Example of Clustering Using K-Means Method in Python

Jelena Milenković[*], Milica Pavlović[*], Vojkan Nikolic[*], Aleksandar Jasic[**], Velibor Premceski[**]

[*] *University of Criminal Investigation and Police Studies, Zemun, Republic of Serbia*

[**] *UNIVERSITY OF NOVI SAD , TECHNICAL FACULTY "MIHAJLO PUPIN" ZRENJANIN, ,REPUBLIC OF SERBIA*
vojkan.nikolic@kpu.edu.rs, aleksandar.jasic96@gmail.com , velci.aspire@gmail.com

Abstract: The goal of clustering is to sort the data by similarity, based on a predefined number of clusters. We will demonstrate an example of clustering as such. The idea is to implement an application that, based on the input data from Excel spreadsheets, will display results after clustering. In particular, we worked on random data, which we enter from two tables. The results will be presented in 2D as well as in the 3D model.

***Key words**: clustering, k-mean, division, similarities*

I. INTRODUCTION

One of the most basic characteristics of human beings is the ability to group similar objects into groups, the so-called. classification. The grouping of similar objects into categories dates from the very inception of the planet earth, when the first humans had to distinguish different objects that had some of the same or similar characteristics. Classification can be defined as the process of classifying subjects in a field of science into classes, groups of subjects having a common characteristic that differ from other groups in that characteristic, whereas cluster analysis or clustering is the main task of research data retrieval and a common technique for statistical analysis. data, which is used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, data compression, and computer graphics.

II. THE CONCEPT OF CLUSTERING

Cluster analysis is the task of grouping a set of objects in such a way that the objects within the cluster are similar to each other and thus different from the objects of other clusters. Unlike classification, we do not have an "exact" solution here:

- Algorithm performance evaluation is much more difficult than classification
- The suitability of the solution depends on the domain and the application case
- one and the same solution can be evaluated differently in different application cases
- requires the involvement of domain experts to evaluate the solution

The cluster analysis process consists of two basic steps:

1) selection of an appropriate measure of distance (similarity),

2) the choice of a clustering algorithm, that is, a series of procedures for grouping elements so that there are small differences within the cluster and large between clusters. There are different algorithms for solving clustering problems. However, there is no objectively the best algorithm for clustering, because a particular algorithm can produce good results on one dataset and bad on another because clustering depends on the dimensionality, structure and type of data. There are hierarchical and non-hierarchical methods, including the k-mean method, which is the subject of our research. Non-hierarchical sub-clustering methods, more reliable than hierarchical ones, assume that the number of clusters known in advance, or as with some methods, varies during the clustering process.

III. K-MEANS METHOD

The k-means method is one of the simplest, but also the most well-known classification algorithms. Over time, several algorithms have been developed that deal with the clustering process, such as X-means, Kohonen SOM, DB Scan, hierarchical cluster algorithms, and certainly the most popular is the K-means algorithm. K-means is an algorithm that groups data into K clusters, where the number of K clusters is determined in different ways and depending on the preferences of the decision maker. Because it is difficult to determine what is the true number of clusters in the data, the algorithm is most often implemented multiple times, so based on the measure of cluster quality or on the basis of confirmation of cluster quality

by the decision maker decides that the result is satisfactory.

The idea behind partitioning a dataset is to provide a partition of n objects in k disjunct clusters. By definition, an object can belong to one cluster and each cluster must have at least one object (otherwise they would have fewer than k clusters). The classification algorithm is usually iterative; the initial step is usually improved in each subsequent step as long as there are improvements. Defining the initial partition requires a priori specification of the number of clusters. Suppose that a "measure" of how good a partition is represented by a function J whose value is reduced as far as possible to achieve further optimization of results. A general algorithm for all methods of this type would be:

1. Determine the initial partition in the k cluster and derive a value for the function,

2. Change the partition to reduce the value as much as possible, leaving k unchanged,

3. If new reductions are not possible, the process will stop and the number of clusters existing at that moment will be the final number of clusters. Otherwise, we go back to the previous step.

