage. In the work of Na et al. (2010), the algorithm saves the running time directly to computing the distance of each data object to the cluster centres frequently. According to Wu et al. (2008), this is one of the most influential data mining algorithms in the research community.

## 2. Problem Statement

The FOREL-2 algorithm by Zagoruiko (1999) is a modification of the basic FOREL algorithm that was proposed by Zagoruiko and Yolkina (Zagoruiko, 1999). If we compare it with the k-means algorithm, we can find a lot in common. Clusters obtained by these algorithms have a spherical shape. The results of the algorithm work depend on the data normalization method.

In the algorithm of Zhuravleva, the features of the objects are initially normalized in such a way that the values of all parameters are in the range from zero to one. In comparison with k-means, the FOREL-2 algorithm has a range of specific features. The number of clusters depends on the radius of the sphere. The more it is, the less clusters are obtained.

In the general statement of the problem, of the FOREL algorithm searches for such a partition of m objects into k clusters so that the criterion of similarity F to the center for all clusters is minimal. Let us denote the coordinates of the center of the j-th cluster by $C_j$. The sum of distances $p(C_j, a_i)$ between the center of a sphere with radius $R_0$ and all $m_j$ points $a_j$ of this cluster is $p_j=\sum p(C_j,a_j)$, where $i=1\ldots m_j$ and the sum of such internal distances for all k clusters is $F=\sum p_j$, $j=1\ldots k$.

https://doi.org/10.15405/epct.23021.45
Corresponding Author: F. G. Ahmatshin
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2672-8834

For the modified FOREL-2 algorithm, the clustering quality function F uses an additional criterion f corresponding to the value of the exact number of clusters

$$F = f\left(k_i\right)\sum_{j=1}^{k} p_j\,, \tag{2}$$

where $f(k_i)=1$, if $k_i=k$, or $f(k_i)=\infty$, if $k_i \neq k$. The best option corresponds to the minimum value F.

The condition for joining a point to a cluster $x \in K_j$ is determined by the formula

$$p\left(C_j, a_i\right) \leq R_0 R\,, \tag{3}$$

where R is the radius of the search for local concentrations.

For the modified FOREL-2 algorithm, it is necessary to obtain a precisely specified number of clusters. In all variations of the FOREL algorithm, there is a basic procedure coinciding with the k-means algorithm at steps 1, 2, 3 and 4. Step 5 is different because all observations of test trials occurring inside the sphere are selected according to (3), as well as in the method for determining the convergence of the objective function (2).

## 3. Research Questions

The following questions were posed during the study:

- How to choose the search radius for local concentrations in the FOREL-2 clustering algorithm with an initial number of clusters?
- How to improve the accuracy and stability of the result, such as identifying homogeneous batches of industrial products?

## 4. Purpose of the Study

The answers to these questions will lead us to find the best accuracy, according to the obtained Rand index with a larger value of the objective function in comparison with the k-means algorithm.

At the same time, being compared to the classical implementation of the k-means algorithm, the use of greedy heuristics in the FOREL-2 algorithm shows greater efficiency.

## 5. Research Methods

Orlov and Fedosov (2016) used the data from the results of tests conducted at the test center in order to isolate presumably homogeneous batches with some accuracy. For two types of integrated circuits (ICs), 140UD25A and 140UD26A, belonging to different homogeneous batches, the initial data is a combination of parameters and consists of 18 batches (9 batches of the first and second IC types, respectively). Products (devices) in each batch are described by 18 input measured parameters. For the first batch of IC 1, 3, 8 and 9 data vectors were selected in the amount of n=807. For the type of the ICs, data vectors of batches 4, 5, 7 and 9 were selected in the amount of n=532. The data set was decimated, discarded every 2, 3 and 4 data

https://doi.org/10.15405/epct.23021.45
Corresponding Author: F. G. Ahmatshin
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2672-8834

on the test results, the number was n=201 and n=132 for the first and second types of the ICs, respectively. In the work of Ahmatshin and Kazakotsev (2020), the selected test data were normalized to the standard deviation and to the allowable values of the parameter.

