Customer Segmentation in User Behavior Analysis: A Comparative Study of Clustering Algorithms

Yingze Liu*

College of Information Engineering, Xi'an University, Xi'an, China

*Corresponding author: liuyingze@stu.xawl.edu.cn

Abstract. A thorough understanding of customer behavior patterns is still essential for corporate success in today's digital environment. To segment clients in-depth, this study used three distinct clustering algorithms: k-means, hierarchical clustering, and dbscan. By carefully examining the age, yearly income, and consumption score This study goes beyond conventional views and creates a comprehensive and distinct perspective when disclosing numerous consumer attributes using data derived from the mall consumer Segmentation Data set. It offers insightful information that can be used to adjust a company's marketing plan. The strategic application of customer insights is redefined in this study, empowering stakeholders to decide wisely and improve market performance. The road to greater competitiveness and relevance in a developing market segment is mapped using real-world data and powerful clustering techniques. Based on research of this dataset, it was discovered that Dbscan performed best on this dataset with unequal density. It highlights the usefulness of these algorithms in today's complicated business world.

Keywords: Customer segmentation; user behavior analysis; clustering algorithms; K-means; hierarchical clustering, DBSCAN.

1. Introduction

Businesses now have quick access to enormous volumes of data on client behavior thanks to the Internet's explosive growth. Researchers and companies are increasingly using clustering algorithms to analyze customer behavior and uncover patterns and trends in order to make better use of this data [1]. Numerous academics have already investigated the use of clustering algorithms to examine customer purchasing trends, preferences, and spending practices. Researchers can better grasp the features of various market segments by clustering clients who share similar traits, enabling businesses to create more individualized and targeted marketing plans [2].

The significance of customer classification in this area, though, is more noticeable after presenting the existing study on customer behavior analysis. Customer classification is the process of breaking down large, complicated customer groupings into smaller categories that have comparable traits and behavior patterns. This classification gives businesses the chance to learn more about consumer wants and behaviors as well as aids in more precise market targeting. Businesses can better understand the demands, purchasing patterns, and possible market opportunities of various kinds of consumers by categorizing their clients [3]. Through more targeted product and service launches and more precise market strategy formulation, businesses are better able to compete in their respective markets and satisfy their customers.

Customer classification has a lot of potential for user behavior analysis, but there isn't a lot of research on it right now. Although some studies have started to investigate the use of various clustering algorithms in customer classification, more research is still required to determine the real effects and viability in various market and industrial situations. This paper has the chance to further investigate the possibilities of clustering algorithms in customer behavior research and provide businesses more intelligent and forward-looking market judgments as the fields of data science and artificial intelligence continue to develop. In order to give solid backing for marketing decisions, this study intends to provide businesses with a more scientific technique of customer behavior analysis through the thorough comparison of various clustering methods [3-5].

2. Data

The dataset used in this study, Mall Customer Segmentation Data, was made available on the Kaggle website by author Vijay Choudhary. This data set was created primarily to demonstrate the concepts of customer segmentation and market basket analysis. This dataset contains basic consumer data collected by a supermarket using loyalty cards, such as customer ID, age, gender, yearly income, and expenditure score. Let's now see how this dataset was visualized using some straightforward Python machine learning techniques. Plot the gender statistics, the relationship between the columns and rows, and the histograms in the following Figs 1-6.

