International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 03 | March - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

Customer Segmentation Using Enhanced K -Means Algorithm

Dr. Nagarajan .R , Jothi S , Harish M

Assistant Professor Department of Computer Science, Sri Ramakrishna College of Arts &Science PG

Student, Department of Computer Science, Sri Ramakrishna College of Arts &Science

PG Student, Department of Computer Science, Sri Ramakrishna College of Arts &Science

1. ABSTRACT

In the highly competitive retail sector, gaining insights into customer behavior plays a vital role in enhancing business growth and customer satisfaction. Retail organizations continuously generate large volumes of customer-related data through billing systems, loyalty cards, and online shopping platforms. However, manual analysis of such large datasets is complex, time- consuming, and inefficient. Customer segmentation provides an effective solution by grouping customers into meaningful categories based on their purchasing patterns and behavioral characteristics. Clustering algorithms, particularly K-Means, are commonly used for customer segmentation because of their simplicity and computational efficiency. Despite its popularity, the traditional K-Means algorithm has several limitations, including random selection of initial centroids, sensitivity to outliers, and inconsistent clustering outcomes. These drawbacks negatively impact the accuracy and reliability of customer segmentation. To address these challenges, this project adopts an Enhanced K-Means approach using the K-Means++ algorithm, which improves the centroid initialization process. The proposed system segments customers using key attributes such as purchase frequency, total expenditure, and product preference patterns. The system is developed using Python, and experimental analysis indicates that K-Means++ produces more stable, accurate, and well-defined clusters compared to conventional K-Means. The results highlight the effectiveness of enhanced clustering techniques in supporting targeted marketing, customer retention strategies, and personalized retail services.

Keywords: Customer Segmentation, K-Means++, Clustering, Machine Learning, Retail Data Analysis, python

2. INTRODUCTION

In the present digital age, retail businesses operate in a highly competitive environment driven by rapidly evolving customer expectations and market trends. Customers engage with retailers through multiple channels, including physical stores, e-commerce websites, mobile applications, and loyalty programs. These interactions generate vast amounts of data that hold valuable information about customer preferences, buying behavior, and spending habits.

Customer segmentation is a key analytical technique that divides customers into distinct groups based on shared characteristics and behavioral patterns. Effective segmentation enables businesses to design focused marketing campaigns, enhance customer satisfaction, and maximize profitability. Rather than applying uniform strategies to all customers, segmentation allows retailers to identify high-value customers and offer personalized products and services.

Machine learning techniques have significantly improved the effectiveness of customer segmentation. Among these techniques, clustering algorithms are widely used because they can discover hidden patterns in data without the need for labeled samples. The K-Means algorithm is one of the most widely adopted clustering methods

due to its simplicity and speed. However, traditional K-Means suffers from several limitations, including dependence on random centroid initialization, vulnerability to noise, and inconsistent clustering results.

To overcome these shortcomings, K- Means++, an improved variant of the K- Means algorithm, was proposed. K- Means++ enhances clustering performance by selecting initial centroids in a structured and distance-aware manner. This leads to faster convergence and improved cluster quality. This project focuses on applying the K-Means++ algorithm to retail customer data and evaluating its effectiveness in generating meaningful and reliable customer segments.

3. LITERATURE REVIEW

Customer segmentation has been extensively explored in the areas of data mining and machine learning. Early segmentation techniques primarily relied on demographic attributes such as age, gender, and income level. Although these methods provided basic insights, they were highly subjective and lacked the ability to scale effectively with large datasets.

As data mining techniques evolved, clustering algorithms such as Hierarchical Clustering, DBSCAN, and K-Means gained significant attention. Among these methods, K-Means emerged as a popular choice due to its low computational complexity and ease of implementation. However, numerous studies have pointed out its major limitation, particularly the issue of poor centroid initialization, which often leads to suboptimal clustering results.In the present digital age, retail businesses operate in a highly competitive environment driven by rapidly evolving customer expectations and market trends. Customers engage with retailers through multiple channels, including physical stores, e-commerce websites, mobile applications, and loyalty programs. These interactions generate vast amounts of data that hold valuable information about customer preferences, buying behavior, and spending habits.

