Deployment of Hierarchical clustering model for Data Analysis

R. Raj Kumar

Assistant Professor, School of Competitive Coding Koneru Lakshmaiah Education Foundation Vaddeswaram, Guntur District- 522302, Andhra Pradesh, India, Email : reddyrajkumar@kluniversity.in, rajkumarreddy123@gmail.com

Received: 13.04.2024 Revised : 18.05.2024 Accepted: 20.05.2024

ABSTRACT

Data retrieval concerns have seen a range of adaptive applications of grouping. Currently, clustering constitutes one of the more actively researched and developed fields. The goal of clusters is to identify the collection of major groups in which members of all of them are closer linked to the other versus members of other categories. The ensuing clusters may provide a framework for arranging huge text collections for effective surfing and searchable. Finding naturally occurring clusters or collections among numerous dimensions using an indicator of similarity is known as dataset clustering. A basic phenomenon across a wide range of fields is clustering. As a result, the topic of clustering is being extensively investigated by scholars from several domains. If conclusions or understanding that may be obtained through the data is unable to deduced, then the information has become meaningless. According to a set of standards, the clustering technique divides facts into significant, useful, or both groups (clusters) according to common traits. Data analysis in the fields of algorithmic learning, computational biology, statistics, and detection of patterns, to name a few, has involved employing the techniques of clusters and segmentation. This paper focuses on the concept of clustering. It discusses about the Hierarchical Clustering algorithm. Eventually, it provides applications of clustering methods.

Keywords: Clustering, Data Mining, Hierarchical Clustering, unsupervised and data analytics.

1. INTRODUCTION

One of the main areas of data mining is groupings, which additionally serves as a crucial technique for splitting or combining data. These days, cluster is used in many fields, including biology, online categorization, market evaluation, and trade.[1] The goal of the statistical mining approach is to extract pertinent details from large, unwieldy data sets and transform it into a form that is suitable for further use. In analysis and data mining uses, cluster is an important activity. It involves arranging a collection of items so that members of the same group have a stronger relationship with one another than do members in other groupings (clusters). Mining information may be accomplished by following different stages. Both supervised as well as unsupervised learning may be used in mining. Unsupervised learning is what the phenomenon of clustering is. Substantial superior clusters that have substantial intra-class and little inter-class resemblance are the result of a well-designed method for clustering.[2]

2. LITERATURE REVIEW

According to Pritika Talwar et al., the work in their paper describes the concept of clustering, how clustering is used in machine learning, an importance of clustering, clustering algorithm and types of clustering algorithm such as Hierarchical Clustering, Partitioning-based clustering, Density-based clustering, Model-based clustering, Fuzzy clustering and their respective clustering algorithms. The paper also describes the application of clustering model such as Banking and Healthcare etc. Further, it describes difficulties such as sensitivity to initialization, handling of missing values, and high-dimensional data. Further, it concludes the research gaps of fixing the difficulties by preventive measures by selecting appropriate methods or techniques selection and execution.[5]

According to Anil K. Jain, the paper discusses the fundamental concepts of clustering, its evolution over the past decades, and various clustering algorithms. It provides insights into the objectives, methods, and applications of clustering in data analysis and pattern recognition. This foundational reference is highly regarded in the field of clustering and serves as a valuable resource for understanding the theoretical underpinnings and practical applications of clustering techniques.[6]

According to Yogita Rani et al., the work of their paper elaborates the data mining, hierarchical clustering algorithm, CURE (acronym Clustering using representatives), BIRCH (acronym Balanced Iterative Reducing and Clustering using Hierarchies), ROCK (acronym Robust Clustering using links), CHEMELEOM Algorithm, Linkage Algorithms, Leaders-Subleaders and Bisecting K-Means. It is concluded with the challenges of existing algorithm. It is overcoming the problems with the modified algorithm of their work.[7]

