Hierarchical Clustering Algorithms: A Comprehensive Review and Analysis

Shaik Soniya

PG Scholar, Dept. of Computer Science Sri Venkateswara University, Tirupati

Abstract— *Hierarchical clustering is a fundamental data analysis technique widely used in various fields, including data mining, pattern recognition, and bioinformatics. In this research paper, we investigate the application of hierarchical clustering algorithms on shopping data to explore customer segmentation patterns. The dataset comprises customer transaction records from an online retail platform, containing information on customer demographics, purchase history, and shopping preferences. We conduct a comprehensive study by employing various hierarchical clustering techniques, including agglomerative and divisive methods, with different distance metrics and linkage criteria. Through our experiments and analysis, we aim to identify meaningful customer segments and gain insights into their shopping behaviors and preferences.*

I. INTRODUCTION

Numerous information mining and AI applications going from PC vision to science issues have as of late confronted a blast of information. Information mining is a cycle used to transform crude information into helpful data. Different procedures and calculations have been utilized for extricating significant data from enormous informational collections. Grouping is one of the fundamental procedures for examination in information mining. It is a course of collection comparable information things together. There are numerous calculations utilized in grouping. Different calculations can be utilized for Bunch examination that be different expressively in their impression of makes a gathering likewise how to capably find them [1]4[2]. Pervasive thoughts of bunches contain bunches with minor distances between the gathering objects, information space in thick regions, explicit mathematical conveyances [3]. In bunch examination we look for designs in an informational index by gathering the multivariate perceptions into groups. The objective is to find an ideal gathering for which the perceptions or articles inside each group are comparative yet the bunches are not at all like each other [6]. Bunch examination is a more crude procedure in that no suspicions are made concerning the quantity of gatherings or the gathering structure. Gathering is finished based on similitudes or distances.

The gatherings can hence communicated as a multi-objective improvement trouble. As a result it has become progressively vital to create powerful, exact, vigorous to commotion, quick, and general grouping calculations, open to designers and scientists in a different scope of regions. In progressive bunching the objective isn't to track down a solitary dividing of the information, however an order (by and large addressed by a tree) of segments which might uncover fascinating construction with regards to the information at different degrees of granularity. The most broadly utilized progressive strategies are the agglomerative bunching methods.

II. METHODOLOGY

Progressive strategies are among the customary procedures of group examination. They comprise in progressive collection or division of the perceptions and their subsets. Coming about because of this sort of system there is a tree-like construction, which is alluded to as dendrogram. The agglomerative methods start from the arrangement of perceptions, every one of which is treated as a different group. Bunches are collected as per the diminishing level of closeness (or the rising level of divergence) until one, single group is laid out [4][5].

III. HIERARCHICAL TECHNIQUES

Progressive bunching calculations are utilized to develop the various leveled relationship among information things to frame groups. At the point when the data on different degrees of bunch structure is required, these calculations work effectively to decipher results. It consolidates different levels in a consecutive of steps. The consequence of progressive bunching can be graphically displayed in a tree like design called dendrogram.

Progressive calculations make bunches recursively by isolating an information base D of N objects into various degrees of settled parceling, indicated by a dendrogram [7]. A dendrogram is a two layered chart or tree and gives a total progressive portrayal of how items are like each other on various levels. It very well may be inspected at a specific level to address an alternate grouping of the information [8]. There are two kinds of progressive calculations: agglomerative calculations and

disruptive calculations. Agglomerative calculations fabricate the tree base up, for example consolidating the N objects into gatherings. Disruptive calculations develop the tree base by isolating the N objects into better groups [5]. Base up or agglomerative grouping, the more normally utilized procedure, regards each item as a bunch of size 1.

3.1 Agglomerative Strategy

It's also called as AGNES (Agglomerative Settling). It stirs in a base up way. i.e, each thing is at first as a free gathering (leaf). At each step the things are consolidated which are two groups that local unit the preeminent comparable neighborhood unit vanquished into a substitution bigger bunch (hubs) [4][5]. This cycle is repeating till entire things are in a solitary gigantic bunch (root).

3.2 Disruptive technique

It's named as DIANA (Divise Examination) it chips away at the standard of big picture perspective. It is a converse methodology of AGNES. It begins with the root with all thing are engaged with a solitary group. During each emphasis, the most different bunch is partitioned into two gatherings here one gathering is called left bunch and one more gathering is called as right group. This cycle will go on till objects are in their singleton [4][5].

