Ilkogretim Online - Elementary Education Online, 2021; Vol 20 (Issue 3): pp. 3230-3237
http://ilkogretim-online.org
doi: 10.17051/ilkonline.2021.03.335

Customer Segmentation Using Machine Learning

[1]Garima Sharma, [2]Ankita Nainwal, [3]Bhaskar Pant, [4]Vikas Tripathi, [5] Mr Akash Chauhan

[1,2,3,4] Department of Computer Science and Engineering, Graphic Era Deemed to be University, Dehra Dun, Uttarakhand, India.

[5] Assistant Professor, Department of Computer Science and Engineering, Graphic Era Hill University, Dehradun.

ABSTRACT

In this business world where customers are increasing daily, it gets difficult to focus on each customer. Analyzing the customers past transactions would help the seller to satisfy customer demands and can easily attract new customers. Companies segments the customers to attain maximum profit and increase in sales. Companies need to understand this data and identify the similarity and dissimilarities in customer's needs, Customer segmentation uses unsupervised learning to segment customers into multiple distinct groups. In this paper we will use algorithms like k-means clustering and hierarchical clustering.

Key words: Customer Segmentation, K-Means, Hierarchical, Clustering Types

1. INTRODUCTION

Managing customers and identifying their likes and dislikes plays a vital role in market business. It has been observed that companies face losses because they are not able to identify the potential customers that will bring them profit. One of the many reasons of these losses are that the companies are using mass marketing tactics that are since whatever we are selling would be liked by everyone. These tactics are time-consuming, expensive, and even proved non-profitable. These strategies are ineffective since every consumer is unique, necessitating the use of some form of algorithm or practice to categories them based on the similarity of their preferences and direct our attention to those groups. To find hidden patterns in data and make future decisions that will be more effective, machine learning is utilized. The hazy idea of which section to target is made clear by the implementation of segmentation. Customer segmentation is the practice of classifying consumers into groups based on similar behavioral patterns and customers into distinct groups based on different behavioral patterns [1]. Suppose a brand focuses on all the customers that are visiting their website but some of them are just browsing their site without intending to buy anything, such people make it difficult for the seller to sell his product as he is targeting everyone. In these types of situation customer

segmentation is used [2].

Customer segmentation is a procedure in which customers are grouped on the basis that customers in the same group are more like each other than customers in other groups. Customer segmentation is based on clustering algorithm. The cluster segmentation is based on geographic, behavioral, demographic, lifestyle, preferences etc. Customer segmentation helps a company to concentrate on a group of people that were earlier buying similar products or might be interested in their product and now these companies can lure the potential customers by advertisements, discounts or updating them regularly about their new product launches. For example, a makeup brand would like to focus more on female customers aged around 18 to 45 who regularly use or buy their makeup products and the brand can give them discounts or even give them vouchers for their own customized products [3].

Any business will have a competitive advantage in offering focused customer services and creating specialized marketing campaigns for clients if it is able to comprehend the needs of each of its customers. This understanding is possible because to methodical consumer segmentation. Customers in each market sector share similar traits. Traditional market assessments, which are frequently ineffective, especially when there are many consumers, have been replaced by an automated method to customer segmentation thanks to the concepts of big data and machine learning [4]. As previously stated, we will segment customers using k-means and hierarchical clustering.

Datasets

We'll be working with the Kaggle-available Mall Customer dataset. This dataset includes some basic data about the clients, like their ID, age in integers, gender, and spending score (1–100) dollars. Figure 1 shows how dataset looks.

	V1	V2	V3	V4	V5
1	CustomerID	Genre	Age	Annual Income (k$)	Spending Score (1-100)
2	0001	Male	19	15	39
3	0002	Male	21	15	81
4	0003	Female	20	16	6
5	0004	Female	23	16	77
6	0005	Female	31	17	40
7	0006	Female	22	17	76
8	0007	Female	35	18	6
9	0008	Female	23	18	94

Figure 1: Mall Customer dataset

First, we pre-process the data, removing 'na' values and outliers and replacing them with mean of the values and using minorizing method that is to set a benchmark and replace all the outliers with bench mark. Since numeric columns in our dataset have different ranges of values, we will normalize our numeric value using

$$X = \frac{X - Xmin}{Xmax + Xmin}$$

	V1	V2	V3	V4	V5
1	CustomerID	Genre	Age	Annual Income (k$)	Spending Score (1-100)
2	1	Male	0.01923077	0	0.3877551
3	2	Male	0.05769231	0	0.81632653
4	3	Female	0.03846154	0.008492569	0.05102041
5	4	Female	0.09615385	0.008492569	0.7755102
6	5	Female	0.25	0.016985138	0.39795918
7	6	Female	0.07692308	0.016985138	0.76530612
8	7	Female	0.32692308	0.025477707	0.05102041
9	8	Female	0.09615385	0.025477707	0.94897959
10	9	Male	0.88461538	0.033970276	0.02040816