According to Oyewole et al., the work in their paper describes the concept of clustering and its use cases. The paper focuses on the Components and classifications for data clustering, Clustering techniques such as Pattern representation, Clustering or grouping process, Performance evaluation, Clustering classification and Clustering algorithms. The paper also classified the Continuation of selected clustering algorithms based on identified clustering.[9]

3. Clustering In Machine Learning

The technical description of clustering includes the arrangement of things that occurs when the associations between the elements in the provided data are either unknown or poorly understood. The goal of clusters is also to highlight any fundamental classifications that exist in the information. In addition, clusters is a method for classifying unlabeled information into separate categories with minimal or no guidance.[3] Clustering, which is a potent method in analysing information and machine learning, holds the key to the solution. We may organize points of information according to their level of commonalities using clustering algorithms, which helps with a variety of activities like visualization and consumer decomposition.[4] The diagram materializes to exemplify the vital components and collaborations surrounded by the field of Machine Learning.

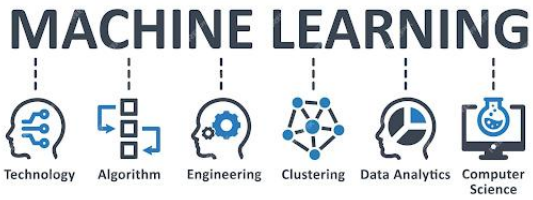


Fig 1. Clustering in Machine Learning[4]

3.1 Components of Machine Learning

3.1.1 Technology: This is probably a representation for the software, gadgets, as well as coding languages that support learning from machines. It depicts an individual's skull with designs resembling circuits. [5]

3.1.2 Algorithm: This is represented by interlocked squares, which stand for the fundamental mathematical concepts and algorithms that underlie artificial intelligence procedures.

3.1.3 Engineering: The technical parts of automated learning, such as system planning, advancement, and execution, are represented by this illustration of an individual's head fitted with gearing.

3.1.4 Clustering: This represents the approaches to clustering employed by algorithmic learning to put pertinent information in an arrangement by showing them as linked nodes.

3.1.5 Data Analytics: This illustrates the function of statistical analysis in comprehending as well as deriving conclusions from information to achieve automated learning goals, using a diagram of a pie inside the human being's skull as a metaphor.

3.1.6 Computer Science: This highlights how technological ideas and techniques form the basis of automated learning, as demonstrated by a computer interface attached to a vial.

3.2 Relationships

It is implied that every one of those parts collaborate in the discipline of computational learning by the dashed paths that link them and creating a structure of interconnection.

4. Clustering Approach

4.1 Unsupervised Learning: The technique used for segmentation usually doesn't need data with labels for instruction, making it an unsupervised teaching problem. Rather, the process pinpoints innate patterns or connections within the information. [6]

4.2 Similarity or Distance Measure: To establish the extent to which connected items are, methods of clustering employ an analogy or distance metric. The distance of Euclid, the cosine proximity, as well as correlation factors are examples of frequently used measurements.

4.3 Grouping into Clusters: A group of clusters to operate, each made up of items more identical to one another than to those found in different groupings, is the result of a method for clustering.
4.4 Applications: Numerous domains, including the mining of data, artificial intelligence, recognition of patterns, visualization, computational biology, and knowledge search, heavily rely on grouping. It is useful for systems of suggestions, division, identifying anomalies, and interactive study of data. All things considered, clustering is an essential method for identifying hierarchy in data, providing understanding of the connections and properties of items inside intricate datasets.

