**Clustering with empty clusters.**

1. **Abstract**

Cluster data analysis is used in many applied tasks of machine learning and data analysis: medicine, sociology, economics, cybersecurity. The absence of observations is not always the absence of information, so it is assumed that the presence of empty clusters can tell us about our data, as well as real observations. The study of empty clusters has already been implicitly considered in the literature, although with certain limitations. The aim of the work is to start with a basic one-dimensional case and propose an algorithm that performs one-dimensional clustering for a given number of clusters and demonstrates an area from an empty cluster. An algorithm is proposed to determine potential empty clusters and their sizes depending on the initial division of the sample into the number of clusters. A method is implemented to fill in these gaps and estimate the displacement of the centroids of the initial clustering when taking into account an empty cluster.

1. **References and sources.**

The review article (Xu D., Tian Y. A., 2015) discusses current SOTA approaches to cluster analysis. More classical algorithms are considered (Raschka, S. and Mirjalili, 2015) in chapter 11 of the book on machine learning. Empty clusters can be explained in different ways, depending on the problem under consideration. Often, the initialization of an empty cluster is a feature of the implementation of the original algorithm and this case is considered as a disadvantage of the method, for example, k–means, and such cases are tried to minimize due to the modernization of classical algorithms. This situation may occur when the correct initial conditions are not met and the optimizer finds a local minimum, or when the input data is represented by binary or categorical features (Raykov Y. P. et al., 2015). (Yadav A., Dhingra S., 2014, Pakhira M., 2009) proposes a modification of the k-means algorithm that does not initialize empty clusters. (Hua C. et al., 2019) also propose an alternative to the k-means method, the genetic XK-Means, which also does not initialize empty clusters. (Tavallali, P., Singhal, M., 2021) note that the initialization of empty clusters is detected as a result of the application of the random forest algorithm. An article is devoted to the development of a visualization tool for dot scatter diagrams with the allocation of empty clusters (Giesen J. et al., 2015)

To account for the effect of the presence of empty clusters, an implementation procedure is proposed (Audigier et al., 2021). His idea is to first populate the dataset with observations corresponding to otherwise unidentifiable clusters. The disadvantage of the approach is that it requires prior knowledge of where new observations are being added. Two articles are also devoted to the problem of empty clusters (Forina M. et al., 2003) in which the authors propose a statistical test to assess the quality of agglomerative clustering, and also propose an index of the informativeness of "empty spaces", taking into account the greatest distance between clusters. (McGee G. et al., 2020) investigate long-term health effects as a result of external influences and note that approaches using estimation equations necessarily exclude empty clusters and, therefore, give biased estimates of marginal effects.

1. **Methodology**

To test the hypothesis about the algorithm's operability, a test data set of 30 observations was used, set manually, to explicitly highlight a skip, or, in our notation, an empty cluster. Next, evenly distributed noise was added to this set. For subsequent testing, the make\_blobs data generator of the sklearn library was used. For convenience, all generated data has been converted to an integer type. In the one-dimensional version, such data sets can be interpreted as mixtures of normal distributions with a given mode and standard deviation. Since our task assumes a manually set number of clusters, we used the following methods to find the optimal initial separation: Distortion Measure (Elbow-method), Silhouette Analysis, Calinski Harabasz, Gap-statistics.

The paper proposes to consider a three-step procedure for initializing and filling an empty cluster and assessing the impact of its presence. It is assumed that those observations that fell into a different cluster from the first one during the new cluster partitioning should be subject to additional verification. The procedure consists of the following steps:

• at the first step, it is proposed to estimate the range of the number of clusters;

• at the second step, based on the entire range of the number of clusters from the previous step and their characteristics (minimum, maximum, average value within the cluster, the number of observations in the cluster), the size of the empty cluster and its characteristics are estimated – the estimated centorid, the number of observations and the spread of values. A data set with the characteristics of an empty cluster is generated;

• at the third step, the clustering is reinitialized with the addition of an "empty" cluster to the source data set and the clustering results are compared with the first one. The observations whose cluster label differs from the first clustering procedure are highlighted.

1. **Results and the novelty**

An algorithm for allocating an empty cluster and generating it based on the expected characteristics is implemented. When varying the size of an empty cluster with constant modes and standard deviation, it is noted that if: its size and width are less than the average of the original ones, then the centroids practically do not shift relative to the original ones, if equal or greater, then vice versa, there is a shift of the centroids and more objects get a new label. Thus, the algorithm for allocating and generating an empty cluster should be parametric.

Further research suggests:

• Add empty clusters not only inside the existing space, but also outside the boundaries of the original space;

• Develop an algorithm for multidimensional clustering with different metrics of the distance between clusters and their boundaries.

1. **Literature**
2. Xu D., Tian Y. A Comprehensive Survey of Clustering Algorithms // Ann. Data. Sci. 2015. Т. 2. № 2. С. 165–193.
3. Raschka, S. and Mirjalili, V. Python Machine Learning. Machine Learning and Deep Learning with Python, scikitlearn, and TensorFlow 2. Packt, 3rd edition, 2019
4. Yadav A., Dhingra S. An Enhanced K-Means Clustering Algorithm to Remove Empty Clusters // 2016. Т. 4. № 4. С. 7.
5. Pakhira M. A Modified k-means Algorithm to Avoid Empty Clusters // International Journal of Recent Trends in Engineering. 2009. Т. 1.
6. Raykov Y. P. и др. What to Do When K-Means Clustering Fails: A Simple yet Principled Alternative Algorithm // PLOS ONE. 2016. Т. 11. № 9. С. e0162259.
7. Hua C. и др. A Genetic XK-Means Algorithm with Empty Cluster Reassignment // Symmetry. 2019. Т. 11. № 6. С. 744.
8. Tavallali, P., Singhal, M. K-means tree: an optimal clustering tree for unsupervised learning. The Journal of Supercomputing, 77:5239–5266, 2021.
9. Giesen J., Kühne L., Lucas P. Sclow Plots: Visualizing Empty Space // Computer Graphics Forum. 2017. Т. 36. № 3. С. 145–155.
10. Audigier V., Niang N., Resche-Rigon M. Clustering with missing data: which imputation model for which cluster analysis method? // 2021.
11. Forina M., Casolino C., Lanteri S. Cluster analysis: significance, empty space, clustering tendency, non-uniformity. I--Statistical tests on the significance of clusters // Ann Chim. 2003. Т. 93. № 1–2. С. 55–68.
12. Forina M., Lanteri S., Casolino C. Cluster analysis: significance, empty space, clustering tendency, non-uniformity. II--Empty Space index // Ann Chim. 2003. Т. 93. № 5–6. С. 489–498.
13. McGee G. и др. Informatively empty clusters with application to multigenerational studies // Biostatistics. 2020. Т. 21. № 4. С. 775–789.