Study of Customer Spending Behavior using K-Means Algorithm

Aishwarya Raut, Student, BVIMIT, Navi Mumbai,India, csaishwaryaraut@gmail.com

Prof.Priya Chandran, BVIMIT, Navi Mumbai,India, priyaci2005@gmail.com

Abstract: Huge customer transaction data is available due to the wide use of information technology. Business organizations are implementing different tools to evaluate customer behavior and thereby promoting their products. In our proposed research study we propose the use of data mining algorithm to study customer spending behavior. We have used K-means algorithm and results are formed as clusters based on age and annual income of the customers.

Keywords — Centroid, Clustering, data mining, Information mining, K-means,

I. INTRODUCTION

Grouping examination strategy is one of the fundamental explanatory techniques in data mining. Clustering is an information mining method to make a group of similar behavior items into classes or clusters. Recognizing the quantity of clusters is a significant task for any grouping issue by and by yet it must be looked with numerous operational challenges. Data Clustering is an unsupervised learning issue. In this paper we aim to study the customer behavior using clustering method and used k-means algorithm for the same. For this study, the information is collected from shopping center. Optimal number of clusters are identified using elbow method.

II. LITERATURE REVIEW

Application of data mining algorithm in customer behavior analysis is an emerging trend. Several techniques are used for clustering implementation. In this paper, the initial segment of K-implies grouping calculation, the underlying centroids are resolved in order to create clusters with better precision. The second part utilizes a productive path for relegating information to clusters[6] [7]. The uniform conveyance of the information focuses is talked about that how this methodology lessen the time intricacy of the K-implies grouping calculation. Blume, Matthias, et al. studied the customer financial behavior and marketing responses and compared the results with clustering and nearest-neighbor and with the combinations of these methods [1]. Vivek K et al. used data from mobile phone to study social interaction features and hence to predict the spending behavior of adults [2]. Khajvand et al. studied the estimation of customer lifetime value based on RFM (Recency, Frequency, and Monetary) analysis of customer purchase behavior to predict the behavior of customers those who invest in share market [3]. Farajianet. Al used data from bank database to analyze customer's behavior to increase number of clients [4].Hung, Jui-Long, and Ke Zhang.uses data mining techniques to analyze various patterns of online learning behaviors, and to make predictions on learning outcomes [5]. L. Ertoz, M. Steinbach, and V. Kumar, studied about challenges when the clusters are of different shapes, sizes, and densities to overcome this they used novel clustering technique where they first finds the nearest neighbors of each data point and then redefines the similarity between two centroids.

The results forming similar type of data with less difference between sizes and density [6]. Ngai et. al studied and analyzed the use of different data mining techniques in customer relationship management [8]. Yadav et.al, studied the customer behavior for e-commerce and used k-mean algorithm for finding relationship between web mining and e-commerce [9]. Quilumba, Franklin L., et al. studied pattern of the actual power consumption by customer and results in forming clusters of customer with similar load consumption [10].

III. RESEARCH METHODOLOGY

We have used data collected from shopping center for the research study. The data includes client information which has client name, gender, annual salary, expenditure and age. In this paper, the unsupervised algorithm, K-means, is used to predict the customer spending behavior. K-means algorithm gives the cluster as output of the prediction. Similar behavior items are grouped as one cluster.

First we feed the data into system, preprocess and then we define the different type of clusters Such as sensible, careless, careful according to age and income. After processing customer data , they are allocated to different clusters that are defined according to their nature and difference between centroids and item. These clusters represent the customer behavior and accordingly we can Judge the behavior of customer this helps shopping center for their marketing. Once we define clusters we apply algorithm for identifying groups First we choose one centroid after choosing centroid the minimum distance between item and centroid are calculated and accordingly

 DOI : 10.35291/2454-9150.2019.0204

clusters are form This loops goes on till data gets similar once we found two data are similar System stops the loop and last output with clusters are form Which is than useful for identifying the customer behavior .The overall goal of this system is to extract information from large data set And transform it into understandable form for further use. Clustering is important so that we can get useful information about Customer. The goal of this task is to determine a particular attribute that is annual income based on another attribute that is Age. In this system we feed large and raw data than it goes through algorithm i.e clustering process and finally we have clustered data that is understandable and useful data. Data in same clusters are similar with each other but are dissimilar to other clusters. This data is than represented in elbow graph This system is useful for small data sets for identification of data.

$$j(v) = \sum_{i=1}^{o} \sum_{j=1}^{oi} (\|x(i) - v(j)\|)^2$$

where,

||x(i)-v(j)|| is the distance between x(i) and v(j)

C(j) is the number of data points in it cluster.