The FOREL-2 algorithm was compared with the k-means algorithm on a system with 8 GB of RAM, an Intel Core i3-3220 processor with a frequency of 3.3 GHz, and an Ubuntu 18.04.6 LTS operating system. We conducted 30 experiments for each algorithm. All results are presented with an average of 30 runs. Each launch lasted 1, 10 and 60 seconds.

The average results of clustering on the speed of the algorithms are given in Table 1: maximum (max), minimum (min), median value (median), average value (avr) and standard deviation (st.dev) for the Rand index and objective function. For the objective function, the coefficient of variation (var) and the span factor (spn) are also calculated.

**Table 1.** Accuracy of clustering 1, 3, 8, and 9 batches of ICs 140UD25A data

| | k-means | | | FOREL-2 | | |
|---|---|---|---|---|---|---|
| time, sec. | 1 | 10 | 60 | 1 | 10 | 60 |
| Mixed batch (n=201) with normalization by standard deviation | | | | | | |
| Rand index | | | | | | |
| max | 0.605 | 0.605 | 0.605 | 0.794 | 0.778 | 0.799 |
| min | 0.55 | 0.542 | 0.563 | 0.605 | 0.641 | 0.656 |
| mean | 0.596 | 0.593 | 0.591 | 0.712 | 0.711 | 0.741 |
| st.dev | 0.01 | 0.014 | 0.01 | 0.052 | 0.039 | 0.04 |
| Objective function | | | | | | |
| max | 85.1 | 80.6 | 77.2 | 85 | 81 | 78.4 |
| min | 66.9 | 66.8 | 66.8 | 76.1 | 72.5 | 72.2 |
| mean | 69.4 | 69 | 69.7 | 81.1 | 77.1 | 75.1 |
| st.dev | 3.5 | 2.9 | 3.2 | 2.5 | 2 | 1.3 |
| var | 0.051 | 0.041 | 0.046 | 0.03 | 0.026 | 0.017 |
| spn | 18.2 | 13.8 | 10.4 | 9 | 8.5 | 6.2 |
| Mixed batch (n=807) with standard deviation normalization | | | | | | |
| Rand index | | | | | | |
| max | 0.6 | 0.599 | 0.6 | 0.76 | 0.777 | 0.781 |
| min | 0.557 | 0.58 | 0.581 | 0.539 | 0.599 | 0.67 |
| mean | 0.584 | 0.594 | 0.595 | 0.669 | 0.718 | 0.738 |
| st.dev | 0.013 | 0.004 | 0.004 | 0.063 | 0.045 | 0.026 |
| Objective function | | | | | | |
| max | 326.5 | 296.9 | 298.4 | 411.6 | 348.2 | 336.7 |
| min | 273 | 269.7 | 269.8 | 325 | 323.5 | 312.7 |
| mean | 296.5 | 274 | 273.8 | 365 | 335.4 | 324.3 |
| st.dev | 14.6 | 6 | 5.9 | 21.1 | 7 | 5.7 |
| var | 0.049 | 0.022 | 0.022 | 0.058 | 0.021 | 0.018 |
| spn | 53.5 | 27.3 | 28.6 | 86.5 | 24.8 | 24 |

https://doi.org/10.15405/epct.23021.45
Corresponding Author: F. G. Ahmatshin
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2672-8834

## 6. Findings

The approach using the FOREL-2 algorithm with a greedy heuristic for choosing the radius of the search for local concentrations demonstrated the best accuracy, according to the obtained Rand index with a larger value of the objective function in comparison with the k-means algorithm.

For solving the problem of clustering electronic radio products based on test data, the accuracy and speed of the software implementation of the algorithm are quite acceptable. Compared to the classical implementation of the k-means algorithm, the use of greedy heuristics in the FOREL-2 algorithm showed greater efficiency.

