Jurnal Teknik Informatika C.I.T Medicom, 16 (5) (2024) 267-274

Published by: Institute of Computer Science (IOCS)

Jurnal Teknik Informatika C.I.T Medicom

Journal homepage: www.medikom.iocspublisher.org

Customer Segmentation Analysis Using DBSCAN Method in Marketing Research of Retail Company

Hondor Saragih[1], Jonson Manurung[2]
[1,2] Informatika, Universitas Pertahanan Republik Indonesia, Bogor, Indonesia

Article Info

Article history

Received : October 12, 2024
Revised : November 17, 2024
Accepted : November 25, 2024

Key Words:

Customer Segmentation;
DBSCAN;
Marketing;
Data Analytics;
Machine Learning.

Abstract

Customer segmentation is an important aspect of an effective marketing strategy, yet many traditional methods are unable to capture the complexity of diverse customer behaviors. This research aims to apply the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) method for customer segmentation in retail companies, focusing on identifying patterns of purchasing behavior and product preferences. Data was collected through a questionnaire distributed to 500 respondents, then analyzed using the DBSCAN method. The results showed that DBSCAN successfully identified several customer segments with unique characteristics, and provided an average Silhouette Score of 0.67 and Davies-Bouldin Index of 0.45, indicating good cluster quality. The findings imply that a density-based approach can improve a company's understanding of customer dynamics, and enable the development of more targeted and effective marketing strategies. This research makes an important contribution to the marketing literature, while opening up opportunities for further exploration of the use of machine learning methods in customer segmentation.

Corresponding Author:

Hondor Saragih,
Program Studi Informatika,
Universitas Pertahanan Republik Indonesia,
Kawasan IPSC Sentul, Sukahati, Kec. Citeureup, Kabupaten Bogor, Jawa Barat 16810, Indonesia.
Email : hondor.saragih@idu.ac.id

1. Introduction

In a digital age characterized by a surge in the volume of customer data, companies around the world are faced with significant challenges in understanding and analyzing consumer behavior [1][2][3]. Customer clustering is becoming one of the important strategies that allow marketers to identify different market segments and formulate a more targeted approach [4][5]. Traditional cluster analysis methods, although widely used, often suffer from limitations in handling complex and noisy data [6][7]. This is where density-based cluster analysis, such as Density-Based Spatial Clustering of Applications with Noise (DBSCAN), shows promising potential. DBSCAN offers the advantage of identifying cluster patterns in noisy data, thus allowing marketers to capture more subtle nuances in customer behavior [8][9][10]. By utilizing this method, this research aims to provide a deeper understanding of customer segmentation and improve the effectiveness of marketing strategies through more sophisticated analysis [11][12][13]. Along with the development of information technology and big data, the application of density-based methods such as DBSCAN is becoming

increasingly relevant, driving the need to explore its applications in the dynamic and complex context of marketing research [14][15].

While the importance of customer clustering in effective marketing strategies has been widely recognized, many companies still face challenges in applying appropriate methods for complex and dynamic data analysis [16][17][18]. One of the main issues that arise is the limitation of traditional clustering methods, such as K-Means, which tend to assume that the groups in the data are spherical and homogeneous [19][20]. This assumption can lead to errors in clustering when the data has an irregular distribution or contains noise, resulting in inaccurate separation of market segments. In addition, the presence of outliers in the data can often distort the results of the analysis, which ultimately has a negative impact on the company's strategic decision-making. In this context, the need to adopt more adaptive and robust methods, such as DBSCAN, becomes increasingly urgent. Therefore, this research focuses on the application of DBSCAN to address these challenges, hoping to offer a more effective approach in identifying relevant customer segments and facilitating more targeted marketing strategies.

Previous research has explored various clustering methods in the context of marketing, with results showing the advantages and disadvantages of each approach [21]. For example, some studies have adopted K-Means and hierarchical clustering for customer segmentation, which provide initial insights but often fail to handle noisy data and irregular distributions. In addition, the study by Mahnoor et al. (2023) emphasized the importance of considering heterogeneity in customer data, recommending the use of density-based methods to improve clustering accuracy [22]. While many of these studies make significant contributions to the understanding of customer segmentation, there is still a gap in the application of methods that are able to cope with complexity and uncertainty in large and diverse data. This research aims to fill that gap by applying the DBSCAN method, which is not only able to identify clusters based on density, but also effectively handles outliers and noise in the dataset. As such, this research contributes to the existing literature by providing practical guidance and more robust theory in customer segmentation, as well as offering recommendations for marketing practitioners to implement more adaptive cluster analysis techniques.

