Customer segmentation application based on K-Means

Jiaqi Zhao

School of Accounting, Southwestern University of Finance and Economics, Chengdu, China

42013011@smail.swufe.edu.cn

Abstract. Customer segmentation(CS) is a crucial aspect of customer relationship management, widely utilized by industries, banks, and consulting companies. However, the intricate data relationship between individuals presents significant challenges in customer segmentation research. Fortunately, machine learning has made remarkable progress in processing big data, and its exceptional performance has captivated the attention of business analytics researchers. Based on this, numerous customer segmentation methods based on machine learning have been proposed. This paper aims to review the papers published after 2010 on customer segmentation, and summarize the current status and importance of customer segmentation in implementing marketing strategies. Additionally, it introduces two primary types of customer segmentation scenarios, and summarizes the common combination of analysis models and machine learning algorithms in customer segmentation. Finally, the paper introduces a customer segmentation method based on k-means and provides a perspective on the future development of customer segmentation.

Keywords: RFM Analysis, Customer segmentation, K-Means Clustering.

1. Introduction

Customer segmentation is a business method that aims to divide customers into specified groups and provide refined operation accordingly, which is widely used by corporations, banks, and consulting companies. It has been taken immense attention and has extensively been used in strategic marketing [1]. The basic process of applying customer segmentation is that business entities study the difference in characteristics of customers in several aspects, such as geography, demography, and purchasing behavior, and then decide which way to group the customers is meaningful for achieving the entity's marketing goals. By performing this process of customer segmentation, business entities can produce, budget, and market with an insightful understanding of the market's states, features, and demands, therefore increasing precision and efficiency in resource allocation. Customer segmentation can also provide some knowledge of forming business strategies that are mostly compatible with the industry and the target market, which is helpful for the entity's operation in a long-term consideration. In summary, customer segmentation is crucial in customer relationship management (CRM), which is gradually becoming essential in business operations nowadays.

In business operations, managers and analysts establish a specific customer segmentation method suitable for a particular business project. First, they determine factors that the customers are grouped by, and the weight and interrelationship between these factors establish the fundamental thinking for the project’s marketing strategies and other subsequent business process that builds upon. To perform customer segmentation, analysts often use data science techniques, including data collecting, data cleaning, data processing, data analytics, etc., with abundant information on the project’s target market, potential customers, and the segmentation market as a whole. Last but not least, bind with business knowledge about marketing, business management, etc.

Data mining and machine learning are the most widely used techniques for customer segmentation nowadays, thanks to the development of data science in the past 2 to 3 decades. Combined with the rapid development of computer science and technology, the relevant technical fields involved in customer segmentation have been relatively mature. As for customer segmentation’s requirement for a mass amount of data, there are abundant quantities of open resources of datasets online, such as the statistics that the government reveals on their official websites and companies’ annual reports on SEC websites. In today's era of big data, there is an overwhelming amount of information available on the internet, with tens of exabytes of data being generated every day. This abundance of data makes it extremely convenient for analysts to access and analyze the required information. Raw data can also be acquired by using web data crawling- commonly by using coding techniques like Python codes,- to collect legal data that lacks collection and proper organization. However, a lot of data generated from institutions will not be open to the public out of information secrecy and other reasons, which considerably limits the ability for practitioners to acquire datasets for analytic use.

A crucial process in customer segmentation is determining the factors in the thinking model. Although business entities share the same purpose in applying customer segmentation techniques, which is to divide heterogeneity into homogeneous forms [2], they may group their customers based on considerably different mindsets. One of the most common ways of customer segmentation is by geographic differences, which considers customers within identical countries, states, or other areas to be classified in a segment. Another commonly used way of customer segmentation is by demographic features, which divides segment markets based on the population within a specific area. A way to perform customer segmentation more precisely and accordingly is by economic behaviors. It is based on the idea that people who have different ages, genders, fields of jobs, educational backgrounds, cultural backgrounds, etc. tend to have explicitly other purchasing behaviors. A common customer segmentation model, RFM (recency, frequency, monetary) analysis, is based on purchasing behavior. Nowadays, with the mass production of data and the advancement of marketing knowledge, it is becoming more and more important for business entities to provide customized products and services. This drives companies to combine various features of customers into studies, forming knowledge of customers with a more diverse perspective and comprehensive understanding.

