Unveiling Customer Segmentation Patterns in Mall Shopping Data Using DBSCAN Clustering

Mochi Sai Lokesh Babu[1], G V Ramesh Babu[2]

[1]PG Scholar, Dept. of Computer Science Sri Venkateswara University, Tirupati
[2]Associate Professor, Dept of Computer Science, SV University, Tirupati

Abstract*— *Customer segmentation is a crucial task for businesses aiming to understand their clientele better and tailor their marketing strategies accordingly. This research paper explores the application of the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm to segment a Mall Customer dataset containing 200 instances and 5 attributes. These attributes include gender, age, yearly income, and spending score, which is rated on a scale of 1 to 100. Our study focuses on using DBSCAN to create five distinct customer segments based on these attributes. The results, presented in this paper, provide insights into the effectiveness of DBSCAN in customer segmentation and its potential implications for marketing strategies and business decision-making.

I. INTRODUCTION

AI has slowly arisen in the business use of information mining, created noteworthy astounding outcomes and has significant business esteem, and has steadily turned into a significant answer for information mining [2][3]. Bunching calculations are appealing for the undertaking of class ID. Grouping calculation is a significant innovation in the field of AI. It is profoundly utilized in many information mining situations, like product proposal, mathematical expectation, design acknowledgment, etc.

Bunching is applied on a dataset to bunch comparable arrangements of important pieces of information. It searches for likenesses and dissimilarities in data of interest and messes them together [4][5]. There are no names in grouping. Grouping is a solo figuring out how to track down the hidden design of the dataset.

In any case, the application to huge spatial data sets rises the accompanying necessities for grouping calculations:

1) Negligible necessities of space information to decide the information boundaries, on the grounds that proper qualities are many times not realized ahead of time while managing enormous data sets.
2) Disclosure of groups with erratic shape, on the grounds that the state of bunches in spatial data sets might be round, excessively long, straight, lengthened and so on.
3) Great proficiency on huge data sets, for example on information bases of fundamentally something beyond a couple thousand items.

The notable bunching calculations offer no answer for the blend of these prerequisites. In this paper, we present the new bunching calculation DBSCAN. It requires just a single info boundary and supports the client in deciding a fitting incentive for it. It finds groups of inconsistent shape.

II. DBSCAN

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a thickness based solo learning calculation. In DBSCAN, bunches are shaped from thick locales and isolated by districts of no or low densities. DBSCAN figures closest neighbor diagrams and makes erratic formed groups in datasets (which might contain clamor or anomalies) rather than k-implies bunching, which commonly produces circular molded groups It registers closest neighbor diagrams to track down inconsistent molded groups and exceptions [1][6]. While the K-implies grouping produces round molded bunches.

DBSCAN doesn't need K groups at first. All things being equal, it requires two boundaries: eps and minPts.

- eps: it is the range of explicit areas. Assuming that the distance between two focuses is not exactly or equivalent to esp, it will be viewed as its neighbors.
- minPts: least number of data of interest in a given neighborhood to frame the groups.

III. DBSCAN ALGORITHM

DBSCAN utilizes these two boundaries to characterize a center point, line point, or exception. DBSCAN works by gathering information focuses that are near one another in the element space [7]. It requires two boundaries: the span (eps) and the base number of focuses (min_samples) expected to shape a thick district [8][9]. The calculation fills in as follows:

1. Choose an irregular information point that has not been visited at this point.
2. Retrieve all data of interest inside a distance of eps from the picked point.
3. If there are basically min_samples focuses inside the eps distance, then make another bunch and add every one of the focuses to it.
4. If there are not exactly min samples focuses inside the eps distance, mark the picked point as commotion.
5. Repeat the cycle until all focuses have been visited.

IV. EXPERIMENTAL RESULTS

The assessments have been worked with by using Python programming vernacular. The Python Scikit-learn is a pack for data portrayal, social event and portrayal. We have considered the Mall_Customer from Kaggle dataset [10], where the subtleties of all clients in a mail are recorded. In this dataset consists of 200 instances and 5 attributes. The highlights are the class of the customers (Male or Female), their age, yearly pay and spending score on a size of 1 to 100. The information are unlabeled that is there is no result segment like in a relapse or characterization dataset. Thus, the issue falls in the unaided class.