Customer segmentation is a key analytical technique that divides customers into distinct groups based on shared characteristics and behavioral patterns. Effective segmentation enables businesses to design focused marketing campaigns, enhance customer satisfaction, and maximize profitability. Rather than applying uniform strategies to all customers, segmentation allows retailers to identify high-value customers and offer personalized products and services.

Machine learning techniques have significantly improved the effectiveness of customer segmentation. Among these techniques, clustering algorithms are widely used because they can discover hidden patterns in data without the need for labeled samples. The K-Means algorithm is one of the most widely adopted clustering methods due to its simplicity and speed. However, traditional K-Means suffers from several limitations, including dependence on random centroid initialization, vulnerability to noise, and inconsistent clustering results.

To overcome these shortcomings, K- Means++, an improved variant of the K- Means algorithm, was proposed. K- Means++ enhances clustering performance by selecting initial centroids in a structured and distance-aware manner. This leads to faster convergence and improved cluster quality. This project focuses on applying the K-Means++ algorithm to retail customer data and evaluating its effectiveness in generating meaningful and reliable customer segments.

4. METHODOLOGY

This project implements Customer Segmentation using Enhanced K-Means (K- Means++) entirely using Python in the Jupyter Notebook environment. Jupyter Notebook provides an interactive platform that allows step-by-step execution, visualization, and analysis of clustering results, making it suitable for machine learning–based projects.

The methodology consists of multiple stages starting from data collection to cluster evaluation.

3.1 System Architecture Overview

The overall workflow of the system follows these steps:

Customer data collection Data preprocessing

Feature selection and normalization Enhanced K-

Means++ clustering Cluster analysis and

visualization Performance evaluation

3.2 Data Collection

Customer data is collected from publicly available datasets or retail transaction records. The dataset includes attributes such as:

Customer ID Purchase frequency

Total spending amount Recency of purchase

Product category preferences

The dataset is stored in CSV format and loaded into Jupyter Notebook using Python libraries

3.3 Data Preprocessing

Raw customer data may contain missing values, duplicate records, and inconsistent formats. Data preprocessing is performed to improve clustering accuracy.

Steps involved:

Removing duplicate customer records

Handling missing values using mean or median methods

Converting categorical data into numerical form (if required)

Scaling numerical features using normalization techniques

Python libraries such as Pandas and NumPy are used for data preprocessing.

3.4 Feature Selection and Normalization

Only relevant attributes that influence customer behavior are selected for clustering.

Examples: Annual spending

Purchase frequency Recency

Since K-Means is distance-based, feature scaling is mandatory.

Standardization is applied to ensure that all features contribute equally.

3.5 Enhanced K-Means++ Clustering

Traditional K-Means initializes centroids randomly, which can lead to unstable clusters.

To overcome this limitation, K-Means++ is used.

Advantages of K-Means++: Better initial centroid

selection Faster convergence

Reduced clustering errors Improved stability and

accuracy

The clustering model is implemented using Scikit-learn in Python.

3.6 Cluster Formation Process

Initial centroids are selected using K-Means++ strategy

Each customer is assigned to the nearest centroid

Centroids are recalculated based on cluster mean

Steps 2 and 3 are repeated until convergence Final

clusters are generated

3.7 Visualization of Clusters

Data visualization helps in understanding customer segments clearly.

Graphs such as:

Scatter plots and Cluster chats are generated using Matplotlib within Jupyter Notebook.

3.8 Performance Evaluation

The clustering performance valuated using

Inertia value (Within-cluster sum of squares)

Silhouette score

These metrics help measure cluster compactness and separation.

