

Paper Review

Main idea

Title: "K-Means Clustering Approach for Intelligent Customer Segmentation Using Customer Purchase Behavior Data"

Authors: Kayalvily Tabianan, Shubashini Velu, and Vinayakumar Ravi

Motivation: This paper aims to improve the understanding of customer purchase behaviors and to tailor marketing strategies effectively in order to maximize the profit of each customer by selling the right product to the right customer. The study is significant in the context of both the growing importance of personalized marketing in online retail to reduce the risk of losing customers and exploration of gaining more potential customers to increase e-commerce business exposure in this competitive world.

Summary

The paper presents a K-Means clustering approach to segment E-commerce customers based on purchase behavior data. It highlights the importance of understanding customer interactions for targeted marketing. The study details the use of customer data from a Malaysian E-commerce platform to demonstrate the effectiveness of the K-Means method over other algorithms. Key findings include patterns in product views versus purchases and peak active times for online shopping. The paper suggests improvements in data handling and proposes deep learning for future research to enhance customer segmentation.

Approach and Contributions

The authors use Malaysia's E-commerce dataset in the last 3 years for accuracy in this analysis, and choose the sample which should show the existence of inequalities in marketing by considering $k=3$ main clusters (event type, products, and categories) and revolve around three target groups of age, gender, working adults or students age. Besides, they also use results from quantitative research questionnaires as fundamentals and perform the data pre-processing before using the improved K-Means clustering approach.

The paper consists of two different variations of K-Means algorithm, the *SAPK + K-Means algorithm* (as an improved way with smaller square error result by searching for the optimal centroids in each cluster and also maximizing the objective profit function) and *AP + K-Means method*, and compares their respective results in terms of accurate rates and time spent. In this case, the SAPK + K-Means algorithm won based on a higher accurate rate but reasonably with a longer time.

The main findings from the graphs reveal that customers predominantly view products, particularly in the appliances and electronics categories, more frequently than adding them to the cart or purchasing. This suggests a pattern of extensive product comparison before purchase decisions. The application of the K-Means algorithm for segmenting customer groups based on event types (view, purchase, cart) and product categories (favorite to least favorite) provides insights into customer preferences and behaviors. The time analysis indicates that the peak activity on the online store occurs from around 2 pm to 4 pm. The temporal analysis reveals peak activity times, which can be crucial for scheduling marketing campaigns or promotions.

The contributions from the analysis shown in the figures are important as they provide a data-driven way to understand customer behaviors through machine learning. This enhances the efficiency of targeting efforts, leading to potentially higher conversion rates and better customer retention rates, which are key metrics for E-commerce success.

The paper builds on understanding of hierarchical clustering, K-Means clustering, and mean shift clustering, and chooses the most appropriate method of K-Means clustering for customer segmentation. It not only stands with stability that shows less variability in the construction of the three clusters but is also representative of the same segment of municipality.

Areas for Improvement

The study could further explore the applicability of its methods across different markets and E-commerce platforms to generalize its findings. The volatility in small datasets can significantly impact clustering and model classifications. To improve, future work should focus on smoothing or standardization processes for more effective data mining. Additionally, exploring deep learning methods for customer segmentation could offer enhancements over traditional machine learning approaches, given their recent successes in various applications. Further research should also aim to include larger and more varied datasets to mitigate volatility and improve model robustness.