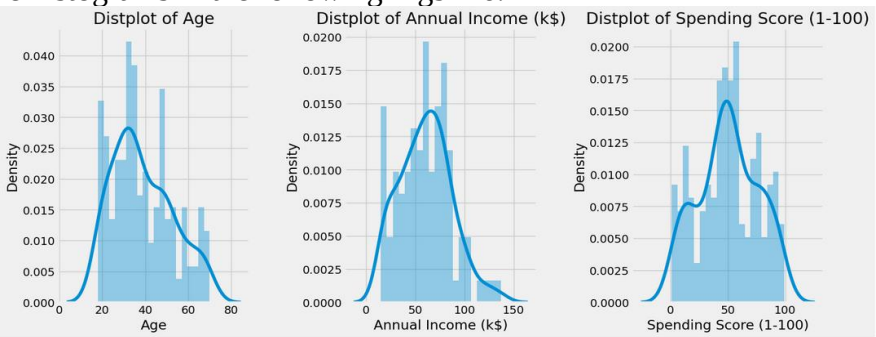

Fig. 1 Histograms

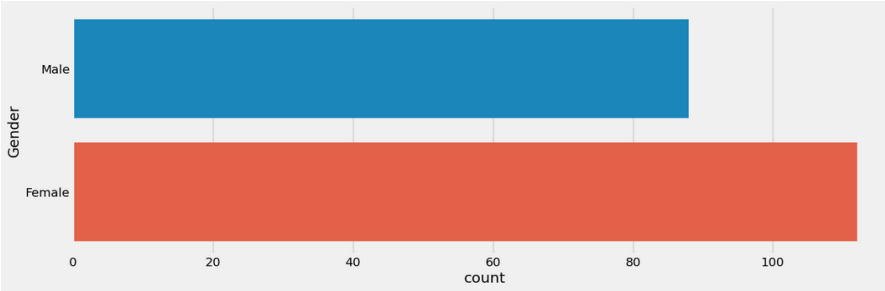

Fig. 2 Count Plot of Gender

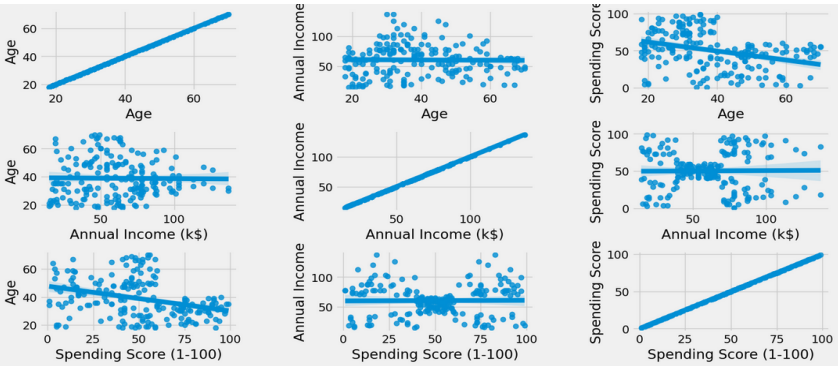

Fig. 3 The relation between Age, Annual Income and Spending Score

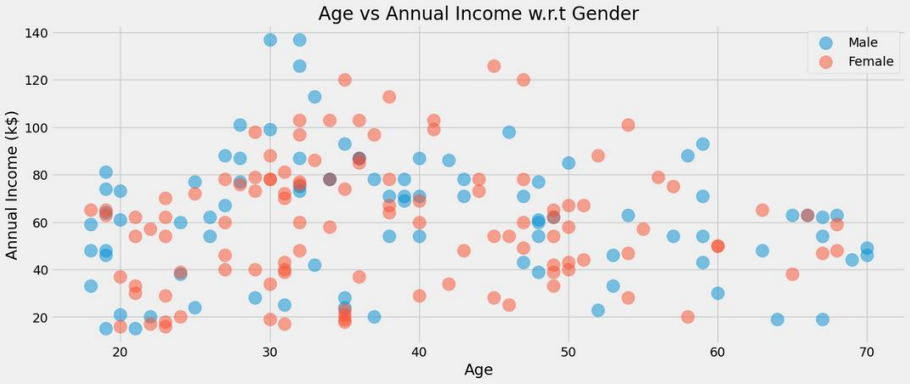

Fig. 4 The relation between Age, Annual Income and Spending Score

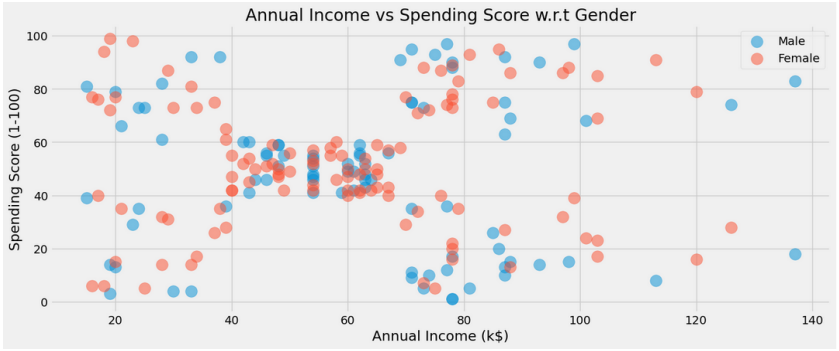

Fig. 5 The relation between Age, Annual Income and Spending Score

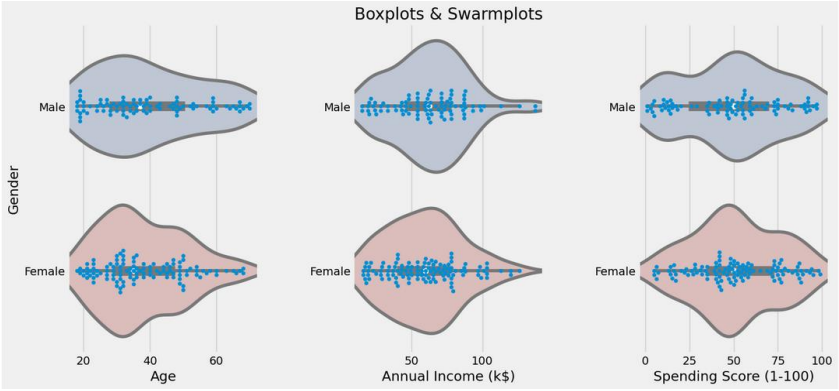

Fig. 6 Distribution of values in Age, Annual Income and Spending Score according to Gender

