Mesopotamian journal of Computer Science
Vol. 2024, **pp**. 104–109
DOI: https://doi.org/10.58496/MJCSC/2024/008 ; ISSN: 2958-6631
https://mesopotamian.press/journals/index.php/cs

Research Article

Data Mining Utilizing Various Leveled Clustering Procedures on the Position of Workers in a Data Innovation Firm

Hussein Alkattan[1,2, *], Noor Razzaq Abbas[3], Oluwaseun A. Adelaja[4], Mostafa Abotaleb[5], Guma Ali[6],

[1] *Department of System Programming, South Ural State University, Chelyabinsk, Russia*

[2] *Directorate of Environment in Najaf, Ministry of Environment, Najaf, Iraq*

[3] *Al-Furat Al-Awsat Technical University, Najaf, Iraq,*

[4] *Information Communication and Technology Department, Lagos State University, Lagos, Nigeria*

[5] *Intelligent Systems and Machine Learning Lab, Shenzhen 518000, China*

[6] *Department of Computer and Information Science, Muni University, Arua, Uganda*

ARTICLE INFO

Article History

Received 04 May 2024
Accepted 05 Jul 2024
Published 26 Jul 2024

Keywords

Data Mining

Clustering Algorithm

Python

Clustering Analysis

ABSTRACT

The objective of this paper is to clarify dynamic clustering, the divisive and agglomerative dynamic clustering techniques. It fundamentally centers on the concept of the divisive different leveled shapes as well known as the top-down approach by creating a workflow appear, dendrograms, clustered data table which accumulated the clusters based the chosen property, and appear the isolated between each cluster with the assistance of an data mining device called Python. The DIANA dynamic approach utilized data tests of the list of laborers in a Data Advancement firm to induce clusters from the position column inside the data test table. In this work, we in addition executed genuine infers by creating bar chart that shows up the ages of the chosen agent sets plotted against the positions which are the Engineers, Assistants, Workers and Troughs.

1. INTRODUCTION

Data mining is an integral part of modern management and decision-making methods in organizations. In recent years, the use of cluster analysis has become one of the main techniques for identifying structures and patterns in large amounts of data [1]. This becomes especially important for companies involved in data innovation, where a deep understanding of internal and external processes is required to improve operational efficiency and strategic planning [2]. Level clustering, or hierarchical cluster analysis, is a data clustering technique that allows you to identify hidden structures at different levels of nesting [3]. This method is widely used in various fields including marketing, bioinformatics, and social network analysis [4-6]. In the context of human resource management, grouping levels can help identify key groups of employees based on different characteristics such as professional skills, productivity and engagement [7]. For companies involved in data innovation, knowing and understanding internal employee structures becomes crucial. For example, identifying groups of high-performing employees allows you to improve recruitment and retention processes [8]. In addition, level analysis allows us to identify groups that need additional training or support, which helps create a more productive work environment [9]. Using different grouping procedures on employee positions in a company can also help improve communication and coordination between different departments [10]. This is especially important in an innovative business environment, where synergy between teams plays a key role in the successful implementation of projects [11].

The purpose of this paper is to explore the application of different-level clustering procedures to analyze data related to employees' jobs in a data innovation company. In particular, we will consider clustering and partition clustering techniques, as well as their practical application for identifying large groups of employees [12-14]. The advantages of combinatorial methods lie in their ability to create hierarchical structures that allow flexible control of the level of detail of analysis [15]. On the other hand, partitioning methods can be more effective for large data sets, as they can quickly divide the data into groups [16]. It is important to note that choosing the appropriate methodology depends on the specific objectives of the analysis and the characteristics of the data [17]. To conduct the study, we will use real data about company employees, including information about positions, skills, experience, and performance [18]. The analysis will be performed using

*Corresponding author. Email: alkattan.hussein92@gmail.com

specialized cluster analysis software such as Wika, which will allow us to obtain reliable and reproducible results [19]. As a result of this study, we expect to obtain a deep understanding of the structural characteristics of the company's employees, which will help improve management and decision-making processes [20]. The results may also be useful for other data innovation companies seeking to improve their operational efficiency and competitiveness [23-21]. Therefore, data mining using different aggregation procedures at company employee job levels is a promising tool for improving the efficiency of personnel management and strategic planning [22-24][31][32].

