2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)
Amity University, Noida, India. Sep 3-4, 2021

2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO) | 978-1-6654-1703-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICRITO51393.2021.9596118

A Survey on Customer Segmentation using Machine Learning Algorithms to Find Prospective Clients

Vaidisha Mehta
CCE, SCIT,
Manipal University, Jaipur
vaidisha.s.mehta2001@gmail.com

Ritvik Mehra
CCE, SCIT,
Manipal University, Jaipur
ritvikmehra17@gmail.com

Sourabh Singh Verma*
CCE, SCIT,
Manipal University, Jaipur
ssverma80@gmail.com

***Abstract*— Numerical classification has occupied an important position in economic development due to the rapid growth of the world's capital and society. The focus is to enhance strategic governance by improving some common virtues regarding service quality systems. Market segmentation not only grasps the customer's geographic, demographic, psychographic and behavioral traits but also governs the market according to the consumer's needs, by undertaking the correlating procedure to establish a better retailer-consumer relationship. This can be achieved by a concept that comes under unsupervised learning, known as cluster analysis.**

In this method, customers are divided into 'n' number of groups based upon the above-mentioned traits, clustering means grouping the information based on similarities in the dataset. Customers belonging to a particular group have some common traits. Customers are grouped in a way that a customer belonging to a particular group shares a common interest with other customers of the same group. In this paper, we have studied and explained various algorithms related to clustering which help segment customers according to their needs. With respect to their datasets, these algorithms are analyzed and compared using metrics like Silhouette and Davies Bouldin.

Keywords*— *Customer Segmentation, k-means, clustering, categorization, Silhouette, Davis Bouldin

I. INTRODUCTION

During COVID-19 we have seen companies and proprietors suffering and going through a huge loss because they were unable to cope up. So, the main motive behind this study is to analyze their potential and profitable customers and in saving resources by approaching only those customers who are genuinely interested in purchasing their products. The categorization of prospects and customers into distinct groups is based on a variety of features and variables such as their age, gender, needs, buying characteristics, habits, income, etc.

Customer segmentation comprises of various stages including:

Foundation: Establishing business goals aligning with an overall strategy. The scope can be defined by either the number of customer accounts to be analyzed or by the maximum number of segments to be created.

Analysis: Determining a set of criteria for inferring customer value across the market depending on the nature of business and service.

Data collection: Data, the most important marketing tool can be collected by developing a detailed research plan that defines the resource and method from which the data can be extracted. Be aware not to include data without integrity such as incomplete data, outdated data, non-standardized data, and data that necessitates quantitative judgment.

Synthesis: Uncover what your target customers are like by discovering their common interests, habits and so on to define a look-alike audience. Use all this data to create a market segment which then helps to deliver one's demand generation goals by allowing one to focus resources on the most profitable customers.

To handle a large customer base, segmentation is done to find similar patterns to gain a competitive advantage in terms of cost and efficiency within each segment. Grouping the information based on some similarities in the dataset is known as clustering where the instance of each data point belongs in exactly one cluster. However, some sort of measure is also required to apply in the network so that any attack can be avoided or recovered else it will hamper the result of the analysis. Clustering algorithms have been classified into hierarchical and partitional clustering algorithms. Clusters are created based on top-down hierarchy referred to as divisive or bottom-up hierarchy referred to as agglomerative, whereas in partitional algorithms various partitions are created and evaluated based on some criterion [1-6][13-15]. Behavioral segmentation is one of the methods which suggest clustering based on customer habits [1]. Such clusters help to penetrate the right customer base for advertising and product identification. We found that most researches are being done to implement an algorithm to implement customer segmentation but only a few have done any such comparison of different methods. This motivates us to write this article that includes comparisons of different such methods.

II. TYPES MARKET SEGMENTATION

A. Demographic segmentation

Demographic segmentation is a process of splitting customer groups based on traits such as age, gender, ethnicity, income, level of education, religion, and profession [1][9]. By using demographics as a metric, targeting customers get more straightforward. The strategic development process becomes easier as end-user outliers are easy to understand. Whereas receiving faulty data within a given region may produce undependable assumptions thereby affecting the accuracy of marketing methods.

B. Psychographic segmentation

Psychographic segmentation allows incredibly effective marketing by grouping customers at the more personal level by defining their hobbies, personality traits, values, life goals, lifestyles, and beliefs [1][10]. It provides a detailed analysis of customers' likes and dislikes and their buying preferences thus providing more accurate details regarding their potential customers. insight into the motive behind consumer behavior, providing a more accurate picture of what makes their potential customer. However, this type of data is difficult to obtain and so to ensure consistency in this method, the data interpretation must be done with well-defined rules.