5. Hierachical Clustering Technique

The goal of the cluster assessment technique known as the hierarchical clustering is to create an ordered sequence of groupings. Once a combination or division of decisions was successfully carried out and a pure top-down clustering technique's efficiency is compromised by lack of ability to make adjustments. After that, it won't reverse prior actions or switch items across clusters. Therefore, if a merging or split choice is made poorly at any point, it might result in clusters that are not very high quality. Combining hierarchy-based clustering with additional approaches to perform several phase aggregations is a possible way to enhance the level of grouping produced by hierarchy approaches.[7]

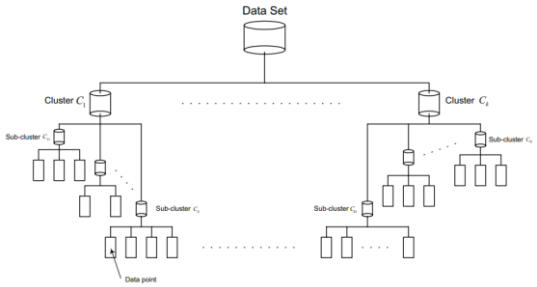


Fig 2. Pictorial illustration of Hierarchical Clustering [8]

Grouping a collection of data so as to optimize intracluster resemblance while organizing groups within a natural order is the aim of the clustering method. The process consists of two stages: (1) clustering points of data forming a potential subcluster; plus (2) combining the subclusters based on an identity score. Cluster or instances of data inside an identical category appear more identical to one another versus those in distinct groups at all stages of the hierarchical.

5.1 Categories of Hierarchical Clustering

It is categorized into 2 primary forms of hierarchical grouping exists as follows:
5.1.1 Agglomerative hierarchical clustering: This clustering algorithm follows a Bottom to Upward strategy. The figure 3 at first, we designate every single point with a separate group. The nearest pair of groups is then marked at every repetition, and we continue in this manner till only one group remains. Therefore we combine the nearest clusters at every phase. Thus, another name for it is hierarchy additives clustered.

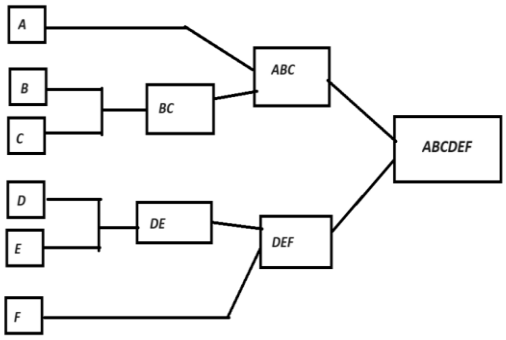


Fig 3. Agglomerative Hierarchical Clustering[10]

5.1.2 Divisive Hierarchical clustering: It's an organized method called the top to Bottom Strategy. The figure 4 operates in an opposing manner. Rather than beginning using 'n' groupings, it begins with an individual cluster that contains every point. We separate the group's furthest member at every repetition

and keep going till every cluster has just one point within it. Divided grouping by hierarchy gets its name from the fact that we separate (or divide) groups at every step.

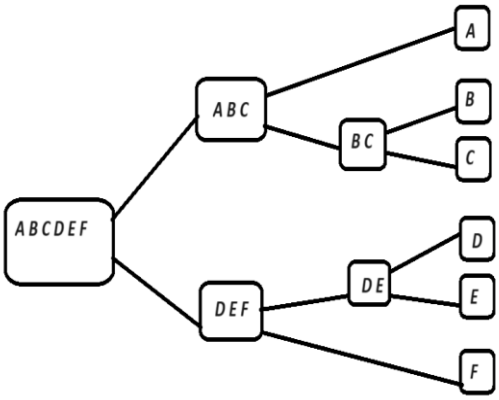


Fig 4. Divisive Hierarchical Clustering[10]

6. Deploying Workflow

Deploying a clustering model involves several considerations and steps to ensure its effectiveness in real-world applications. Here's a brief literature review on key aspects of deploying clustering models:[11]

6.1 Algorithm Selection: Different clustering algorithms suit different types of data and objectives. K-means clustering is widely used due to its simplicity and efficiency, while hierarchical clustering offers flexibility in cluster shapes. Density-based methods like DBSCAN are useful for identifying clusters of varying shapes and densities. Literature often compares these algorithms in terms of scalability, performance metrics, and suitability for different data types.[12]