Figure 2: pre-processed dataset

2. RELATED WORKS

Online shopping is very popular, and many companies use it to market and sell their goods. using information, techniques, and procedures from studies on consumer segmentation to review customer segmentation. Server logs, cookies, and survey results were categorized as external data, whilst customer profiles and purchase history were categorized as internal data. In this study, different approaches were categorized as Simple, RFM, Target, and Unsupervised, and the process was generalized in identifying the business aim, gathering data, preparing data for analysis, analyzing variables, processing data, and evaluating performance[5].The study [6] establishes the most effective method for customer segmentation and extrapolates associated rules for this based on recency, frequency, and monetary (RFM) considerations as well as demographic factors. To improve the understanding of consumer segmentation, the effects of RFM and demographic variables were investigated in this study. The study's summary of findings and proof of its empirical implications show how to perfect extracted rules through effective model factors and variants for effective and efficient marketing.

Another study examines the two data mining (DM) methods of subgroup finding and clustering for consumer segmentation. The two DM algorithms that are used, K-Medoids and CN2-SD, create useful tools for better comprehending consumer preferences and tastes [7]. In a research, firm items are distributed as a product tree, with the internal nodes (apart from the root node) indicating various product categories and the leaf nodes representing goods to sale. Based on this tree, a proposal for a "personalised product tree," also known as the purchase tree, is proposed to reflect a customer's transaction data. The client transaction data set can be compressed using the resulting collection of buy trees. To swiftly cluster purchase trees, PurTreeClust, a partitional clustering technique, is employed. To efficiently compute the distance between two purchase trees, a novel distance measure is proposed. By evaluating the buy trees as probable candidate representative trees with a novel distinct density, the study first selects the top k customers as representations of the top k customer groups [8]. Electric scooters are

becoming more and more well-liked, which has caused the market to expand quickly on a global basis. Based on their budgetary restrictions and personal preferences, clients can use a study to help them choose the cheapest scooter from a particular market category. A dataset containing 42 scooter specs was used in the study. The study performs market segmentation and outlines the structure of the Polish electric scooter market. The quotients of the coefficients in consecutive phases of combining into clusters were used to make an arbitrary judgement to distinguish between the two and four classes of scooters. As the performance of the electric scooter improves, so does the price, according to a comparison of clusters with the chosen price ranges [9].

3. Methodology

An exploratory data analysis method called clustering is employed to look at the underlying structure of the data. When objects are grouped together, it means that they have comparable traits [10]. Our study uses two clustering methods namely k-means clustering and hierarchical clustering for the segmentation of customers.

I. K-Means Clustering

K-Means clustering is the most simpler and popular clustering algorithm. It is an iterative algorithm that divides the data into k non-overlapping groups. K Means uses intra cluster distance to group points into a cluster. It assigns points into a cluster such that the sum of the arithmetic mean of all the points that belong to that cluster is minimum. For a set of observation X = {x1, x2. . .. xn} k-means clustering partition it into k clusters where one point belongs to only one cluster and each cluster is unique. Mathematically algorithm works as follows: -

1. We first select 'k' cluster centers Y = {y_1, y_2…y_k} on a random basis.
2. Use Euclidean distances to determine the separation between each data point and its centers.
3. The data point will now be assigned to the cluster center with the shortest distance to it.
4. We now re-calculate the cluster center using

$$Y_i = (1/k_i) \sum_{j=1}^{c_i} x_i$$

where 'c_i' represent the number of data points in the cluster i.

5. Up until there are no changes between the reallocated cluster centers, we will recalculate the distance between each data point and the centers.

Now the question is how to determine the number of cluster 'k'. We can use elbow method or silhouette width for identifying k. The elbow method plots the variation as a function of number of clusters. The number of 'k' is determined by picking the elbow of the curve. From the figure 3 we can see that there is a curve at 5 at x axis referring number of optimal clusters should be 5.