2. Customer Segmentation Application

The procedure of forming a marketing strategy consists of several steps, including the establishment of the product’s core competitiveness, identifying potential customers, and setting up the product’s distribution network. Using analytic techniques to study customer data and identify target customers, customer segmentation serves a vital role in the preparation phase of marketing. As the first and foremost step in the formation of marketing strategy, customer segmentation answers what group of people, which could be age group, gender group, career group, income group, group of people who share the same hobby or group of people who share similar daily routine, etc., are most likely to purchase the particular product that the company is promoting. However, in reality, marketing staff usually takes multiple factors into account and creates a multi-dimensional method of measurement, because the selection of appropriate methods can reveal useful hidden patterns in customer segments [3].

Customer segmentation also serves as the foundation of a marketing strategy. Only after the marketing department understands the customers that they are most possibly serving by using customer segmentation, they can successively make a plan of how they want to sell the products and later implement it. In summary, customer segmentation identifies potential customers and builds a foundation for the subsequent marketing procedure. The quality of customer segmentation significantly affects the efficiency and productivity of a marketing strategy in implementation.

2.1. Entering a New Market

One of the customer segmentation goals is to choose a new market to enter. This could be applied to two roughly sorted conditions: either when a start-up company with one fully developed core product makes its first try to enter the market, or when a company has developed a new type of product and the existing marketing strategies suitable for old products doesn't adapt to the new product. Start-up companies and new product departments have almost zero target customer information. In this situation, the prime priority for marketing managers' CS decision is to accurately locate a group that is suitable to be the target customers. The idea of choosing a new market is relatively simple, which is to examine a big scale of information about potential customers and narrow the options down by a selected accordance. When applying customer segmentation techniques, usually, a large proportion of external customer information is used. For example, start-up companies use demographic data from official government websites and other open resource datasets for customer segmentation. Also, without accumulated experience in marketing particular products, those companies and departments could appear to be inclined to generality as feature selection in segmentation.

2.2. Re-allocating Marketing Resources

Another customer segmentation goal is reallocating the marketing resources invested in one or more products. As the number of transactions grows, companies increasingly accumulate first-hand information on their customers, which allows them to examine the collected information and use the analytic result as an aid to improve their marketing strategy. This situation could be applied to medium or large corporations that have already developed a certain amount of market shares and require further, better subdivision of the market to increase operational efficiency and reduce costs. In this scenario, customer segmentation helps the company analyze both external customer information acquired from marketing research and internal customer information from the company's historical sales records. Therefore, the company's marketing department can better understand the features of customers who are attracted to the company's products. Customer segmentation may also help study the difference between the historical target customer group and the historical actual customer group, from which the marketing department can identify the existing deficiency and discrepancies in their customer targeting so that they can make reasonable adjustments to their current marketing strategy.

3. Customer Segmentation Methods

In today's business world, as digital transformation picked up stream [4], business acts are increasingly receiving practical assistance from computerized and analytic tools. Naturally, this trend brings significant influence to customer segmentation techniques. In the past decade, most of the customer segmentation tasks were performed with an analytical model and one machine learning algorithm. The RFM (Recency, Frequency, and Monetary) analysis model helps the marketing department to build the basic thinking framework on how the customers' information will be analyzed. The model determines the factors that the potential customers will be valued by. The K-means clustering algorithm is a machine learning algorithm that conducts the segmentation task and puts out the segmentation result.

3.1. RFM Analysis

RFM analysis, first proposed in 1994 by Hughes, is one of the most commonly used customer segmentation models in today's business world. RFM is the abbreviation for recency, frequency, and monetary, which respectively represents the time of last purchase, the frequency of purchase in a specific period, and the revenue received from a particular customer. Attributed to its focus on studying customers' purchasing behavior, RFM analysis shows a high level of relativity and accuracy in most of its business analysis applications, such as customer lifetime value measurement, customer segmentation, and behavior analysis.