C. Geographic segmentation

Geographic segmentation allows many different kinds of considerations when advertising to consumers by grouping them based on their geographic location such as their country, region, city, and even postal code [1][8]. It saves resources by selecting a particular target area. Focusing marketing on a specific area helps to gain a deeper knowledge of customers and competitors. It is easy to determine the location of customers. The development of a promotional mix often depends on geographic variables. Although companies do not often rely on geographic segments to determine their target market as they usually combine demographic and psychographic variables to produce more accurate market targeting.

D. Behavioral segmentation

Behavioral segmentation is perhaps the most useful of all e-commerce businesses as most of this data such as customers' spending habits, browsing habits, purchasing habits, loyalty to a brand, interactions with the branch, and previous product ratings can be gathered via the website itself [1][7]. Based on their spending scores, their purchasing habits, how often they interact with the brand and how loyal they are with these brands, brands build their targeted consumer segments. One needs to be more precise while collecting this type of information as behavior and motives vary greatly from person to person making resultant data difficult to interpret.

III. Major Technologies For Customer Segmentation

A. K-Means Algorithm

This algorithm is based on partitioning principle. It introduces the elbow method to identified 'k', the total number of partitions. Sum Square Errors is used to measure nearest centroid to the given data points, where it may use different values of 'k'. The optimize K value is calculated such that SSE is minimum [1][4][11-13].

Sum Square Errors (SSE) is:

$$\sum_{i=1}^{k} \sum_{X_j \in S_i} ||Xj - \mu i||^2$$

Xj = data point in Si cluster
μi = centroid of the cluster

Then Euclidean data points are allocated to the closest centroid forming the cluster. Then a repetitive process of calculating barycentre's by the means of the cluster is carried out until there is no change in centroid position. This is the quickest centroid-based algorithm, it is very lucid and can scale up a large amount of dataset and it also minimizes intra-cluster variance measure but still suffers from local minimum problem. This algorithm cannot deal with datasets of non-convex shapes and faces complications to predict the best k value. K-means algorithm is used when there is even cluster size, flat geometry, not too many clusters and for general purposes.

B. Agglomerative Algorithm

This algorithm uses dendrograms which are based on formal hierarchy and help to form a memory for recording how the clusters have been formed. N clusters are formed for different data points and merge the closest data points in a way that each step consists of one less cluster than the previous cluster [1][4][13]. Embedded flexibility regarding the level of granularity using dendrogram. This algorithm can handle any form of similarity or distance, but it is computationally expensive, and it cannot handle outliers. Also, the word's algorithm usually generates equal size clusters. An agglomerative algorithm is usually used when there are possible connectivity constraints and is based on non-Euclidean distance having many clusters.

C. Divisive Algorithm

This is a top-down hierarchical approach in which all the observations begin in one cluster and are then repeatedly divided into different clusters and the results are in the form of dendrograms. In each step, based on the distance between two clusters called the specific linkage criteria, the two closest clusters are merged. From the range of maximum distance and placing a cut-off line at that position, the required number of clusters can be selected from the dendrogram. It simply indicates that the formed clusters are at maximum distance from each other and divergence can be made among them [4][13].

D. Mean-Shift Algorithm

It is a process of iteratively shifting data points towards mode to form a hill of the clusters[1][5][15]. Where mode is the highest density of data points. It is different from k-means as it does not require knowing the number of clusters in beginning, the number of clusters is derived from data.

$$f(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

Where h = bandwidth
K = kernel

Bandwidth represents the number of data points belonging to the cluster where the small value of it will make convergence hard to achieve, whereas convergence can be easily achieved by a larger value of it.

E. GMM

In one dimension Gaussian Distribution, the probability density function of a is given as below:

$$G(X|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

Where μ = mean of the distribution
σ^2 = variance of the distribution

For Multivariate i.e. d-variate Gaussian Distribution, the probability density function is given as below:

$$G(X|\mu,\sum) = \frac{1}{\sqrt{(2\pi)\ |\sum|}} \exp\left(-\frac{1}{2}(X-\mu)^T\sum^{-1}(X-\mu)\right)$$

Here μ represents a d dimensional vector denoting the mean of the distribution and $\sum$ represents the d X d covariance matrix [5][15]. This algorithm is robust to outliers, it provides BIC score for selecting parameters and it converges fast given good initialization. Although this algorithm is highly complex and can be slow. GMM is good for density estimation and flat geometry.

F. DBSCAN Algorithm

It helps to identify noise by classifying the border datapoints and core data points. It uses two basic components these are eps and min points. Eps is the measurement of distance for a given point in a close neighborhood. The configuration of a cluster can be changed by changing eps and min points in the radius. [5][15].