2. RELATED WORK

Odelia Yim and Kylie presented a stepwise cluster analysis framework using SPSS software. They focused on this method, in which clustering is done incrementally by systematically combining similar groups using connectivity and exclusion measures. They also compared different measures of connectivity (single connectivity, complete connectivity, average connectivity) and related them to mental information to achieve accurate results [6].

Tian Zhang and colleagues conducted a study on the BIRCH (Reduced Iterative Reduction and Pecking Classification) strategy, which is an ideal agglomerative clustering strategy for large databases. BIRCH has been proven to be most effective when dealing with large and chaotic collections in databases, allowing the creation of a single collection using a single information filter while maintaining quality. The effectiveness of BIRCH has been evaluated through numerous experiments conducted by the creators, demonstrating high efficiency in managing input data and organizing groups [27].

Vera M.B considered the idea of clustering and used stepwise agglomerative calculation as a tool for extracting information from the capital market in order to study the Bulgarian stock exchange and identify global trends in the behavior of traded shares. She also created a dendrogram of stock groups using conventional connectivity measures [26].

Anna S. and her colleagues proposed progressive probabilistic clustering strategies used in supervised and unsupervised learning for information extraction. Their probabilistic clustering is based on generalized Gaussian mixing, and the proposed approach relies on agglomerative clustering with L2 removal [28].

In their study, Sudipto G. and colleagues presented a stepwise calculation that handles groups of non-spherical shapes and with wide variations in discretization. This approach combines irregular scanning and partitioning to work with large databases, where clusters are defined by sparse points and then reduced towards the center of the clusters [29].

Fathi HS and colleagues compared different agglomeration calculations based on assessing the quality of clusters created using different advanced agglomeration strategies. Their research results showed that agglomerative calculation, using I1 as a model for selecting clusters, produces higher group quality compared to other methods based on entropy and virtue measures as external criteria[30].

3. HIERARCHICAL CLUSTERING STRUCTURE

Partitioning and agglomerative clustering don't require a preparatory determination of the number of clusters k, but for them it is fundamental to decide a strategy for calculating the separate between clusters. As appeared in Figure 1, the top-down approach from root to leaf is known as various leveled divisive clustering (DIANA), whereas the switch strategy, from leaf to root, is known as agglomerative progressive clustering (AGNES). Dendrograms resulting from the method of divisive clustering reflect diverse levels of pecking order and the degree of likeness at each level. Within the case of agglomerative clustering, the method begins with one information point and slowly combines two or more clusters. In differentiate, divisive progressive clustering begins with a expansive cluster and successively isolates it into littler ones [25].

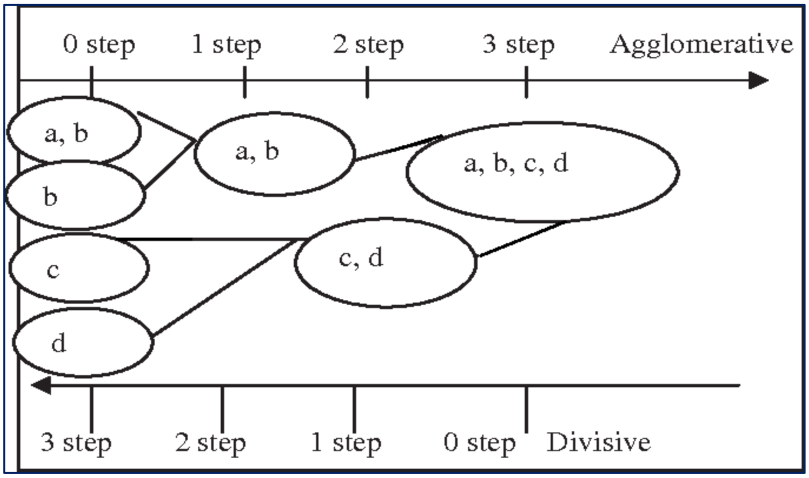

Fig. 1. Hierarchical Clustering Techniques

4. DENDROGRAM

Dendrogram might be a chart that talks to the dynamic relationship between challenge. From the dynamic clustering method in Figure 2, the dendrograms are gotten inside the taking after portion get ready plan 1-8 for the levels:

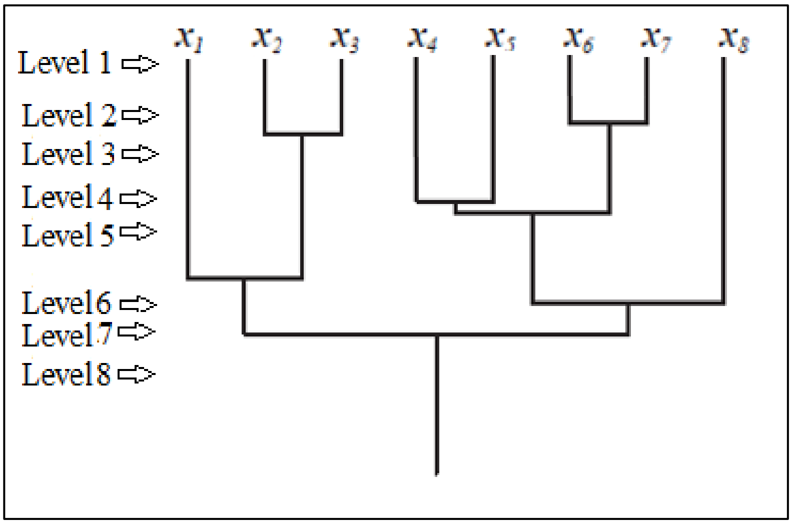

Fig. 2. Dendrogram appearing the Progressive Part Relationship

4.1 Clustering Algorithm

A cluster clustering algorithm is an unsupervised machine learning technique that is used to group data into groups or clusters, such that the data in each cluster is similar to each other and different from the data in other clusters. Cluster clustering can be used in many fields such as data analysis, pattern learning, dimensionality reduction, and market segmentation.

Types of cluster clustering algorithms

- Hierarchical Clustering
- K-means clustering
- Density based clustering (DBSCAN)

Hierarchical Clustering

Hierarchical cluster clustering builds a hierarchy of clusters. This hierarchy can be either agglomerative or divisive. In the clustering method, each element starts out in a separate cluster, and the clusters gradually merge until all the elements are combined into one large cluster. In the partitioning method, all the elements start in one large cluster and are gradually divided until all the elements are in separate clusters.

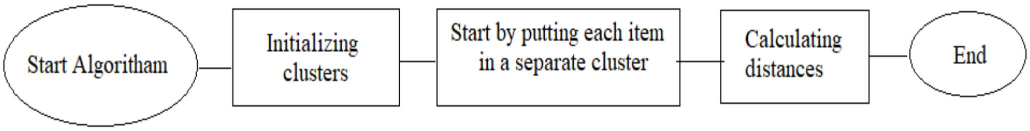

Fig. 3. Show the algoritham structure.

To Calculate the distances between all clusters using a certain distance criterion such as the (Euclidean Distance):

$$\sqrt{{}^{2}(\ {}_{jk}x_{ik} - x)\sum_{k=1}^{n}} = d(i,j) \qquad (1)$$

Where $d(i,j)$ is the distance between element i , element j , $_{jk}x$ and x_{ik} They are the feature values k of elements i and j respectively.

After merging, calculate the new distances between the new cluster and the rest of the clusters:

$$d(i,j)\sum_{j\in B}\sum_{i\in A}\frac{1}{|A||B|}=d(A,B) \qquad (2)$$

Where A and B are clusters, $|A|$ and |B| are the number of elements in the clusters.

5. RESULTS

We created a DataFrame containing employee names, their roles, and their scores in two variables (Score1 and Score2).

Choosing the variables used in the analysis:

We chose the variables Score1 and Score2 to perform the cluster analysis.

Scale variables:

We used StandardScaler to standardize variables.