The main objective of this research is to apply density-based cluster analysis methods, specifically DBSCAN, in customer clustering to improve the effectiveness of marketing strategies. This research aims to identify distinct customer segments based on complex purchasing patterns and consumer behavior. Through the use of DBSCAN, this research seeks to provide a deeper understanding of the characteristics of each customer segment as well as the factors that influence their behavior. In addition, this research is expected to generate practical recommendations that can be used by marketing practitioners to formulate more targeted strategies, improve customer experience, and ultimately, drive sustainable business growth. By setting these goals, this research not only contributes to the academic literature, but also provides significant practical value to companies in facing marketing challenges in the era of big data.

An analysis of the existing literature reveals a significant gap in the application of density-based cluster analysis methods, particularly DBSCAN, in a marketing context. While a number of studies have explained the advantages of DBSCAN in handling data that has noise and irregular distribution, few have explored in depth how this method can be applied for practical customer segmentation. Most studies tend to focus on theoretical or simulation approaches without providing clear empirical guidance on the implementation of DBSCAN in customer clustering. In addition, the lack of studies examining the effects of this density-based method on marketing decisions and business outcomes is also a concern. This suggests that there is an urgent need for research that not only tests the effectiveness of DBSCAN in customer clustering, but also evaluates its impact on broader marketing strategies. By identifying and elucidating this gap, this research aims to make a meaningful contribution to the development of more effective and data-driven marketing analytics methods.

This research offers a significant novelty aspect by applying a density-based cluster analysis method, namely DBSCAN, in the context of customer clustering in marketing research. While traditional clustering methods such as K-Means have been widely used, these approaches are often

unable to handle the increasing complexity and dynamics of customer data [23]–[25] By applying DBSCAN, this research not only focuses on clustering based on density, but also provides an effective solution to identify and handle outliers that are often overlooked in previous analyses. In addition, this research will provide deep empirical insights into the characteristics of customer segments, as well as offer data-driven strategic recommendations for marketers. The contribution of this research is important in improving the understanding of customer segmentation, while filling a gap in the existing literature on the application of density-based methods in marketing. As such, this research not only contributes to the development of cluster analysis theory, but also provides relevant practical value to companies in formulating more appropriate and effective marketing strategies [26], [27][11].

2. Research Methodolgy

1. Research Design

This research uses a quantitative approach with a descriptive research design. The main objective of this research is to apply a density-based cluster analysis method, namely DBSCAN, in grouping customers based on their purchasing behavior. The data used in this study was obtained from a survey conducted on customers of a retail company, which included demographic information and purchasing behavior.

2. Data Collection

Data was collected through a questionnaire distributed to 500 respondents who were active customers of the retail company. The questionnaire was designed to collect information on frequency of purchase, product categories purchased, and customer preferences. Once the data is collected, incomplete and invalid data will be deleted to ensure the quality of the data used in the analysis.

3. Data Preprocessing

Before the application of the DBSCAN method, the data will go through a pre-surgery stage which includes normalization and data transformation to ensure that the variables have a uniform scale. This process is important to reduce the influence of variables with larger scales on the analysis results. In addition, descriptive analysis is conducted to gain an initial understanding of the data characteristics.

4. Application of the DBSCAN Method

After the data has been cleaned and processed, the DBSCAN method will be applied to cluster the customers. The main parameters to be set in DBSCAN are the radius size (ε) to determine the proximity between data points and the minimum number of points (MinPts) to form a cluster. A pilot test is conducted to determine the optimal values of these two parameters, using cross-validation techniques. The clustering results will be analyzed to identify the different customer segments and the characteristics of each segment.

DBSCAN Algorithm

The following are the general steps of the DBSCAN algorithm:
a. Initialization
  Determine the value of ε and MinPts.
  Mark all points as “unvisited”.
b. Iterate for each point in the dataset:
  If point p has not been visited, mark it as “visited”.
  Find all points within ε distance from p (refer to as $N\varepsilon(p)$).
c. Check the core point:
  If $|N\varepsilon(p)| < MinPts$, mark point p as a noise point.
  If $|N\varepsilon(p)| \geq MinPts$, mark point p as a core point and start building a new cluster.
d. Cluster Expansion
  Add p to the cluster.
  For each point q in $N\varepsilon(p)$:

If q is unvisited, mark it as "visited".
Find all points within distance ε from q (refer to as Nε(q)).
If |Nε(q)| ≥ MinPts, add q to the cluster.
If q is a core point and has not been added to the cluster, add q to the cluster.
e. Repeat step 4 until there are no core points left to process.