The process of building the RFM model is as follows. First, use each dimension of the RFM model to sort the studied database and divide the customer into several equal segments. The factor used to quantify the three dimensions varies flexibly with the database. For example, the monetary dimension could be measured by both the dollar amount of a single customer's average spent per purchase and the total spent during a certain period. Though it was suggested that it is better to use average spent rather than the total spent for statistical feasibility, but in some cases the reverse is true. Secondly, create a scoring method for sorted data. Rather than simply giving a score to each number with one multiplier in a linear method, the scoring method could use unequal intervals to be more compatible with the database. Lastly, make a single RFM score with a reasonable weighting scheme. The composite RFM value is the product of the normalized RFM score of each data point and the correspondent weight of that RFM factor [5].

The advantage of applying the RFM model rather than other segmentation methods that mainly consider geography factors or population factors, is that the RFM model focuses on customers' consuming behavior that can portray the customers in a more specific, multi-angled way. In most cases, behavioral factors appear to be better indications of future customer purchasing patterns. The RFM model also quantifies the factors so that they can be explicitly and objectively measured. However, based on prior studies, RFM values are inclined to be firm-specific and based on the nature of products [5]. It might suggest a certain degree of subjectivity in RFM analysis, whose analytic result, to some extent, depends on analysts' judgment and professionalism.

3.2. K-Means Algorithm

The age of big data has resulted in more and more open access to machine learning algorithms for business applications. Among all sophisticated ML algorithms, clustering algorithms are the most compatible in accomplishing customer segmentation tasks which are already widely used in business analysis. The most popular clustering algorithm used in customer segmentation is k-means clustering.

Data points are clustered when they have least Euclidean distance in K-means algorithm. K-means uses the average value as the center of the cluster. In application, firstly, analysts choose the k value as the number of clusters they think is appropriate for the segmentation case. Then, the computer program randomly divides the data into k clusters and obtains the cluster centers using Euclidean distance. After that, the program recalculates the cluster center. The recalculation repeats until there is no modification to the cluster. The result is k groups of data that are closest to their cluster center in the sense of average distance considering all dimensions of the data.

The simplicity and reliability of k-means clustering theorem make k-means algorithm both easy to apply and also ideal in validity. In application, its simplicity allows high processing speed, which makes the analytic process more efficient. In addition, the k-means algorithm promises a certain level of goodness of fit. Although k-means clustering could be trapped into local optimum, in most cases, a k-means clustering result is considered qualified for a business analysis problem. However, there are some disadvantages to k-means clustering. Firstly, the result of k-means clustering has a high level of randomness. A different choice of parameter k could lead to a different clustering result. K-means clustering is also sensitive to the initiate cluster center value whose variety causes the clustering result to be unstable. The clustering result could also be altered by anomalous values. Secondly, the effectiveness of the k-means clustering result is limited. Since the clustering theorem of this algorithm is by average distance, this algorithm can work effectively on samples that are aggregative but might be less efficient on dispersed data. In addition, since all dimensions of data are being averagely considered in k-means clustering, it is possible that the importance of certain variables is overestimated or underestimated. This inaccuracy of the weight of variables could make the clustering result less fitting to the real-life situation represented by the data.

Fortunately, in application, some remedies can be done to make up for the disadvantages of the k-means algorithm. For example, to lower the uncertainty of the clustering result, the sample points could be preprocessed and cleaned by eliminating isolated data points and low-density data points [6]. This process--data cleaning--also has practical meaning in addition to statistical processing efficiency because, in target marketing, it is more efficient to focus on homogeneous customer groups and not allocate much of the resources to single, centrifugal individual customers.