Implementing hierarchical clustering:

We used the linkage function from the scipy.cluster.hierarchy library to implement the hierarchical cluster clustering algorithm using the “rose” method.

Create a tree diagram:

We used the dendrogram function to create a dendrogram to display the results.

Thus, we have prepared and implemented the hierarchical cluster clustering algorithm and presented the results using a tree diagram.

The dendrograms and partitions gotten from the clusters portion for the examiner, Company positions after analyzing the clusters and the thrust ID for each position are showed up in figure 3 below:

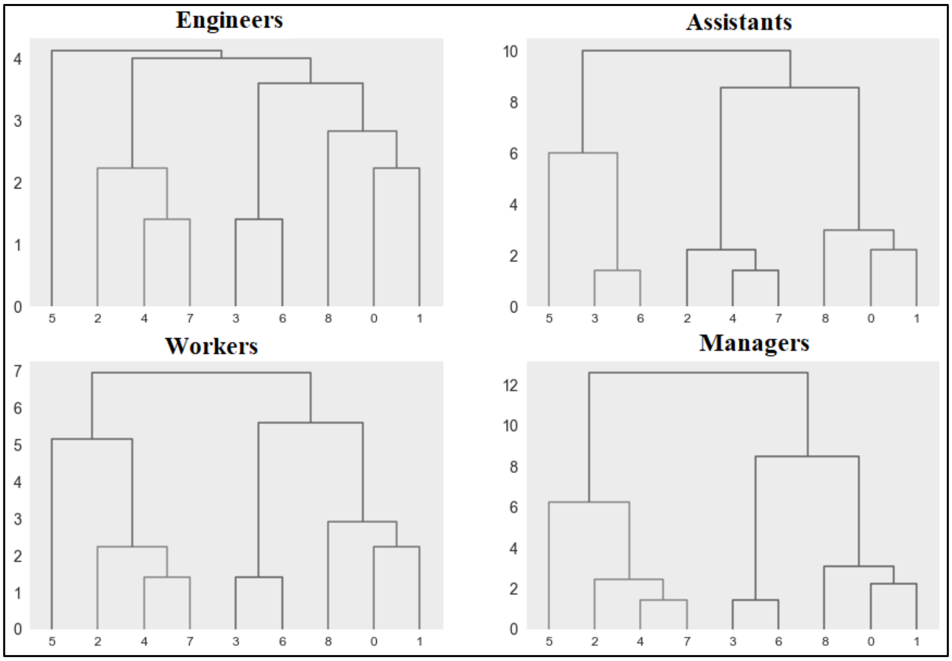

Fig. 3. Dendrogram of Clusters for Enginners, Assistants, Workers and Managers Position.

Genuine infers through bar chart was utilized to look at the ages the agents and the position from the clustered data table was as well actualized as shown in figure 4 below:

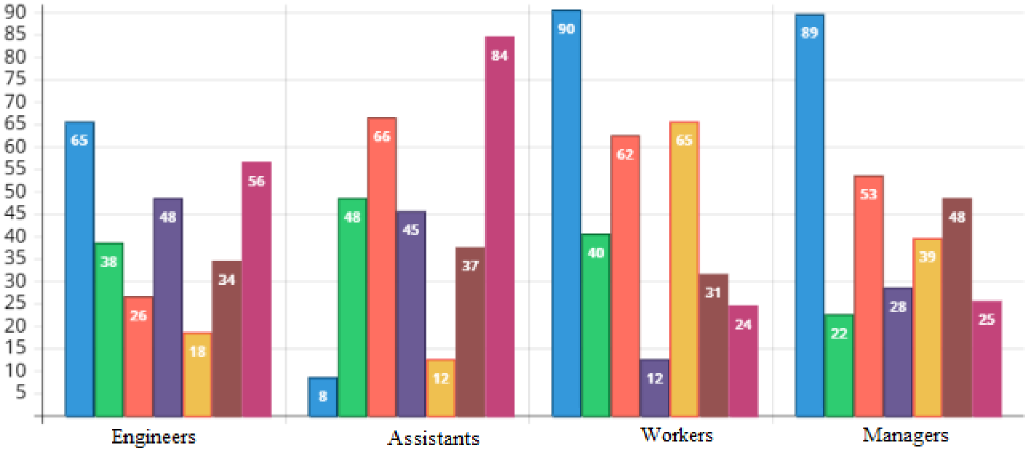

Fig. 4. Statistical clusters for each position.