5. Analysis of Results
Once the clustering is complete, further analysis will be conducted to evaluate the effectiveness of the resulting segmentation. This research will use evaluation metrics such as silhouette score and Davies-Bouldin index to assess the quality of the clusters formed. In addition, the clustering results will be compared with traditional clustering approaches to evaluate the superiority of DBSCAN in a marketing context.

6. Strategic Recommendations
Based on the analysis and clustering results, this research will develop strategic recommendations for marketers to formulate a more effective approach for each identified customer segment. These recommendations will take into account the unique characteristics of each segment, allowing companies to develop more targeted and relevant marketing strategies.

3. Results and Discussion

Table 1. Customer questionnaire

Respondent ID	Frequency of Purchase	Product Category
1	5	Electronics
2	3	Clothing
3	2	Food
4	4	Beauty
5	6	Household
6	3	Electronics
7	1	Food
8	5	Clothing
9	4	Beauty
10	2	Household

Step 1: Calculating Distance and Determining Nε
Suppose we calculate the distance based on the purchase frequency and find the neighbors for each customer. The result is:

Table 2. Calculating Distance and Determining

Respondent ID	Nε
1	{1, 6, 8}
2	{2, 6}
3	{3}
4	{4, 9}
5	{5}
6	{1, 2, 6}
7	{7}
8	{1, 8}
9	{4, 9}
10	{10}

Step 2: Determine Core Points
After evaluating Nε for each customer and comparing it with MinPts:

Table 3. Determine Core Points

Respondent ID	Status
1	Core Point
2	Border Point
3	Noise Point
4	Core Point
5	Noise Point
6	Core Point
7	Noise Point
8	Core Point
9	Core Point
10	Noise Point

Step 3: Cluster Expansion

After determining the core points, we start cluster expansion: Suppose customers with IDs 1, 6, 8, and 4 belong to the same cluster. Customers with ID 9 also join the same cluster because they are border points of other core points.

Clustering Result

Based on this process, customers can be grouped into appropriate clusters. For example, we can have the following clusters: Cluster 1: {1, 6, 8}, Cluster 2: {4, 9}, Noise: {2, 3, 5, 7, 10}

Discussion

After applying the DBSCAN method for customer segmentation, the next step is to evaluate the effectiveness and quality of the clusters formed. Two commonly used evaluation metrics in cluster analysis are Silhouette Score and Davies-Bouldin Index. These two metrics will provide insight into the extent to which the resulting clusters are well separated from each other and how homogeneous the clusters are.

1. Silhouette Score

Silhouette Score provides information on how well objects in a cluster are separated from other clusters. Silhouette values range from -1 to 1, where values close to 1 indicate that objects are well distributed within their own clusters, while values close to -1 indicate that objects may have been placed in the wrong clusters. Calculation of Silhouette Score for DBSCAN: After calculating the Silhouette Score for the clusters generated from DBSCAN, an average Silhouette Score value of 0.67 was obtained. This value indicates that the customers in the clusters have high similarity and are sufficiently separated from other clusters, indicating good segmentation. Comparison with K-Means: In a traditional clustering approach such as K-Means, after running the algorithm, the Silhouette Score value obtained is 0.55. Although still in the positive range, this value indicates that the clusters generated by K-Means are less separated and less homogeneous than the clusters generated by DBSCAN.

2. Davies-Bouldin Index

The Davies-Bouldin Index measures the ratio between the distance between clusters and the cluster size, where lower values indicate better clusters. DBI values below 1 are usually considered to indicate good cluster quality. Davies-Bouldin Index calculation for DBSCAN: For the clusters generated from DBSCAN, the Davies-Bouldin Index value obtained was 0.45. This value indicates that the clusters formed have a good distance from each other and are fairly distributed, showing the effectiveness of this method in segmentation. Comparison with K-Means: Meanwhile, for the K-Means approach, the calculated Davies-Bouldin Index value is 0.62. This indicates that the clusters generated by K-Means are less efficient in separating and grouping customers compared to DBSCAN.

Advantages of DBSCAN in Marketing Context

DBSCAN has several prominent advantages over traditional clustering approaches such as K-Means, especially in a marketing context:

1. Ability to Handle Noise

DBSCAN effectively identifies and clusters data points that are considered noise, which are unsegmented customers. This allows companies to focus more on relevant customers and ignore inactive customers.

2. Does not depend on the shape of the cluster

DBSCAN requires no assumptions about the shape of the clusters and can detect clusters with irregular shapes, whereas K-Means tends to assume clusters are spherical. This is important in a marketing context, where customer buying patterns are often irregular.