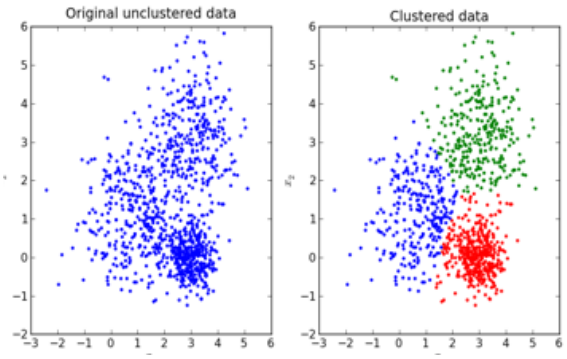
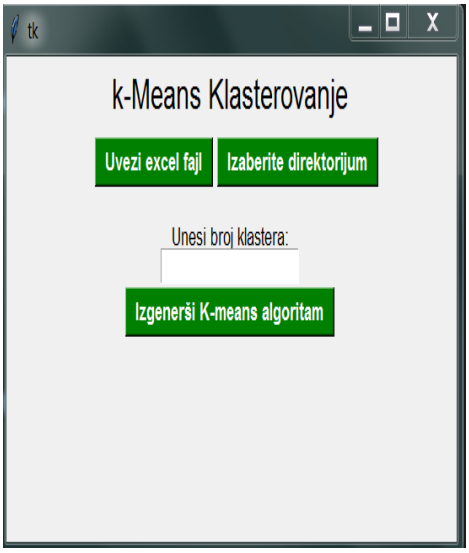


Figure 1: Example of data before and after clustering

IV. GRAPHICAL USER INTERFACE (GUI)

At the very beginning, we will create a user interface with the Tkinter library. We will specify the appropriate width and height of the GUI. When the application starts, a new window opens. The window will contain one label (where we enter the name of the application). Below the name will be two buttons that import data (one button enters a file, the other a directory with multiple files). Then, the user is allowed to enter the desired number of clusters in a specific field. Finally, we will have another button that will generate the k-means algorithm and display the data graphically.

Figure 2: Main window layout



V. 2D MODEL DISPLAY

Clicking the "Import Excel file" button opens the file dialog, window for selecting a file that is called via the getExcel function. The pandas library reads an Excel file from the given columns, in this case x and y. In the event that the file cannot be loaded in the specified manner, an error is thrown.

When we select the desired file, enter the desired number of clusters in the Enter cluster number field. The field is blank by the time the number is entered. When the number of clusters is taken, the code defines the variable that k-means will make. Receives these clusters as a parameter and loads the document. A variable is created that will contain the centers of the cluster, as well as a label that will list what the centers are. After that, the real canvas that will accept that figure is to define the figure that is currently blank and adjust the appearance of that figure. The x and y axes from the file are defined, the label in the k-means itself is adjusted, as well as the color settings, and then the centroids around which the data will be distributed. After that, pressing the "Generate K-means algorithm" button results in a clustering result in the form of a graphical 2D view, using the getKmeans function, which will be shown in Figure 3.

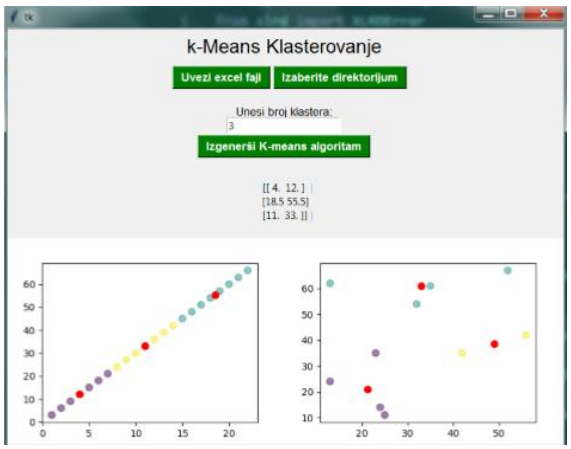


Figure 3. 2D display of clustering results

VI. 3D MODEL DISPLAY

To get a 3D view of the results, we must first add one more column of z to our tables. The clustering process is done in the same way as for 2D display, you just need to make some changes to the code. The left side of the next image shows a clustering result of data entered from a single file, while the right side of the image shows a clustering result imported from a directory.