6.2 Preprocessing: Data preprocessing plays a crucial role in clustering accuracy and efficiency. Techniques such as normalization, standardization, and handling missing values are frequently discussed in the literature. Effective preprocessing enhances clustering quality by making data suitable for algorithms that rely on distance measures.[13]

6.3 Evaluation Metrics: Assessing clustering quality is challenging due to the absence of ground truth labels. Internal evaluation metrics like silhouette score and Davies-Bouldin index are commonly used to measure compactness and separation of clusters. External metrics, such as Adjusted Rand Index (ARI), are employed when ground truth labels are available.[14]

6.4 Scalability: Scalability is a critical factor for deploying clustering models in large datasets or real-time applications. Literature often discusses parallel and distributed clustering algorithms, as well as techniques for reducing computational complexity and memory usage.[15]

6.5 Visualization: Visualizing clustering results helps interpret and communicate insights from data. Literature covers techniques namely t-SNE (acronym t-Distributed Stochastic Neighbor Embedding) and PCA (acronym Principal Component Analysis) for visualizing high-dimensional data and clustering structures effectively.[16]

6.6 Deployment Considerations: Deploying clustering models involves considerations such as model persistence, integration with existing systems (e.g., databases), and updating models with new data. Real-world case studies and literature reviews often discuss challenges and best practices for deployment in various domains. [17]

6.7 Applications: Clustering finds applications across domains such as customer segmentation, anomaly detection, and recommendation systems. Literature reviews often highlight successful applications, along with specific challenges and adaptations required for different use cases. [18]

6.8 Algorithmic Enhancements: Research in clustering continues to propose algorithmic enhancements, such as robustness to noise, handling high-dimensional data, and incorporating domain-specific constraints. Literature reviews cover advancements and comparative studies to evaluate these enhancements. [19]

6.9 Ethical and Legal Considerations: Deploying clustering models also involves ethical considerations related to privacy, bias, and interpretability. Literature discusses guidelines and frameworks for responsible deployment and the implications of clustering results on decision-making. [20]

6.10 Future Directions: Finally, literature reviews often suggest future research directions, such as hybrid approaches combining clustering with other machine learning techniques, improving interpretability of clustering results, and addressing challenges in dynamic and streaming data environments. [21]

7. Hierarchical Clustering Algorithm

A distance/similarity metric is used in clustering structure to form new groups. The following summarizes the steps involved in the two categories of Hierarchical Clustering:[22][23]

7.1 Agglomerative algorithm

Step 1: Utilizing a certain distance measure, calculate a closeness matrix.
Step 2: The cluster is allocated to every single point of data.
Step 3: Combine the groups using the level of similarity among them.
Step 4: Make an imaginary distance matrices updating
Step 5: Continue from Steps 3 through 4 until there is just one cluster left.

7.2 Divisive Algorithm

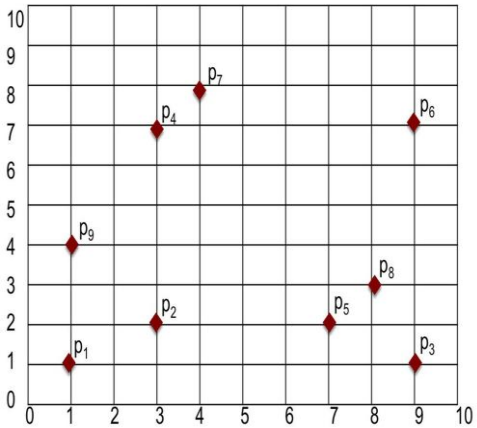
Step1: Think of every data point as belonging to a distinct group.
Step 2: Divide into groups utilizing standard flat-clustering such as K-Mean
Step 3: To divide the groups further, select the most promising cluster by determining which one contains the biggest Sum of Squared Error (SSE).
Step 4: The second and third steps should be repeated until only one cluster remains.