3. Methods

3.1. K-means

Due to its popularity and widespread application in the field of data analysis, K-means is a popular clustering approach [1]. The algorithm's fundamental principle is to split the data set into K clusters of a predetermined number, each of which is represented by a centroid (cluster center). In order to highlight the fundamental structure and pattern of the data, this partitioning technique groups similar data points into one cluster based on their proximity and similarity.

K-means performs admirably in a variety of circumstances, particularly when trying to discover globular clusters [3]. For instance, K-means can successfully group consumers who tend to organize into clusters based on similar purchase preferences in particular market segmentation circumstances. K-means can quickly categorize customers into distinct groups by providing the expected number of clusters K, which helps organizations better understand the traits and requirements of various client groups.

3.2. Hierarchical Clustering

Because it creates a hierarchy of clusters instead of requiring us to predetermine the number of clusters, hierarchical clustering is a potent clustering technique that enables us to group data at various fine-grained levels. Greater flexibility for in-depth data analysis is made possible by this. Hierarchical clustering has the advantage of handling clusters of varying sizes and forms, better adjusting to different data distributions [6].

The similarity between clusters in hierarchical clustering is determined by the separation or similarity between the data points. Starting with a single data point, the algorithm gradually groups the data points into larger clusters, resulting in the formation of a hierarchy of clusters. This structure can be presented as a tree or "dendrogram," which offers an intuitive manner to comprehend the connections between the data.

3.3. DBSCAN

A well-known density-based clustering technique that offers strong analytical tools to the field of data analysis is called DBSCAN (Density-Based Spatial Clustering of Applications with Noise). The algorithm's ability to automatically locate clusters of various sizes and shapes, as well as to successfully identify and remove noise points from the data set, is what makes it so impressive [7]. DBSCAN outperforms conventional K-means and hierarchical clustering techniques in some circumstances.

One benefit of DBSCAN is that it splits clusters based on the density of the data points rather than requiring a predetermined number of clusters. DBSCAN is able to adapt to variations in the number and size of clusters in various datasets as a result, giving data analysis more freedom. DBSCAN has the apparent advantage of being able to precisely divide clusters with varying densities, which is especially advantageous when working with data sets with uneven density distributions.

4. Result

4.1. K- means

Firstly, the columns 'Age' and 'Spending Score (1-100)' are selected from the DataFrame as input features and converted to numpy array X1. Data is clustered using the K-means clustering technique, and inertia is calculated for different numbers of clusters (1 to 10). The K-means technique uses the inertia performance index, which sums the distances between each data point and the cluster's center to determine how compact the cluster is. In the loop, each iteration adds the value of the current number of clusters to the list Inertia, and finally draws the "Elbow Method" diagram (Fig. 7) to select the best number of clusters.

Secondly, K-means clustering is used to cluster the two features of 'Age' and 'Spending Score (1-100)' in 'Mall_Customers.xls' data set, and the clustering results are visualized. The graph shows raw data points, clustering centers, and decision boundaries. Through such visualization, clustering structure of the data is understood (Fig. 8).