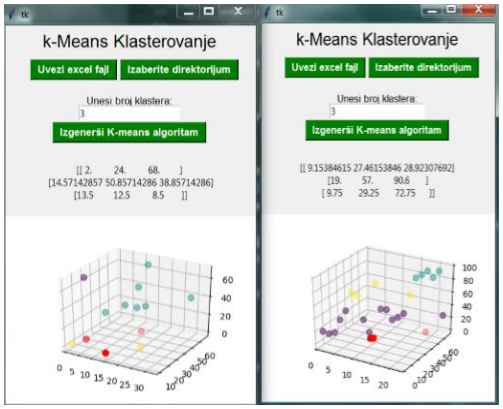


Figure 4. 3D view of clustering results

CONCLUSION

Cluster analysis is widely used in many scientific fields. Cluster analysis techniques are used to find clusters. groups, in an a priori unclassified multivariate data set. Although cluster analysis techniques are very useful for data analysis itself, they require careful attention in order to avoid the wrong solutions. Many cluster analysis methods have been developed and numerous studies have shown that there is no best method, but it depends solely on what we want to get. we show. In order to successfully apply the cluster analysis technique, it is necessary to study well the data from which the conclusion will be drawn and the direction in which the analysis itself will go in order for the appropriate model to be applied. Clustering respects the orderliness of nature itself, where objects do not tend to be distributed randomly and evenly. Nature has the property of grouping. Clustering is performed when classes for object sorting are not known in advance. What makes clustering even more possible is to reduce the number of cases being analyzed by treating clustered cases in a cluster equally, and it is sufficient to analyze only the representative of each cluster, etc. As clustering is applied in more and more fields (psychology and other social sciences, biology, statistics, mechanical learning, data research, etc.), existing algorithms need to be refined and new algorithms with less time and space complexity need to be refined. To achieve this, it is also necessary to discover new techniques and data structures that will be used in algorithms and contribute to their efficiency.

REFERENCES:

[1] Suknovic M. , Delibašić B. Poslovna inteligencija i sistemi za podršku odlučivanju, FON, Belgrade, 2010.

[2] Bradley, P. S., Bennett, K. P., & Demiriz, A. (2000). Constrained k-means clustering (Technical Report MSR-TR-2000-65). Microsoft Research, Redmond,WA.

[3] Jain, A. K., & Dubes, R. C. (1988). Algorithms for clustering data. Prentice Hall.

[4] E. Rasmussen, “Clustering Algorithms”, in Information Retrieval Data Structures and Algorithms, Frakes and Baeza-Yates (Eds.), pp. 419-442, New Jersey: Prentice Hall, 1992.

[5] Lucke, J., & Forster, D. (2019). k-means as a variational EM approximation ofGaussian mixture models.Pattern Recognition Letters, 125, 349e356.https://doi.org/10.1016/j.patrec.2019.04.001

[6] Yen, H., Park, S., Arnold, J. G., Srinivasan, R., Chawanda, C. J., Wang, R. Y.,...Zhang, X. S. (2019). IPEAT plus : A built-in optimization and automatic cali-bration tool of SWAT.Water, 11(8).https://doi.org/10.3390/w11081681. ARTN168https://www.displayr.com/what-is-k-means-cluster-analysis/ (08.01.2020.)

[7] Zou, Q., Cui, P., He, J., Lei, Y., & Li, S. S. (2019). Regional risk assessment of
debrisflows in China-An HRU-based approach.Geomorphology, 340, 84e102. https://doi.org/10.1016/j.geomorph.2019.04.027.

[8] Lipor, J., & Balzano, L. (2020). Clustering quality metrics for subspace clustering .Pattern Recognition, 104, 107328.https://doi.org/10.1016/j.patcog.2020.107328.

[9] Sheshukov, A. Y., Douglas-Mankin, K. R., Sinnathamby, S., & Daggupati, P. (2016).Pasture BMP effectiveness using an HRU-based subarea approach in SWAT.Journal of Environmental Management, 166,2 76e284.https://doi.org/10.1016/j.jenvman.2015.10.023

[10] Lotz, T., Opp, C., & He, X. (2018). Factors of runoff generation in the Dongting Lakebasin based on a SWAT model and implications of recent land cover change.QuaternaryInternational,475,54e62.https://doi.org/10.1016/j.quaint.2017.03.05

[11] Tian, K., Li, J. H., Zeng, J. F., Evans, A., & Zhang, L. N. (2019). Segmentation of tomatoleaf images based on adaptive clustering number of K-means algorithm.Computers and Electronics in Agriculture, 165.https://doi.org/10.1016/j.com-pag.2019.104962. ARTN 104962

[12] Mengistu, A. G., van Rensburg, L. D., & Woyessa, Y. E. (2019). Techniques for cali-bration and validation of SWAT model in data scarce arid and semi-aridcatchments in South Africa.Journal of Hydrology-Regional Studies, 25.https://doi.org/10.1016/j.ejrh.2019.100621. UNSP 100621.

[13] Guo, T., Engel, B. A., Shao, G., Arnold, J. G., Srinivasan, R., & Kiniry, J. R. (2019).Development and improvement of the simulation of woody bioenergy crops inthe Soil and Water Assessment Tool (SWAT).Environmental Modelling&Soft-ware, 122, 104295.https://doi.org/10.1016/j.envsoft.2018.08.030