7.3 Calculating a adjacency matrix: Making an impartial matrix is the initial stage in the procedure. A measurement algorithm connecting every couple of items is applied to determine the numerical contents of the matrices. This technique is frequently performed using the Euclidean distance calculation. For an information set of n items, the closest neighbor matrix's architecture will be as below. The separation measurements across pi as well as pj are shown here as d(pi,pj).[23]

Table 1. Measurements across PI and PJ

	p1	p2	p3	...	Pn
p1	d(p1,p1)	d(p1,p2)	d(p1,p3)	...	d(p1,pn)
p2	d(p2,p1)	d(p2,p2)	d(p2,p3)	...	d(p2,pn)
p3	d(p3,p1)	d(p3,p2)	d(p3,p3)	...	d(p3,pn)
...	...	...	...	...	...
Pn	d(pn,p1)	d(pn,p2)	d(pn,p3)	...	d(pn,pn)

7.4 Correlations between cluster: Whether to modify the nearness matrix and figure out the spread between groups is the central topic in the hierarchy of clustering. The solution to such question is provided by several methods. Every strategy has benefits and drawbacks. The decision will be based on the density of the data points, the presence of noise within the information set, and the circularity or non-circularity of the grouping shapes. An example using numbers will assist to clarify the approaches and decisions. We'll make use of a little sample data set (Figure 5) that has a mere nine a two-dimensional feature.



Point	x	y
p_1	1	1
p_2	3	2
p_3	9	1
p_4	3	7
p_5	7	2
p_6	9	7
p_7	4	8
p_8	8	3
p_9	1	4

Fig 5. Sample Data

As seen in Figure 6, let us assume that this particular set of data contains two distinct clusters. The method used to determine the separation among the groups varies. The following is a list of popular techniques.

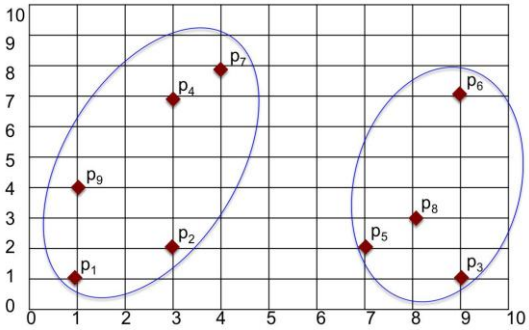


Fig 6. Two clusters

7.5 Min (Single) Linkage: Determining the smallest distance among locations inside clusters can be a means of measuring the spread amongst them. In simple terms, we can determine which location in the initial cluster is closest to which position in the second group and figure out how far apart they are. P2 in a single group as well as P5 in another are the nearest positions in Diagram 7. D(p2,p5)=4 is the amount of time between the points in question and, thus, the length between groups.

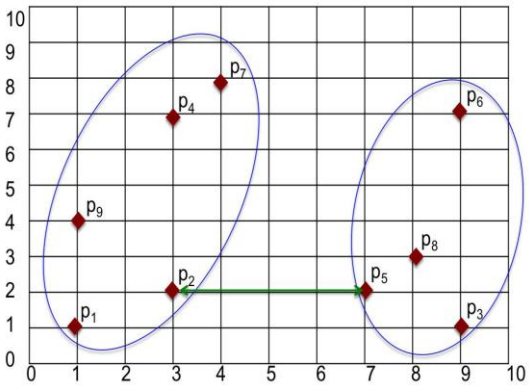


Fig 7. Min Linkage Method

7.6 Max (Complete) Linkage: Finding the greatest length among the edges of both clusters provides an additional method of measuring the separation. Those locations in each group that are the farthest apart may be located, and their distances can be computed. The longest stretch in Figure 8 is found at points P1 as well as P6. The measurement d(p1,p6)=10 represents the amount of time across the two places and, consequently, the length across groups.