## 7. Conclusion

When solving clustering problems, FOREL-2 algorithm often gives a result which is very far from the optimal solution. In this research, we aimed at developing not only fast but also the most accurate algorithm, based on greedy heuristics for selection radius of local concentrations, for solving problems of clustering electronic radio products based on test data, using the FOREL-2 algorithm.

Computational experiments show that the use of the greedy heuristics for the radius selection of local concentrations in the FOREL-2 algorithm improves the accuracy and speed of obtaining results at identifying electronic radio products. Moreover, the best results can be shown by FOREL-2 algorithm compared to the k-means algorithm used medoids as intermediate centres of spheres.

## Acknowledgments

The work was supported by the Grant of the President of the Russian Federation for state support of the Leading Scientific Schools) NSh-421.2022.4.

## References

Ahmatshin, F. G., & Kazakotsev, L. A. (2020). Impact of data normalization methods and clustering model in the problem of automatic grouping of industrial products. *Journal of Physics: Conference Series, 1679*(3), 032085. https://doi.org/10.1088/1742-6596/1679/3/032085

Anderberg, M. R. (1973). *Cluster Analysis for Applications.* Academic Press.

Bobrowski, L., & Bezdek, J. C. (1991). c-means clustering with the l/sub l/ and l/sub infinity / norms. *IEEE Transactions on Systems, Man, and Cybernetics, 21*(3), 545-554. https://doi.org/10.1109/21.97475

Hossain, M. Z., Akhtar, M. N., Ahmad, R. B., & Rahman, M. (2019). A dynamic K-means clustering for data mining. *Indonesian Journal of Electrical Engineering and Computer Science, 13*(2), 521-526. https://doi.org/10.11591/ijeecs.v13.i2.pp521-526

Kazakovtsev, L. A., Antamoshkin, A. N., & Masich, I. S. (2015). Fast deterministic algorithm for EEE components classification. *IOP Conference Series: Materials Science and Engineering, 94,* 012015. https://doi.org/10.1088/1757-899x/94/1/012015

Klosgen, W., & Zytkow, J. M. (1996). Knowledge discovery in databases terminology. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, & R. Uthurusamy (Eds.), *Advances in Knowledge Discovery and Data Mining* (pp. 573-592). AAAI Press/The MIT Press.

Li, Y., & Wu, H. (2012). A clustering method based on K-means algorithm. *Physics Procedia, 25,* 1104-1109. https://doi.org/10.1016/j.phpro.2012.03.206

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE transactions on information theory, 28*(2), 129-137. https://doi.org/10.1109/tit.1982.1056489

MacQueen, J. (1967). Classification and analysis of multivariate observations. In 5th Berkeley Symp. Math. Statist. Probability (pp. 281-297). University of California.

Milligan, G. W., & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika, 50*(2), 159-179. https://doi.org/10.1007/bf02294245

Na, S., Xumin, L., & Yong, G. (2010). Research on k-means clustering algorithm: An improved k-means clustering algorithm. In 2010 Third International Symposium on intelligent information technology and security informatics (pp. 63-67). IEEE. https://doi.org/10.1109/iitsi.2010.74

Orlov, V. I., & Fedosov V. V. (2016). *ERC clustering dataset*. http://levk.info/data1526.zip

Pérez-Ortega, J., Almanza-Ortega, N. N., & Romero, D. (2018). Balancing effort and benefit of K-means clustering algorithms in Big Data realms. *PLOS ONE, 13*(9), e0201874. https://doi.org/10.1371/journal.pone.0201874

Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Yu, P. S., Zhou, Z.-H., Steinbach, M., Hand, D. J., & Steinberg, D. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems, 14*(1), 1-37. https://doi.org/10.1007/s10115-007-0114-2

Zagoruiko, N. G. (1999). *Applied methods of data and knowledge analysis.* Novosibirsk: Institute of Mathematics SD RAS.