PSEUDO CODE FOR ENHANCED K-MEANS (K-MEANS++)

Algorithm: Enhanced K-Means++ for Customer Segmentation

Input:

Customer dataset D Number of clusters K Output:

Clustered customer groups Begin

Load dataset D into Jupyter Notebook Remove

duplicate and missing values Select relevant

features from D Normalize selected features

Initialize centroids using K-Means++ method

Repeat

For each data point in D:

Calculate distance to each centroid

Assign data point to nearest centroid For each cluster:

Recalculate centroid as mean of cluster points

Until centroids do not change Evaluate clustering

performance Visualize clusters using plots End

3.9 Tools and Environment Used Programming

Language: Python 3.x

Development Environment: Jupyter Notebook

Libraries Used: Pandas – Data handling

NumPy – Numerical computation Scikit-learn – K-

Means++ clustering

RESULTS AND ACCURACY

The proposed Enhanced K-Means (K- Means++) algorithm was implemented using Python in the Jupyter Notebook environment and evaluated using the Mall Customer Dataset. The dataset contains customer information such as Customer ID, Annual Income, and Spending Score, which are commonly used attributes for customer segmentation in retail analytics.

After preprocessing the dataset, the K- Means++ algorithm was applied to cluster customers into distinct groups. The optimal number of clusters was determined using the Elbow Method, which indicated that five clusters provide effective segmentation. Each cluster represents a unique customer group with similar purchasing behavior and income patterns.

Compared to the traditional K-Means algorithm, K-Means++ showed improved cluster stability and better separation between customer groups. The enhanced centroid initialization reduced random variations in clustering results and led to faster convergence. Visualization results clearly demonstrate well-defined clusters with minimal overlap.

To evaluate clustering quality, internal performance metrics such as Inertia (Within-Cluster Sum of Squares) and Silhouette Score were considered. The K- Means++ algorithm achieved a lower inertia value and a higher silhouette score compared to standard K-Means, indicating improved clustering accuracy and compactness. These results confirm that Enhanced K-Means provides more reliable and meaningful customer segmentation for retail applications.

After applying the Enhanced K-Means (K-Means++) algorithm on the Mall Customer dataset, customers were grouped into meaningful clusters based on Annual Income and SpendingScore. A sample of the clustered output is shown below:

Customer ID	Annual Income (kS)	Spending Score	Cluster
1	15	39	Cluster 0
2	16	81	Cluster 3
3	17	6	Cluster 1
4	18	77	Cluster 3
5	19	40	Cluster 0

CONCLUSION

Customer segmentation plays a vital role in understanding customer behavior and improving business decision-making in retail organizations. This project successfully implemented an Enhanced K- Means clustering approach using the K- Means++ algorithm to address the limitations of traditional K-Means.

By applying the proposed method to the Mall Customer Dataset, the system effectively grouped customers based on spending behavior and income patterns. The improved centroid initialization in K- Means++ resulted in more stable clusters, better accuracy, and faster convergence compared to standard K-Means.

The experimental results demonstrate that Enhanced K-Means is a suitable and efficient approach for customer segmentation in retail analytics. The project highlights the importance of using advanced machine learning techniques to extract valuable insights from customer data, which can support targeted marketing, customer retention, and personalized service strategies.

FUTURE ENHANCEMENT

Although the proposed system produces effective customer segmentation results, several enhancements can be considered for future work. Additional customer attributes such as age, gender, purchase frequency, and product categories can be included to achieve more detailed segmentation.

Advanced clustering algorithms such as DBSCAN, Hierarchical Clustering, or hybrid models combining clustering with classification techniques can be explored to improve performance further. Thes

The system can also be extended to handle real-time customer data using big data platforms.

Integration with recommendation systems and customer relationship management (CRM) tools can enhance business value. Moreover, deploying the model as a web-based or cloud- based application would allow organizations to perform dynamic and scalable customer segmentation in real-world environments.

REFERENCES

Arthur, D., & Vassilvitskii, S. (2007). K-Means++: The Advantages of Careful Seeding. Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms.

Jain, A. K. (2010). Data Clustering: 50 Years Beyond K-Means. Pattern Recognition Letters.

Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers.

Scikit-learn Documentation. Clustering Algorithms.

https://scikit-learn.org

Kaggle. Mall Customer Segmentation Dataset.

https://www.kaggle.com