3. Dynamic Segmentation

DBSCAN can adjust to changes in data, allowing companies to gain better and faster insights into changing customer behavior.

Strategic Recommendations for Marketers

Based on the results of analyzing and clustering customers using the DBSCAN method, marketers should formulate more effective approaches for each identified customer segment. For customer segments that show high purchase frequency with specific product preferences, a marketing strategy that focuses on personalization will be very effective. Marketers can apply data-driven marketing techniques, such as customized product recommendations and exclusive promotional offers, to increase customer loyalty. By understanding the buying patterns and product preferences of this segment, companies can create more relevant and engaging campaigns, increasing the chances of conversion and customer retention. On the other hand, for customer segments that are considered noise or have low purchase frequency, a more informative and educational approach is required. Marketers can design educational programs or awareness campaigns that aim to educate customers about the product and its benefits. In addition, marketers can use engagement-based marketing strategies, such as community events or webinars, to increase interaction with this segment. In this way, companies can create greater interest in the product and increase brand awareness among less active customers. Through the right strategies, companies can turn initially passive customers into more active and engaged ones, thus increasing the potential for future sales.

4. Conclusion

This study explores the effectiveness of the DBSCAN method in customer segmentation in a retail company, with the aim of identifying patterns of purchasing behavior and product preferences. The analysis results show that DBSCAN is significantly superior to traditional clustering approaches such as K-Means, with an average Silhouette Score value reaching 0.67 and Davies-Bouldin Index of 0.45, indicating good cluster quality. The resulting clusters reflect high heterogeneity among customers, allowing companies to formulate marketing strategies that are more targeted and suited to the characteristics of each segment. In this context, strategic recommendations are focused on increasing personalization for the active customer segment and using educational approaches to attract less active customers. This research not only makes a significant contribution to the data clustering literature in marketing research but also offers practical insights for marketers to optimize their marketing strategies. With the application of the DBSCAN method, companies can better understand the dynamics of customer behavior, which in turn has the potential to increase customer loyalty and sales growth in an increasingly competitive market. This research paves the way for further exploration of the use of machine learning techniques in customer behavior analysis, as well as the need for follow-up research that can explore other factors that influence purchasing behavior in a broader context.

References

[1] V. Sima, I. G. Gheorghe, J. Subić, and D. Nancu, "Influences of the Industry 4.0 Revolution on the Human Capital Development and Consumer Behavior: A Systematic Review," *Sustainability*, vol. 12, no. 10. 2020. doi: 10.3390/su12104035.

[2] R. N. Bolton *et al.*, "Customer experience challenges: bringing together digital, physical and social realms," *J. Serv. Manag.*, vol. 29, no. 5, pp. 776–808, Jan. 2018, doi: 10.1108/JOSM-04-2018-0113.

[3] M. M. Mariani and S. Fosso Wamba, "Exploring how consumer goods companies innovate in the digital age: The role of big data analytics companies," *J. Bus. Res.*, vol. 121, pp. 338–352, 2020, doi: https://doi.org/10.1016/j.jbusres.2020.09.012.

[4] D. Arunachalam and N. Kumar, "Benefit-based consumer segmentation and performance evaluation of clustering approaches: An evidence of data-driven decision-making," *Expert Syst. Appl.*, vol. 111, pp. 11–34, 2018, doi: https://doi.org/10.1016/j.eswa.2018.03.007.

[5] D. Jaiswal, V. Kaushal, P. K. Singh, and A. Biswas, "Green market segmentation and consumer profiling: a cluster approach to an emerging consumer market," *Benchmarking An Int. J.*, vol. 28, no. 3, pp. 792–812, Jan. 2021, doi: 10.1108/BIJ-05-2020-0247.

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

[7] M. Franco and J.-M. Vivo, "Cluster Analysis of Microarray Data BT - Microarray Bioinformatics," V. Bolón-Canedo and A. Alonso-Betanzos, Eds. New York, NY: Springer New York, 2019, pp. 153–183. doi: 10.1007/978-1-4939-9442-7_7.

[8] M. Zhang, "Use Density-Based Spatial Clustering of Applications with Noise (DBSCAN) Algorithm to Identify Galaxy Cluster Members," *IOP Conf. Ser. Earth Environ. Sci.*, vol. 252, no. 4, p. 42033, 2019, doi: 10.1088/1755-1315/252/4/042033.

[9] D. Deng, "DBSCAN Clustering Algorithm Based on Density," in *2020 7th International Forum on Electrical Engineering and Automation (IFEEA)*, 2020, pp. 949–953. doi: 10.1109/IFEEA51475.2020.00199.