IV. Measurement Of Accuracy

A. Silhouette Score

It allows measuring the accuracy of data points allocated in a correct cluster. A higher Silhouette score indicates more accurate clusters.

Silhouette Score = $\frac{b-a}{\max(b,a)}$

Where, b = average distance between the data point and closest cluster data points

a = average distance between the centroid of a cluster and data points embroiled into it

B. Davies Bouldin Score

This score is identified as the average similarity measure of each cluster with its important similar cluster that is based upon the ratio between distance within a cluster and between clusters [12]. Note that If the value of D_{ij} is more, clusters are not very distinct.

TABLE I
EXISTING MODEL COMPARISONS

	Algorithm	**Dataset**	**Clusters**	**Accuracy**	**Testing Result**
[1]	K-Means	Local retail shop – 200 sample	5	Silhouette Score = 0.55	K-Means obtained the highest Silhouette Score. Here, unlabelled dataset has been used.
[8]	K-Means	Mall data – 200 samples	5	In K-Mean's algorithm, the time complexity: O(nki) and space complexity: O(k+n).	K-Means clustering gives better performance than hierarchical clustering in case of larger dataset as the complexity is k-means is O(nki) while that of hierarchical algorithm is $O(n^3)$
[2]	K-Means	Retail supermarket – 2138 samples	4	Null hypothesis is tested by using ANOVA method.	It was clear that the possible segments existed in the given target customer population by the cluster analysis of the chosen sample of respondents.
[1]	Mean Shift Algorithm	Local retail shop – 200 sample	7	Silhouette Score = 0.53	Mean-shift algorithm was unable to cluster the data precisely.
[5]	Mean Shift Algorithm	Crawling data collected from an e-commerce company when customers browse their website.	11		(Age prediction was done based on habit data). The histogram presented showed the access pattern of customers in each group.
[1]	Hierarchical Algorithm: Agglomerative algorithm	Local retail shop – 200 sample	5	Silhouette Score = 0.55	The clusters formed by agglomerative a k-means algorithm were same and hence, there Silhouette scores were similar. Hierarchical algorithm is mainly used for smaller dataset.
[5]	Hierarchical Algorithm	Crawling data collected from an e-commerce company when customers browse their website.	10	Execution time of this algorithm was 24:38 ± 0:42 seconds.	Here we were able to predict gender and age by in depth analysis of their respective datasets.
[3]	Hierarchical Algorithm: Divisive Algorith,	Commercial economic data set: 2000 samples	4	Time complexity: $O(n^2)$	This algorithm is relatively easier to apply and was proven advantageous over other algorithms for clustering the given dataset.

Davies Score (D_{ij}) = $\frac{\overline{d_i}+\overline{d_j}}{d_{ij}}$

Where, $\overline{d_i}$ = average distance between every data point in cluster i and centroid

$\overline{d_j}$ = average distance between every data point in cluster j and centroid

d_{ij} = Euclidean distance between centroids of two clusters

V. Review On Existing Models And Evaluation Comparison

We reviewed various machine learning algorithms through which customer segmentation can be performed. It consists of various algorithms applied on different datasets and the number of clusters formed in each one of them. These methods are compared for accuracy metrics Silhouette score and Davies Bouldin score as well as their time and space complexities followed by testing results of respective algorithms.

VI. Results And Discussion

From the above research, the K-Means algorithm has the best Silhouette and Davies Bouldin score. From this, it can culminate that K-Means Algorithm is more competent for customer segmentation.[1] K-Means takes O(n) time as compared to others like Hierarchical taking O(n^2) time [8]. If variables are huge, then K-Means is mostly computationally faster and it produces tighter clusters. Davies Bouldin and Silhouette Score of various algorithms are as shown in figure 1.