IV. EXPERIMENTAL STUDY

The examinations have been facilitated by utilizing Python programming vernacular. The Python Scikit-learn is a pack for information depiction, social occasion and depiction. We have considered customer-shopping data for experimentation. The dataset contains 200 instances and 5 attributes. Before applying hierarchical clustering, we conducted thorough data preprocessing steps to handle missing values, normalize numerical features, and encode categorical variables. Moreover, feature selection and dimensionality reduction techniques were applied to ensure the effectiveness and efficiency of clustering algorithms.

We want to partition the above dataset into three clusters using Hierarchical Clustering algorithm. The clustering dendrogram results are shown in the figure-1 and figure-2.

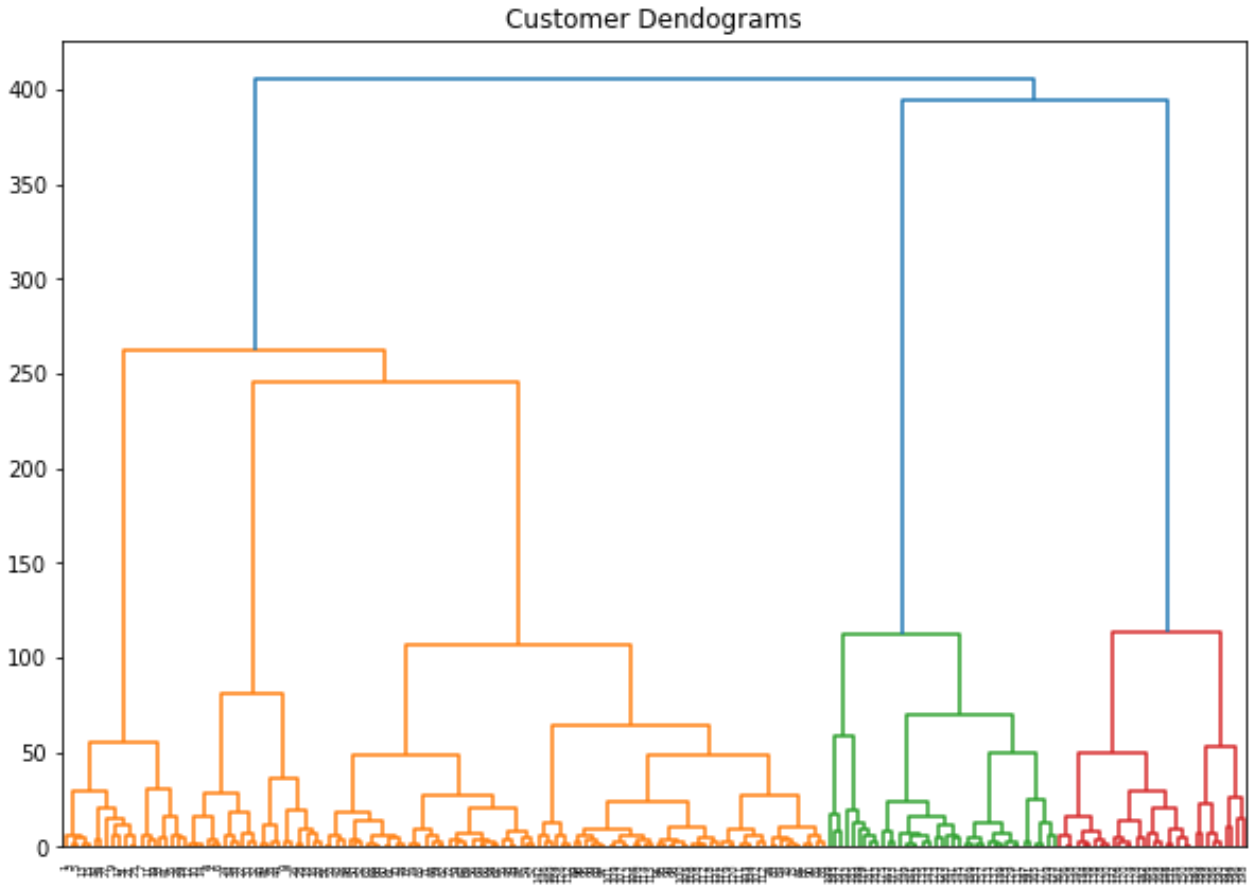

Figure-1: Clustering dendrogram Results

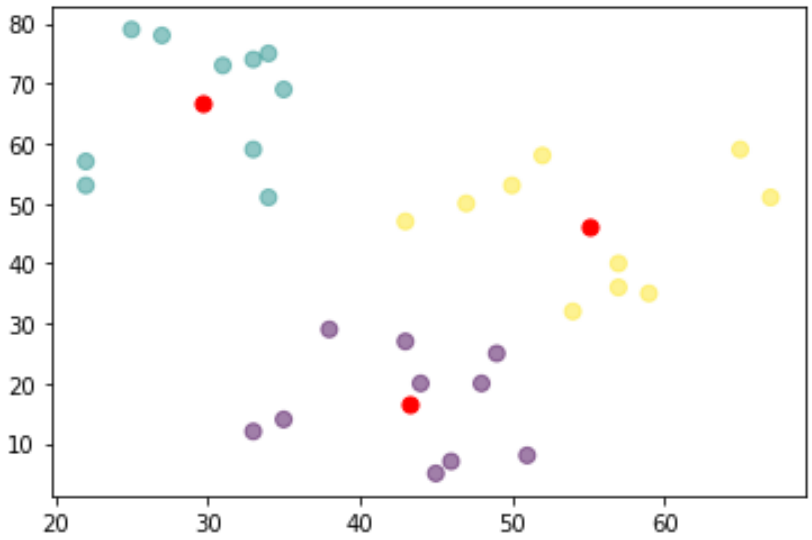

Figure-2: Clustering Results

We observe in the figure-2, there are three clusters of the customer shopping data like cluster-0, cluster-1 and cluster-2 and cluster centroids are as follows:

[1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 0 0 0 0 0 0 0 0 0 0]

[[43.2 16.7]

[29.6 66.8]

[55.1 46.1]]

Hierarchical clustering offered interpretable and actionable insights into the underlying shopping patterns. The hierarchical structure allowed us to observe how smaller subgroups merged to form larger clusters, providing a clear picture of the relationships between customer segments. Retailers can use this information to tailor marketing promotions, optimize product recommendations, and design targeted loyalty programs to increase customer engagement and satisfaction.

V. CONCLUSION