6. CONCLUSION

The cluster algorithm is one of the basic tools in the field of data analysis and machine learning, as it plays an important role in organizing and classifying data based on its similarities and characteristics. Through techniques such as K-Means and Hierarchical Clustering, these algorithms are able to reveal patterns and structures hidden in large, complex data. The importance of the cluster algorithm lies in its ability to simplify data and group it into meaningful groups, making it easier to analyze and make decisions based on accurate insights. Moreover, it contributes to improving the quality of data used in modeling and forecasting, by reducing noise and dividing it into homogeneous groups. However, there remain challenges associated with clustering algorithms, such as determining the optimal number of clusters and dealing with heterogeneous data. Therefore, it is necessary to continuously develop and update algorithmic technologies to ensure that they are maximized in various practical applications. In conclusion, it can be said that the cluster algorithm is a powerful and useful tool in the field of data analysis, allowing practitioners and researchers to understand data in new ways and achieve valuable results in various fields.

Funding

The authors had no sponsor backing.

Conflicts Of Interest

None.

Acknowledgment

The authors extend appreciation to the institution for their unwavering support and encouragement during the course of this research.

References

[1] S. P. Lloyd, "Least squares quantization in PCM," IEEE Trans. Inf. Theory, vol. 28, no. 2, pp. 129-137, 1982. doi: 10.1109/TIT.1982.1056489.

[2] T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: An efficient data clustering method for very large databases," in Proc. 1996 ACM SIGMOD Int. Conf. Management of Data, 1996, pp. 103-114. doi: 10.1145/233269.233324.

[3] A. K. Jain, "Data clustering: 50 years beyond K-means," Pattern Recognit. Lett., vol. 31, no. 8, pp. 651-666, 2010. doi: 10.1016/j.patrec.2009.09.011.

[4] L. Kaufman and P. J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis. Hoboken, NJ, USA: Wiley, 2009. doi: 10.1002/9780470316801.

[5] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques. Burlington, MA, USA: Morgan Kaufmann, 2011. doi: 10.1016/C2009-0-61819-5.

[6] G. Gan, C. Ma, and J. Wu, Data Clustering: Theory, Algorithms, and Applications. Philadelphia, PA, USA: SIAM, 2007. doi: 10.1137/1.9780898718348.

[7] R. Xu and D. Wunsch, Clustering. Hoboken, NJ, USA: Wiley-IEEE Press, 2008. doi: 10.1002/9780470382783.

[8] S. Qureshi, A. A. Shaikh, and S. Qureshi , Trans., “Modified Technique to Solve Degeneracy in Linear Programming Problems”, Babylonian Journal of Mathematics, vol. 2024, pp. 73–77, Sep. 2024, doi: 10.58496/BJM/2024/009.

[9] S. Theodoridis and K. Koutroumbas, Pattern Recognition. Burlington, MA, USA: Academic Press, 2008. doi: 10.1016/B978-1-59749-272-0.X0001-1.

[10] D. Chakrabarti and C. Faloutsos, "Graph mining: Laws, generators, and algorithms," ACM Comput. Surv., vol. 38, no. 1, pp. 2-69, 2006. doi: 10.1145/1132952.1132954.

[11] M. Ester, H. P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in Proc. 2nd Int. Conf. Knowledge Discovery and Data Mining, 1996, pp. 226-231. doi: 10.5555/3001460.3001507.

[12] S. Guha, R. Rastogi, and K. Shim, "CURE: An efficient clustering algorithm for large databases," in Proc. 1998 ACM SIGMOD Int. Conf. Management of Data, 1998, pp. 73-84. doi: 10.1145/276304.276312.

[13] M. R. Anderberg, Cluster Analysis for Applications. New York, NY, USA: Academic Press, 1973. doi: 10.1016/B978-0-12-057650-0.50007-0.