Besides the k-means algorithm, there are several ML algorithms that can be seen used in customer segmentation applications. Some of these alternative algorithms to some extent avoid the disadvantages of the k-means algorithm. For example, to deal with the uncertainty of clustering, sometimes the fuzzy c-means algorithm is used instead of k-means [7]. Sometimes, density-based clustering methods that provide natural protection on outliers are used, such as density-based connectivity clustering or density-function clustering. Some other alternative algorithms are improvements based on k-means algorithms. Recent years have witnessed some exciting developments in ML algorithms specifically used for customer segmentation. In 2017, a group of researchers published their invention of a clustering algorithm called PerTreeClust, which presents and compares each customer's personal purchasing records [8]. In 2021, some researchers published their k-means-based customer segmentation method GPHC (Gaussian Peak Heuristic Clustering), compositing entropy method, genetic algorithm, and hierarchical clustering, whose robustness when facing complicated customer information has been recognized by business professionals [9]. Moreover, other researchers comply with additional algorithms to improve the clustering effect of the K-Means algorithm. Researcher Yue Li, Dong Tian, and their team members use the ALPSO (adaptive learning particle swam optimization) algorithm to prevent the K-means algorithm's dependence on initial cluster centers [10]. Compared to simple k-means clustering, those improved clustering methods have more advantages in the sense of completeness and feasibility. As the newly developed specialized clustering algorithms continue to emerge, the application of customer segmentation will be increasingly sophisticated and intelligent.

4. Conclusion

This study is based on prior academic research on marketing management theory and customer segmentation techniques. The study reviewed two types of applications of customer segmentation and the main method of conducting customer segmentation. Customer segmentation is an essential step in the preparation phase of target marketing. It has two main application scenarios, which are the time for start-up companies to enter a market and the time for medium or large companies to re-allocate their marketing resources based on their historical selling records. Today's customer segmentation procedure mainly consists of the combination of the RFM - analysis model and the k-means clustering algorithm. The techniques can put out reliable results in most cases but are defective because potential subjectivity of model building and the inner uncertainty of the algorithm. Some recent developments of specified customer segmentation algorithms are listed.

References

[1] Hiziroglu A. Soft computing applications in customer segmentation: State-of-art review and critique[J]. Expert Systems with Applications, 2013, 40(16): 6491-6507.

[2] Das S, Nayak J. Customer segmentation via data mining techniques: state-of-the-art review[J]. Computational Intelligence in Data Mining: Proceedings of ICCIDM 2021, 2022: 489-507.

[3] Ernawati E, Baharin S S K, Kasmin F. A review of data mining methods in RFM-based customer segmentation[C]//Journal of Physics: Conference Series. IOP Publishing, 2021, 1869(1): 012085.

[4] Peker S, Kart Ö. Transactional data-based customer segmentation applying CRISP-DM methodology: A systematic review[J]. Journal of Data, Information and Management, 2023: 1-21.

[5] Wei J T, Lin S Y, Wu H H. A review of the application of RFM model[J]. African Journal of Business Management, 2010, 4(19): 4199.

[6] Patel V R, Mehta R G. Modified k-means clustering algorithm[C]//International Conference on Computational Intelligence and Information Technology. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011: 307-312.

[7] Yuliari N P P, Putra I K G D, Rusjayanti N K D. Customer segmentation through fuzzy C-means and fuzzy RFM method[J]. Journal of Theoretical and Applied Information Technology, 2015, 78(3): 380.

[8] Wu J, Shi L, Lin W P, et al. An empirical study on customer segmentation by purchase behaviors using a RFM model and K-means algorithm[J]. Mathematical Problems in Engineering, 2020, 2020: 1-7.

[9] Chen X, Fang Y, Yang M, et al. Purtreeclust: A clustering algorithm for customer segmentation from massive customer transaction data[J]. IEEE Transactions on Knowledge and Data Engineering, 2017, 30(3): 559-572.

[10] Sun Z H, Zuo T Y, Liang D, et al. GPHC: A heuristic clustering method to customer segmentation[J]. Applied Soft Computing, 2021, 111: 107677.