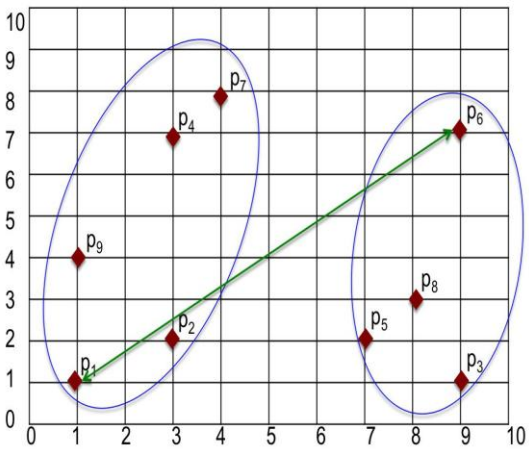


Fig 8. Max Linkage Method

7.7 Centroid Linkage: According to the Centroid technique, the separation across clusters is determined by the spread across their central points. A distance algorithm is used to calculate the length between every cluster's centers once the center of gravity for every group has been determined.

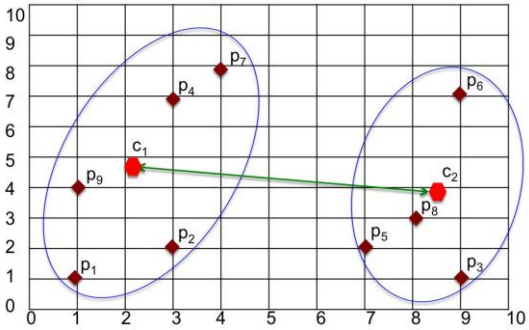


Fig 9. Centroid Linkage Method

7.8 Average Linkage: According to the Mean technique, the mean pairwise range among each pair of points inside a group represents the distance that exists among them. In the following figure, just a portion of the lines that link pairs of points are displayed for clarity.

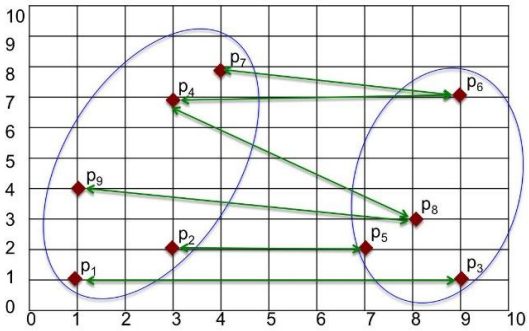


Fig 10. Average Linkage Method

8. Services Of Hierarchical Clustering

In statistical analysis, structured clustering is a popular approach that groups items into clusters according to how similar they are. This clustering technique has several benefits and may be used to a wide range of issues. Ten benefits of structured clustering are as follows:[24]

1. Sturdiness
2. Simple to understand.
3. Adaptable.
4. Expandable.
5. Illustration.
6. Adjustable.
7. Simpler to use.
8. Enhanced precision.
9. The absence of linearity.
10. Numerous output levels.

CONCLUSION

This paper mainly discusses about the concept of data dining and clustering mechanism. It elaborates the conceptualization of Hierarchical Clustering as well as sub categories. It explains the idea of Agglomerative Hierarchical Model and Divisive Model with its algorithmic implementation procedures with the data analysis. It also converse about the Hierarchical clustering applications and advantages. Further, there is a scope of research for Hierarchical clustering in the field of Emerging Trends.

ACKNOWLEDGMENT

I am sincerely thankful to 6th International Conference on Engineering and Advancement in Technology (ICEAT 2024), September 27th & 28th, 2024 organized by Malla Reddy College of Engineering, Secunderabad, India in collaboration with Samarkand State University, Uzbekistan and Manipal University College, Malaysia for providing me with the opportunity to write a research paper on the topic "Deployment of Hierarchical clustering model for Data Analysis". I am also thankful to my colleagues for the constant encouragement. Then, I would like to thank my family members and friends for their kind cooperation and support in completion of the research paper.