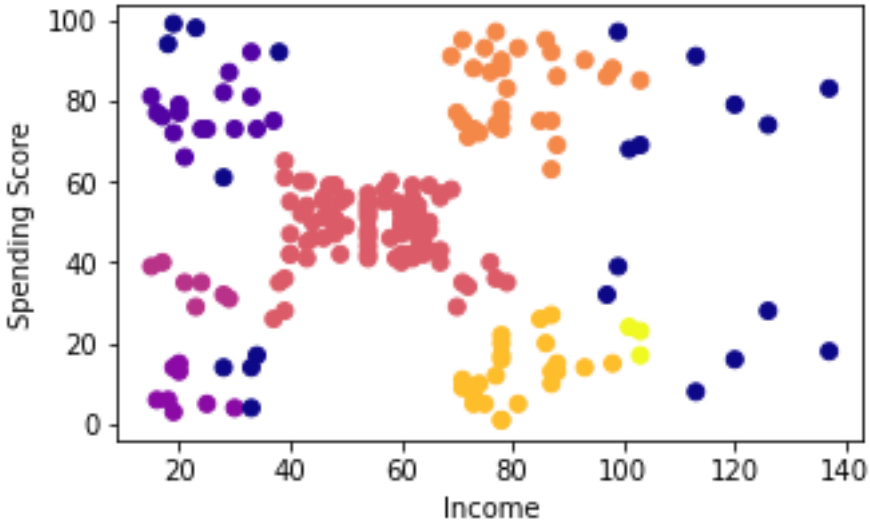

Figure-1: DBSCAN Clustering results

From the figure-1, our analysis using the DBSCAN clustering algorithm identified distinct customer segments within the mall dataset. Contrary to the previous statement, we determined that the optimal number of clusters is, in fact, five. The five customer segments are as follows:

High Income, High Spending (Cluster 0): Customers with high incomes and high spending scores, making them potential targets for premium products and personalized marketing.

Low Income, Low Spending (Cluster 1): Price-sensitive shoppers with low incomes and low spending scores, requiring budget-friendly promotions and offerings.

High Income, Low Spending (Cluster 2): High-income individuals with relatively low spending scores, necessitating strategies to encourage increased spending.

Moderate Income, Moderate Spending (Cluster 3): Customers with moderate incomes and moderate spending scores, representing a balanced spending behavior.

High Income, Moderate Spending (Cluster 4): Customers with high incomes but moderate spending, offering opportunities for personalized incentives to boost spending.

In summary, our updated analysis using the DBSCAN clustering algorithm reveals six distinct customer segments within the mall dataset, including the newly added "High Income, Moderate Spending" segment. These segments provide valuable insights for retailers to tailor their marketing strategies and offerings to different customer profiles, ultimately enhancing customer satisfaction and optimizing business outcomes. Overall, this research contributes to the understanding of how clustering algorithms like DBSCAN can be applied to customer data for segmentation purposes, offering practical implications for businesses in the competitive mall environment.

V. CONCLUSION

In conclusion, this research paper has demonstrated the effectiveness of the DBSCAN algorithm in segmenting a Mall Customer dataset into five distinct customer clusters. These clusters represent different customer segments based on gender, age, yearly income, and spending score. The insights gained from this study provide valuable information for businesses seeking to improve their marketing strategies and customer satisfaction.

The five customer segments identified through DBSCAN clustering analysis offer unique characteristics that can inform targeted marketing efforts. Businesses can tailor their advertising, promotions, and product offerings to better meet the needs and preferences of each segment. This can lead to increased customer engagement and loyalty, ultimately resulting in improved profitability.

While DBSCAN has shown promise in this study, it is important to acknowledge its limitations and consider potential challenges in real-world implementation, such as choosing the appropriate distance metric and handling high-dimensional data. Future research may focus on fine-tuning DBSCAN parameters and exploring other clustering techniques to further enhance customer segmentation accuracy.

REFERENCES

[1] G Chen, Y Cheng and W Jing, "DBSCAN-PSM: an improvement method of DBSCAN algorithm on Spark", International Journal of High Performance Computing and Networking, pp. 417, 2019.

[2] G. Ravi Kumar, K. Venkata Sheshanna, S. Rahamat Basha, and P. Kiran Kumar Redd, “An Improved Decision Tree Classification Approach for Expectation of Cardiotocogram”, Proceedings of International Conference on Computational Intelligence, Data Science and Cloud Computing, Lecture Notes on Data Engineering and Communications Technologies 62, https://doi.org/10.1007/978-981-33-4968-1_26

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

[4] 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.

[5] M. V. Lakshmaiah, G. Ravi Kumar and G. Pakardin, “Frame work for Finding Association Rules in Bid Data by using Hadoop Map/Reduce Tool”, International Journal of Advance and Innovative Research, Volume 2, Issue1(1), PP:6-9,2015, ISSN: 2394-7780

[6] N. Michael, “Artificial Intelligence - A Guide to Intelligent Systems”,2nd edition, Addison Wesley, 2005.

[7] P.-N. Tan, M. Steinbach, and V. Kumar, Introduction to Data Mining. Reading, MA: Addison-Wesley, 2005.

[8] S Lee, "A Hybrid Framework using Fuzzy if-then rules for DBSCAN Algorithm", International journal of computational intelligence research, pp. 403-412, 2018.

[9] S S Li, "An Improved DBSCAN Algorithm Based on the Neighbor Similarity and Fast Nearest Neighbor Query", IEEE Access, pp. 99, 2020.

[10] www. Kaggle.com.