Fig. 7 Elbow Method

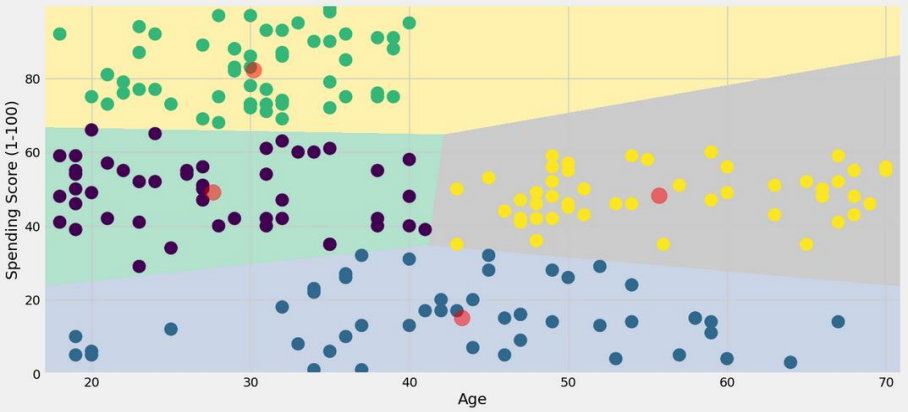

Fig. 8 Clustering using K- means

4.2. Hierarchical Clustering

This paper created an instance of 'AgglomerativeClustering' from the 'scikit-learn' library with the 'n_clusters' parameter set to 5, signifying the division of data into 5 clusters. Next, this paper selected the two features "Age" and "Spending Score (1-100)" for clustering. Then, this paper utilized the 'fit_predict' method to perform hierarchical clustering on the chosen features, assigning each data point to a cluster, and storing the clustering results in a new column named "Hierarchical_Cluster". Finally, a scatter diagram (Fig. 9) is drawn to visualize the clustering results and help us understand the clustering structure among data points.Hierarchical clustering creates a hierarchy of clusters without specifying the number of clusters K in advance.

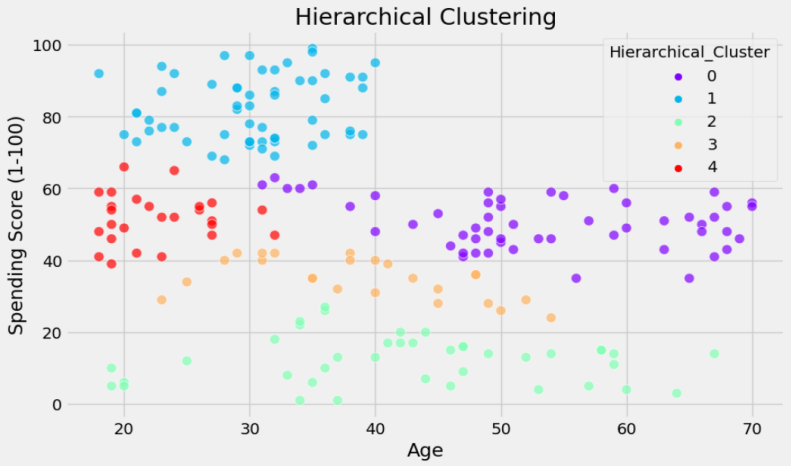

Fig. 9 Clustering using Hierarchical

4.3. DBSCAN

First, the features "Age" and "Spending Score (1-100)" are selected to cluster, creating an instance of 'StandardScaler scaler', which is used to standardize the selected features. Then 'fit_transform' method was used to standardize the selected features, and the feature values were converted to a standard normal distribution with mean 0 and standard deviation 1. Then a DBSCAN instance dbscan was created, and two parameters were set. 'eps' represented the threshold value of the neighborhood radius. 'min_samples' indicates the threshold for the minimum number of samples in the neighborhood. Finally, 'fit_predict' is used to perform DBSCAN clustering on the normalized data, and the clustering results are stored in a new column named "DBSCAN_Cluster". A scatterplot(Fig. 10) is drawn to visualize the DBSCAN clustering results. So that this paper can observe the distribution of different clusters in the feature Spaces of "Age" and "Spending Score (1-100)".Compared with K-means and hierarchical clustering, DBSCAN does not need to specify the number of clusters in advance when the density of data points is uneven.

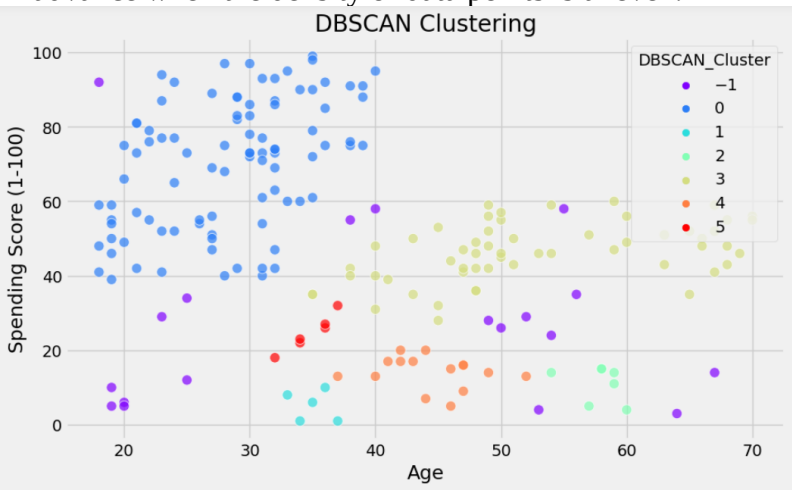

Fig. 10 Clustering using DBSCAN

4.4. Comprehensive

This paper discovered that each method has its benefits and situations in which it may be used by comparing the clustering results of the two variables of age and consumption score. Specifically, DBSCAN excels in datasets with uneven data density, identifies arbitrarily shaped clusters automatically, and can locate noise locations.