[14] Sara salman Qasim and Lubna Mohammed Hasan , Trans., “Mining Utilities Itemsets based on social network”, *BJN*, vol. 2024, pp. 25–30, Mar. 2024, doi: 10.58496/BJN/2024/004.

[15] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, Introduction to Algorithms. Cambridge, MA, USA: MIT Press, 2009. doi: 10.5555/1614191.

[16] A. K. Jain and R. C. Dubes, Algorithms for Clustering Data. Upper Saddle River, NJ, USA: Prentice-Hall, 1988. doi: 10.5555/59541.

[17] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," J. Comput. Appl. Math., vol. 20, pp. 53-65, 1987. doi: 10.1016/0377-0427(87)90125-7.

[18] H. H. Bock, "Origins and extensions of the k-means algorithm in cluster analysis," Electron. J. Hist. Probab. Stat., vol. 13, no. 2, pp. 246-258, 2016. doi: 10.1214/16-EJHPS29.

[19] C. Aggarwal and C. Zhai, "A survey of text clustering algorithms," in Mining Text Data, Boston, MA, USA: Springer, 2012, pp. 77-128. doi: 10.1007/978-1-4614-3223-4_4.

[20] D. H. Fisher, "Knowledge acquisition via incremental conceptual clustering," Mach. Learn., vol. 2, no. 2, pp. 139-172, 1987. doi: 10.1007/BF00058659.

[21] R. T. Ng and J. Han, "CLARANS: A method for clustering objects for spatial data mining," IEEE Trans. Knowl. Data Eng., vol. 14, no. 5, pp. 1003-1016, 2002. doi: 10.1109/TKDE.2002.1033770.

[22] T. K. Asman and H. D. Abd, “Directed Mining of the Dark Web Using Hell9: Advanced Techniques for Exploring and Analyzing Anonymous Networks”, SHIFRA, vol. 2024, pp. 105–112, Jul. 2024, doi: 10.70470/SHIFRA/2024/012.

[23] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in Proc. 5th Berkeley Symp. Math. Statist. Probab., vol. 1, 1967, pp. 281-297. doi: 10.1093/biomet/73.1.45.

[24] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," in Proc. 18th Annu. ACM-SIAM Symp. Discrete Algorithms, 2007, pp. 1027-1035. doi: 10.5555/1283383.1283494.

[25] Chris D. and X. He, "Cluster merging and splitting in hierarchical clustering algorithms," unpublished.

[26] V. M. B., "Using the agglomerative method of hierarchical clustering as a data mining tool in capital market," Int. J. Inf. Theor. Appl., vol. 15, pp. 382-386, 2008.

[27] Z. Tian, R. Raghu, and L. Miron, "BIRCH: An efficient data clustering method for large databases," in Proc. 1996 ACM-SIGMOD Int. Conf. Management of Data, 1996, pp. 103-114.

[28] A. Szymkowiak, L. Jan, and L. K. H., "Hierarchical clustering for data mining," unpublished.

[29] S. Guha, R. Rastogi, and K. Shim, "CURE: An efficient clustering algorithm for large databases," in Proc. 1998 ACM SIGMOD Int. Conf. Management of Data, 1998, pp. 73-84.

[30] [30] F. H. S., O. I. E., and R. E. A., "Comparison of hierarchical agglomerative algorithms for clustering medical documents," Int. J. Softw. Eng. Appl., vol. 3, no. 3, pp. 1-15, May 2012.

[31] H. M. Saleh, “A Comprehensive Review of Data Mining Techniques for Diabetes Diagnosis Using the Pima Indian Diabetes Dataset”, EDRAAK, vol. 2024, pp. 39–42, Apr. 2024, doi: 10.70470/EDRAAK/2024/006.

[32] A. J. Ramadhan, B. A. Kumar, I. Bala, M. M. Mijwil, M. Abotaleb, H. Alkattan, and Z. Albadran, "IoT-integrated multi-sensor plant monitoring and automated tank-based smart home gardening system," in BIO Web Conf., vol. 97, pp. 1-11, Apr. 2024.