REFERENCES

[1] Fang, Chu, and Haiming Liu. 2021. "Research and Application of Improved Clustering Algorithm in Retail Customer Classification" Symmetry 13, no. 10: 1789. https://doi.org/10.3390/sym13101789.

[2] Amandeep Kaur Mann & Navneet Kaur, Review Paper on Clustering Techniques, Global Journal of Computer Science and Technology Software & Data Engineering Volume 13 Issue 5 Version 1.0 Year 2013

[3] Oyewole, G.J., Thopil, G.A. Data clustering: application and trends. Artif Intell Rev **56**, 6439–6475 (2023). https://doi.org/10.1007/s10462-022-10325-y

[4] https://www.analyticsvidhya.com/blog/2023/11/types-of-clustering-algorithms-in-machine-learning/

[5] Pritika Talwar, Shubham, Komalpreet Kaur, EXPLORING CLUSTERING TECHNIQUES IN MACHINE LEARNING, IJCRT, Volume 12, Issue 3, March 2024, ISSN: 2320-2882

[6] Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters, 31*(8), 651-666.

[7] Yogita Rani & Dr. Harish Rohil, A Study of Hierarchical Clustering Algorithm, International Journal of Information and Computation Technology. ISSN 0974-2239 Volume 3, Number 11 (2013), pp. 1225-1232.

[8] Xudong Wang and Vassilis L. Syrmos, Optimal Cluster Selection Based on Fisher Class Separability Measure, 2005 American Control Conference June 8-10, 2005. Portland, OR, USA

[9] J. Oyelade et al., "Data Clustering: Algorithms and Its Applications," 2019 19th International Conference on Computational Science and Its Applications (ICCSA), St. Petersburg, Russia, 2019, pp. 71-81, doi: 10.1109/ICCSA.2019.000-1.

[10] https://www.kaggle.com/code/varinderkm/hierarchical-clustering-mall-customers-dataset

[11] Dasgupta, S., & Long, P. M. (2005). A feature-based approach to modeling and predicting quality in cluster analysis. *Management Science, 51*(4), 612-624.

[12] Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20*, 53-65.

[13] Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence, (2)*, 224-227.

[14] Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD '96)*.

[15] Maaten, L. van der, & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research, 9*, 2579-2605.

[16] Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374*(2065), 20150202.

[17] Zhou, A., & Tao, D. (2018). Machine learning on big data: Opportunities and challenges. *Neurocomputing, 237*, 350-361.

[18] Dua, D., & Graff, C. (2019). UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science.

[19] Bezdek, J. C. (1999). Pattern recognition with fuzzy objective function algorithms. *Plenum Press*.

[20] Aggarwal, C. C., & Reddy, C. K. (2013). Data clustering: Algorithms and applications. *CRC Press*.

[21] Mittelstadt, B. D., & Floridi, L. (2016). The ethics of big data: Current and foreseeable issues in biomedical contexts. *Science and Engineering Ethics, 22*(2), 303-341.

[22] https://builtin.com/machine-learning/agglomerative-clustering

[23] https://www.learndatasci.com/glossary/hierarchical-clustering/

[24] https://codinginfinite.com/hierarchical-clustering-applications-advantages-and-disadvantages/#google_vignette

[25] Zhao, K., Yu, H., & Wang, F. (2020). A review of clustering algorithms for big data: Taxonomy, characteristics, and open issues. *IEEE Access, 8*, 158720-158739.

[26] G. Nathiya, S.C. Punitha and Dr. M. Punithavalli, (IJCSIS) International Journal of Computer Science and Information Security, Vol. 7, No. 3, March 2010

[27] Omran, Mahamed & Engelbrecht, Andries & Salman, Ayed. (2007). An overview of clustering methods. Intell. Data Anal.. 11. 583-605. 10.3233/IDA-2007-11602.