The application of hierarchical clustering algorithms on shopping data demonstrated its effectiveness in uncovering meaningful customer segments and providing actionable insights for retail businesses. By utilizing a combination of agglomerative and divisive clustering techniques, we gained a comprehensive understanding of customer behaviors and preferences, enabling personalized marketing strategies and enhancing customer satisfaction and retention.

REFERENCES

[1] Chris ding and Xiaofeng He (2002), Cluster Merging And Splitting In Hierarchical Clustering Algorithms.

[2] Ian H. Witten and Eibe Frank. Data Mining: Practical machine learning tools and techniques.2nd ed. San Francisco: Morgan Kaufmann, 2005.

[3] J. Han and M. Kamber," Data Mining concepts and Techniques", the Morgan Kaufmann series in Data Management Systems, 2 nd ed. San Mateo, CA; Morgan Kaufmann, 2006.

[4] MarjanKuchaki Rafsanjani, Zahra Asghari Varzaneh, Nasibeh Emami Chukanlo (2012), A survey of hierarchical clustering algorithms, The Journal of Mathematics and Computer Science, 5,.3, pp.229- 240.

[5] Pavel Berkhin (2000), Survey of Clustering Data Mining techniques ,Accrue Software, Inc..

[6] Tian Zhang, Raghu Ramakrishnan, MironLinvy (1996), BIRCH: an efficient data clustering method for large databases, International Conference on Management of Data, In Proc. of 1996 ACM-SIGMOD Montreal, Quebec.

[7] J.A.S. Almeida, L.M.S. Barbosa, A.A.C.C. Pais and S.J. Formosinho (2007), Improving Hierarchical Cluster Analysis: A new method with outlier detection and automatic clustering, Chemo metrics and Intelligent Laboratory Systems, 87, pp. 208-217.

[8] L. Feng, M-H Qiu, Y-X. Wang, Q-L. Xiang, Y-F. Yang and K. Liu (2010), A fast divisive clustering algorithm using an improved discrete particle swarm optimizer, Pattern Recognition Letters, 31, pp. 1216-1225.