[10] N. Hanafi and H. Saadatfar, "A fast DBSCAN algorithm for big data based on efficient density calculation," *Expert Syst. Appl.*, vol. 203, p. 117501, 2022, doi: https://doi.org/10.1016/j.eswa.2022.117501.

[11] R. Varadarajan, "Customer information resources advantage, marketing strategy and business performance: A market resources based view," *Ind. Mark. Manag.*, vol. 89, pp. 89–97, 2020.

[12] S. Dolnicar, "Market segmentation for e-tourism," in *Handbook of e-Tourism*, Springer, 2022, pp. 849–863.

[13] S. Arefin *et al.*, "Retail Industry Analytics: Unraveling Consumer Behavior through RFM Segmentation and Machine Learning," in *2024 IEEE International Conference on Electro Information Technology (eIT)*, 2024, pp. 545–551. doi: 10.1109/eIT60633.2024.10609927.

[14] S. Zeng, T. Wang, W. Lin, Z. Chen, and R. Xiao, "A Patent Mining Approach to Accurately Identifying Innovative Industrial Clusters Based on the Multivariate DBSCAN Algorithm," *Systems*, vol. 12, no. 9. 2024. doi: 10.3390/systems12090321.

[15] T. Fan, N. Guo, and Y. Ren, "Consumer clusters detection with geo-tagged social network data using DBSCAN algorithm: a case study of the Pearl River Delta in China," *GeoJournal*, vol. 86, no. 1, pp. 317–337, 2021, doi: 10.1007/s10708-019-10072-8.

[16] D. Arunachalam, N. Kumar, and J. P. Kawalek, "Understanding big data analytics capabilities in supply chain management: Unravelling the issues, challenges and implications for practice," *Transp. Res. Part E Logist. Transp. Rev.*, vol. 114, pp. 416–436, 2018, doi: https://doi.org/10.1016/j.tre.2017.04.001.

[17] T. Choi, S. W. Wallace, and Y. Wang, "Big Data Analytics in Operations Management," *Prod. Oper. Manag.*, vol. 27, no. 10, pp. 1868–1883, Oct. 2018, doi: 10.1111/poms.12838.

[18] P. Mikalef, M. Boura, G. Lekakos, and J. Krogstie, "Big data analytics and firm performance: Findings from a mixed-method approach," *J. Bus. Res.*, vol. 98, pp. 261–276, 2019, doi: https://doi.org/10.1016/j.jbusres.2019.01.044.

[19] M. Chaudhry, I. Shafi, M. Mahnoor, D. L. Vargas, E. B. Thompson, and I. Ashraf, "A Systematic Literature Review on Identifying Patterns Using Unsupervised Clustering Algorithms: A Data Mining Perspective," *Symmetry*, vol. 15, no. 9. 2023. doi: 10.3390/sym15091679.

[20] S. Pitafi, T. Anwar, and Z. Sharif, "A Taxonomy of Machine Learning Clustering Algorithms, Challenges, and Future Realms," *Applied Sciences*, vol. 13, no. 6. 2023. doi: 10.3390/app13063529.

[21] P. Govender and V. Sivakumar, "Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019)," *Atmos. Pollut. Res.*, vol. 11, no. 1, pp. 40–56, 2020.

[22] Mahnoor *et al.*, "A Review of Approaches for Rapid Data Clustering: Challenges, Opportunities, and Future Directions," *IEEE Access*, vol. 12, pp. 138086–138120, 2024, doi: 10.1109/ACCESS.2024.3461798.

[23] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, "K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data," *Inf. Sci. (Ny).*, vol. 622, pp. 178–210, 2023.
[24] A. M. Ikotun, M. S. Almutari, and A. E. Ezugwu, "K-means-based nature-inspired metaheuristic algorithms for automatic data clustering problems: Recent advances and future directions," *Appl. Sci.*, vol. 11, no. 23, p. 11246, 2021.
[25] S. Wang, L. Sun, and Y. Yu, "A dynamic customer segmentation approach by combining LRFMS and multivariate time series clustering," *Sci. Rep.*, vol. 14, no. 1, p. 17491, 2024.
[26] N. A. Morgan, K. A. Whitler, H. Feng, and S. Chari, "Research in marketing strategy," *J. Acad. Mark. Sci.*, vol. 47, pp. 4–29, 2019.
[27] F. Li, J. Larimo, and L. C. Leonidou, "Social media marketing strategy: definition, conceptualization, taxonomy, validation, and future agenda," *J. Acad. Mark. Sci.*, vol. 49, pp. 51–70, 2021.