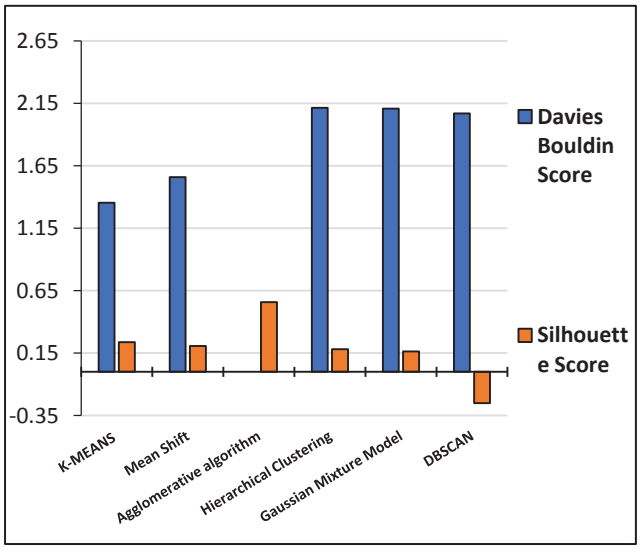

Fig. 1. Result Analysis of different methods using different accuracy scores

VII. Conclusion

For any product, the market is made up of several segments and customers differ a lot in their characteristics as they belong to heterogeneous lots. Thus, feedback from these heterogeneous segments(customers) makes it easier to devise a proper policy for such a market thereby making marketing efficient and economic. By distinguishing customers groups from one another within a market, it benefits not only the marketers by the customers too. Though sometimes market segmentation becomes a costly proposition as a marketer not only faces considerable difficulties but also, must develop different marketing mixes for various segments. Manufacturing a variety of products is costlier as compared to mass production.

Our objective of finding the most suitable algorithm for customer segmentation has been achieved. As we have reviewed above from some research papers, it can be inferred that the K-means algorithm is better than other clustering algorithms for customer segmentation in terms of performance and accuracy both which can also be seen from the obtained Davies and Silhouette scores of various clustering algorithms. The future work will involve more algorithms tested upon diverse datasets to claim even more accurate algorithms, if any.

References

[1] Tushar Kansal, Suraj Bahuguna , Vishal Singh, Tanupriya Choudhury "Customer Segmentation using K-means Clustering", IEEE, Year: 2018.

[2] Kishana R. Kashwan and C. M. Velu "Customer Segmentation Using Clustering and Data Mining Techniques", International Journal of Computer Theory and Engineering, Vol. 5, No. 6, 2013.

[3] Yifei Wang Research on the analysis of commercial economic data based on hierarchical clustering algorithm", IEEE International Conference on Power, Intelligent Computing and Systems (ICPICS), Year: 2020.

[4] Shreya Tripathi1, Aditya Bhardwaj, Poovammal E "Approaches to Clustering in Customer Segmentation", International Journal of Engineering & Technologgy, Year: 2018.

[5] Tran Anh Tuan, Tien-Dung Cao, Tram Truong-Huu "DIRAC: A Hybrid Approach to Customer Demographics Analysis for Advertising Campaigns", 6th NAFOSTED Conference on Information and Computer Science (NICS), Year: 2019.

[6] Chinedu Pascal Ezenkwu, Simeon Ozuomba, Constance kalu "Application of K-Means Algorithm for Efficient Customer Segmentation: A Strategy for Targeted Customer Services", (IJARAI) International Journal of Advanced Research in Artificial Intelligence, Vol. 4, No.10, Year: 2015.

[7] Wilhelmus Hary Susilo "An Impact of Behavioral Segmentation to Increase Consumer Loyalty: Empirical Study In Higher Education Of Postgraduate Institutions At Jakarta", 5th International Conference on Leadership, Technology, Innovation and Business Management, Year: 2016.

[8] Sulekha Goyat "The basis of market segmentation: a critical review of literature", European Journal of Business and Management Vol 3, No.9, 2011.

[9] Meghana N M "Demographic Strategy of Market Segmentation", INDIAN JOURNAL OF APPLIED RESEARCH Volume : 6, Issue : 5, 2016.

[10] Hui Liu, Yinghui Huang, Zichao Wang, Kai Liu, Xiangen Hu and Weijun Wang "Personality or Value: A Comparative Study of Psychographic Segmentation Based on an Online Review Enhanced Recommender System", MDPI, Year: 2019.

[11] Liang Li, Jia Wang, And Xuetao Li "Efficiency Analysis of Machine Learning Intelligent Investment Based on K-Means Algorithm", IEEE, Year: 2020.

[12] Bernad Jumadi Dehotman Sitompul, Opim Salim Sitompul, Poltak Sihombing "Enhancement Clustering Evaluation Result of Davies-Bouldin Index with Determining Initial Centroid of K-Means Algorithm ", The 3rd International Conference on Computing and Applied Informatics 2018.

[13] T. Sajana, C. Sheela Rani and K. Narayana, "A Survey on Clustering Techniques for Big Data Mining", Indian Journal of Science and Technology, vol. 9, no. 3, 2016.

[14] Sourabh Singh Verma, R.B.Patel, S. k. Lenka " Analyzing varying rate flood attack on real flow in MANET and solution proposal "Real Flow Dynamic Queue (RFDQ)" International Journal of Information and Communication Technology, Vol. 10, No. 3, 2017, pp 276-286

[15] dongdong Cheng, "Research on hierarchical clustering algorithm based on natural neighbors. Chongqing University", 2016.