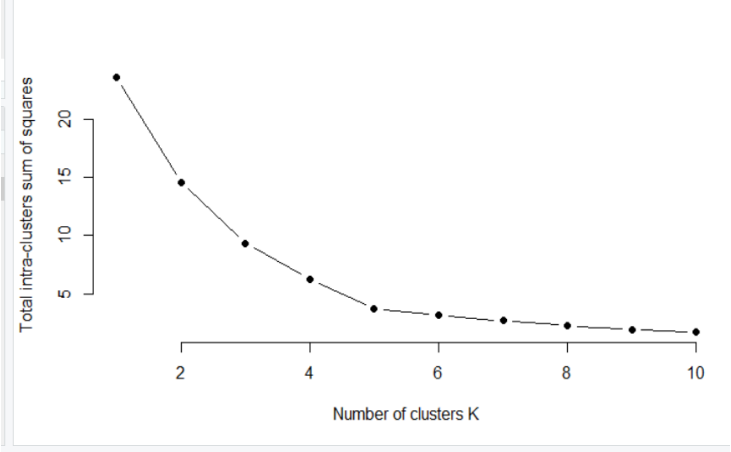

Figure 3: Elbow method demonstration

Silhouette width is a measure of how similar a data point is to its cluster. Its values range from -1 to +1. The likelihood of a data point fitting into its cluster increases as the width value increases. The convergence results are directly proportional to choice of our clustering K-value, which must be made beforehand. To overcome this issue, we focus on four K-value selection algorithms: TEM (The Elbow Method), GS (Gap Statistic), SC (Silhouette Coefficient), and the last one is Canopy [11].

II. Hierarchical Clustering

Hierarchical clustering or HCA is a method where hierarchy of clusters are made. Bottom-up clustering and top-down clustering are two methods for achieving hierarchical clustering. The clustering process in hierarchical approaches might begin with a single cluster of individual data points. To solve the clustering problem, hierarchical clustering algorithms create a binary tree-based data structure called a dendrogram [12]. By splitting the tree at various levels after the dendrogram has been created, it is possible to acquire various clustering solutions for the same dataset without having to perform the clustering method again . It starts with many clusters and combine them one by one at each iteration) and divisive (the top-down approach, it starts with one cluster and divide it into smaller cluster at each iteration). The algorithm works as follows

1. We assume each data point is a cluster, n clusters.
2. Take two closest clusters and combine them into one, n-1 clusters. Distance between the clusters can be calculated using Euclidean or Manhattan distance.

Euclidean distance = $\sum_i (a_i - b_i)^2$

Manhattan distance = $\sum_i |a_i - b_i|$

3. Repeat the above step until we have only one cluster left.

Hierarchical clustering uses matrix of distances and stores the clusters in the form of dendrogram. Dendrogram is a tree like structure, describes the relation between the different data points. Height of the blocks represent the distance between the clusters.

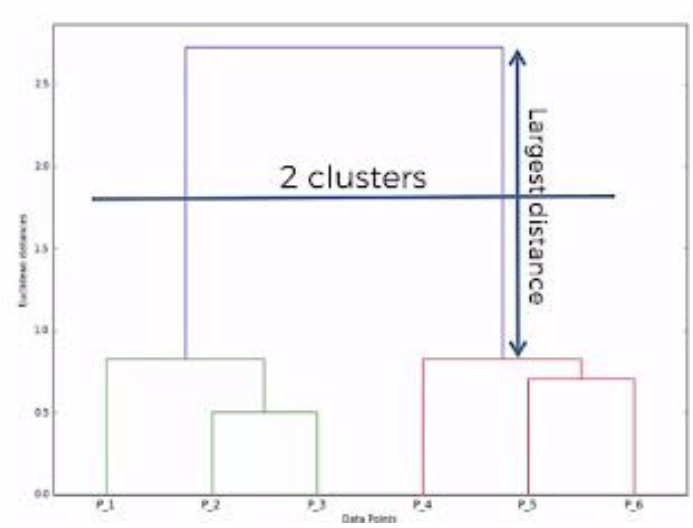

Figure 4: Dendrogram demonstration

Hierarchical clustering doesn't require a pre-defined value 'k' for clustering. It uses dendrogram to identify what number of clusters will give the best result. The number of vertical lines in dendrogram a horizontal threshold cuts is the number of clusters that can be formed. One of the standard methods for identifying number of clusters is to find highest vertical line that threshold cuts.