Additionally, the DBSCAN algorithm can determine the number and form of clusters automatically without requiring the user to predetermine the number of clusters, which eases the load of parameter selection. When working with data that has complicated geometries and uneven density distributions, DBSCAN performs better than conventional K-means and hierarchical clustering techniques [8]. This presents us with the chance to develop a more thorough understanding of the traits and requirements of our clients, which can direct business decisions.

It's also crucial to keep in mind that DBSCAN is highly dependent on the neighborhood radius and minimum sample size parameters that are chosen. To get the optimum clustering effect when using DBSCAN, these settings must be changed in accordance with the data's actual circumstances. Additionally, DBSCAN's performance could be constrained for high-dimensional data, so you should consider its benefits and drawbacks before using it [9].

In conclusion, this paper can understand the benefits and applicability of the DBSCAN algorithm for this specific dataset by comparing several clustering algorithms. DBSCAN is a potent tool for user activity research and consumer segmentation since it can automatically cluster, adapt to clusters of various shapes, and identify noise. To better fulfill the demands of complicated data analysis, future research can examine how to combine various clustering methods and parameter optimization techniques [10].

5. Conclusion

The use of clustering techniques such as k-means, hierarchical clustering, and DBSCAN was utilized in this study to examine client behavior patterns. This article gained novel insights into consumer segmentation by utilizing age, annual income, and consumption score as crucial factors. Particularly when handling datasets with variable data density and unusual cluster geometries, DBSCAN stood out as the top performer. Its versatility was demonstrated by its ability to automatically find clusters and noise locations without using specified cluster numbers.

DBSCAN's sensitivity to parameter selections, such as neighborhood radius and minimum sample size, must be taken into account. It might not be appropriate for data with high dimensions. The integration of various clustering strategies and techniques for parameter optimization should be investigated in further research. Despite its advantages, the clustering algorithm chosen should always be in line with the particular dataset and analysis goals.

In conclusion, DBSCAN provides a powerful tool for consumer segmentation and user activity analysis, but its effective deployment necessitates precise parameter adjustment. These approaches can be improved in future research for more complex and flexible customer insights.

References

[1] Teslenko D, Sorokina A, Smelyakov K, et al. Comparative Analysis of the Applicability of Five Clustering Algorithms for Market Segmentation. 2023 IEEE Open Conference of Electrical, Electronic and Information Sciences, 2023: 1-6.

[2] Fuchs M, Höpken W. Clustering: Hierarchical, k-Means, DBSCAN. Applied Data Science in Tourism: Interdisciplinary Approaches, Methodologies, and Applications. Cham: Springer International Publishing, 2022: 129-149.

[3] Kansal T, Bahuguna S, Singh V, et al. Customer segmentation using K-means clustering. 2018 international conference on computational techniques, electronics and mechanical systems, 2018: 135-139.

[4] Zakrzewska D, Murlewski J. Clustering algorithms for bank customer segmentation. 5th International Conference on Intelligent Systems Design and Applications, 2005: 197-202.

[5] Liang Jingzhang, Huang Xingshu, Wu Lijuan, et al. Clustering method of power load curves based on KPCA and Improved K-means. Journal of South China University of Technology (Natural Science Edition), 2020, 48(6).

[6] Ji Tao, LIU Weijie, Duan Li. Power customer big Data behavior analysis using improved Gaussian Mixture model. Journal of Chongqing University of Technology (Natural Science), 2022, 36(5): 233-240.

[7] Monil P, Darshan P, Jecky R, et al. Customer segmentation using machine learning. International Journal for Research in Applied Science and Engineering Technology, 2020, 8(6): 2104-2108.

[8] Ahmad H P, Dang S. Performance Evaluation of Clustering Algorithm Using different dataset. International Journal of Advance Research in Computer Science and Management Studies, 2015, 8.

[9] Oyelade J, Isewon I, Oladipupo O, et al. Data clustering: Algorithms and its applications. 2019 19th International Conference on Computational Science and Its Applications, 2019: 71-81.

[10] Bartels C. Cluster Analysis for Customer Segmentation with Open Banking Data. 2022 3rd Asia Service Sciences and Software Engineering Conference, 2022: 87-94.