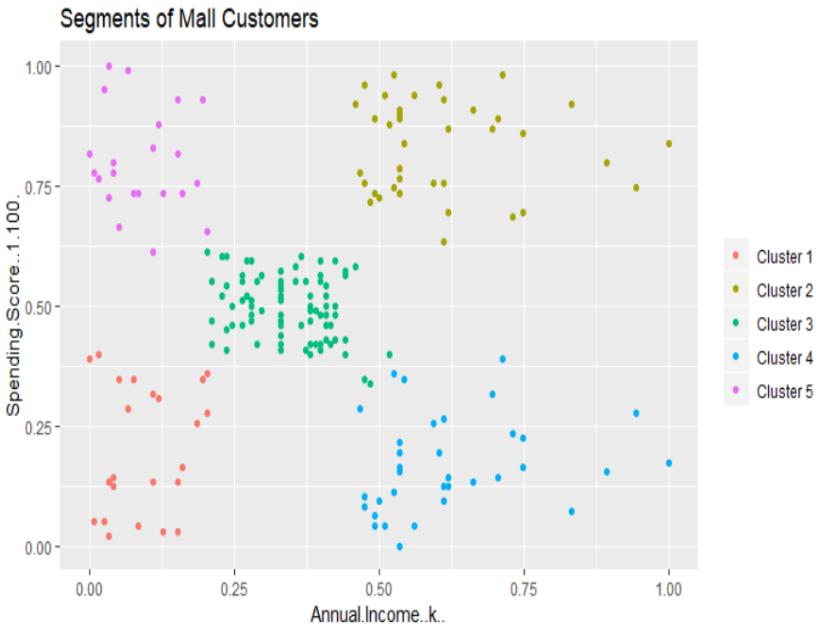

Figure 5: Clustering between annual income and spending score of customers.

According to Figure 5 customers who belongs to cluster 2 will give us the maximum profit as they are the ones whose spending amount and annual income both are more than the other.

4. RESULT AND ANALYSIS

K-means algorithm is used with large dataset as its time complexity is almost linear and even it takes less space as it only requires to store only the data points and centroids. K-means is fast and one of the simplest algorithms that gives best result when data points are distinct but it fails when data points are highly overlapped or are non-linear. Moreover, it requires a pre-defined value of 'k' the number of clusters whereas in hierarchical clustering there is no need of this value and number of clusters to be formed is easily determined by using a dendrogram. However hierarchical clustering doesn't work well with large dataset as its time complexity is O(n3) and requires O(n2) memory as it requires to store dendrogram.

5. CONCLUSIONS AND FUTURE SCOPE

The study mainly helps in understanding customer segmentation and how to apply machine learning methods to make system more effective and efficient. The study reviews various applications wherein the clustering methods are used for better customer identification and segmentation. The study helps in understanding k-means clustering and identifying its advantages using customer data retrieved from kaggle website. Similarly the study again uses the hierarchical clustering method for customer segmentation and helps in understanding the concepts and advantages of the method. With ecommerce development people our more interested in buying products online which makes it easier for companies to gather data, that can be done using cookies, server log, customer logins etc. But this data should be processed carefully as some customers would only be window surfing. Further we can apply supervised learning with decision tree and Quantile Membership which is based on RFM technique. Various advancing algorithms in clustering could also be used and compared to determine the best results.

REFERENCES

[1] T. Kansal, S. Bahuguna, V. Singh and T. Choudhury, "Customer Segmentation using K-means Clustering," 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS), 2018, pp. 135-139, doi: 10.1109/CTEMS.2018.8769171.

[2] Dolnicar, Sara. "Market segmentation for e-Tourism." Handbook of e-Tourism (2020): 1-15.

[3] "Intelligent and Fuzzy Techniques in Big Data Analytics and Decision Making"Proceedings of the INFUS 2019 Conference, Istanbul, Turkey, July 23-25, 2019

[4] Ezenkwu, Chinedu Pascal, Simeon Ozuomba, and Constance Kalu. "Application of K-Means algorithm for efficient customer segmentation: a strategy for targeted customer services." (2015).

[5] Sari, J.N., Nugroho, L.E., Ferdiana, R. and Santosa, P.I.," Review on customer segmentation technique on ecommerce" Advanced Science Letters, 22(10), pp.3018-3022.

[6] Sarvari, P.A., Ustundag, A. and Takci, H., "Performance evaluation of different

customer segmentation approaches based on RFM and demographics analysis" Kybernetes (2016).

[7] Brito, P.Q., Soares, C., Almeida, S., Monte, A. and Byvoet, M., " Customer segmentation in a large database of an online customized fashion business", Robotics and Computer-Integrated Manufacturing, 2015 36, pp.93-100.

[8] Chen, Xiaojun, et al. "Purtreeclust: A clustering algorithm for customer segmentation from massive customer transaction data." IEEE Transactions on Knowledge and Data Engineering 30.3 (2017):pp. 559-572.

[9] Kubiczek, J., & Hadasik, B. "Segmentation of the electric scooter market in Poland" Econometrics. Ekonometria. Advances in Applied Data Analytics, (2020), pp.50-65.

[10] Govender, P. and Sivakumar, V.," Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019)", Atmospheric pollution research ,2020, pp.40-56.

[11] Yuan, Chunhui, and Haitao Yang. "Research on K-value selection method of K-means clustering algorithm." J 2, no. 2 (2019): 226-235.

[12] Reddy, Chandan K., and Bhanukiran Vinzamuri. "A survey of partitional and hierarchical clustering algorithms." 2018, In Data clustering, pp. 87-110.