'c' is the number of cluster center.

3.1 Elbow Method

The Elbow technique is a strategy for understanding and validation of consistency within cluster analysis designed to help finding the appropriate number of clusters in a dataset. This technique takes looks at the percentage of variance explained as a element of the number of clusters. One chooses to pick various clusters so that including another group doesn't give much better modeling of the information.

3.2 Euclidean distance method

Euclidean distance is stand-alone tool, use in application to find the nearest object to form similar type of cluster which becomes easy to identify data. In mathematics, Euclidean distance means the straight line that means difference between two points that will be measured using dimension, and is given by the Pythagoras formula.

Euclidean distance is

$$\sqrt{(a_{i1} + a_{j1}) + (a_{i2} + a_{j2}) + \cdots + (a_{in} + a_{jn})}$$

Where i = $(a_{i1}, a_{i2} \ldots\ldots, a_{in})$

j=$(a_{j1}, a_{j2} \ldots\ldots, a_{jn})$

n is dimension object.

IV RESULTS AND DISCUSSION

We have used python to implement the proposed system. The input data set is preprocessed first and then K means algorithm is used to obtain results. The results were presented as clusters. Identifying the optimal number of clusters is a big challenge in implementing clustering method. We have used elbow method to identify the optimal number of clusters.

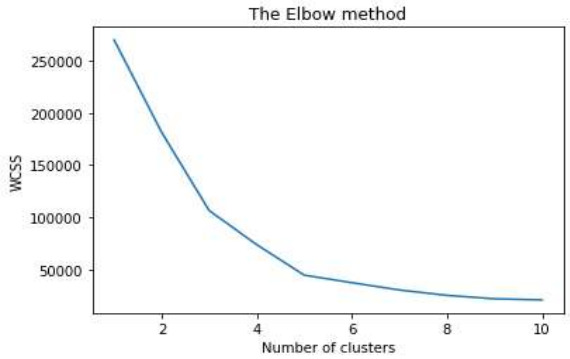

Figure 4.1 Elbow Graph

Figure 4.1 shows elbow method .Number of cluster is represented as X-axis and WCSS (within-cluster sums of squares) is represented as Y-axis. As number of clusters increase, WCSS decreases. WCSS measures the squared average distance of all the points within a cluster to the cluster centroid. To calculate WCSS, Euclidean distance between a given point and the centroid to which it is assigned is estimated. This system takes five clusters which have wcss value as 50000.

Clusters are form according to groups. This all clusters have cendroid. In this we have five clusters. These five clusters are made according to expenditure in mall with respect to age groups and the annual income. This system helps us to differentiate between expenditure of different according to age and annual income. This helps us to know peoples perspective. The proposed system form five clusters and are given below:

1.Careful

2.Standards

3.Sensible

4.Careless

5.Target

When we take input information in our framework it initially pick introductory centroids and then begin computing separation between variables. It shapes the bunch of those information which have less separation. At the point when separate between two factors gets rehashed

 DOI : 10.35291/2454-9150.2019.0204

around then the framework gets stop and last clusters are created.

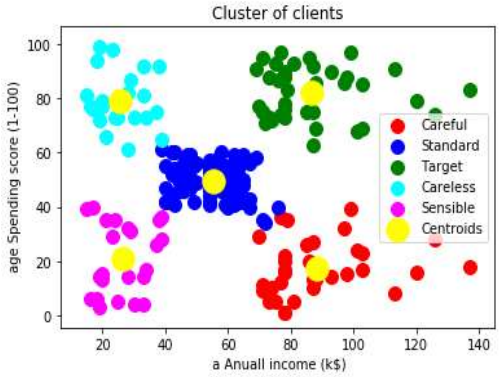

Figure 4.2 Clusters

Figure 4.2 shows the clusters are formed in different group. There are five groups of clusters on bases of client's age and annual income.

Table 4.1 Cluster data

Age\Income(k)	0-20	20-40	40-60	60-80	80-100	100-120	120-140
0-20	Sensible	-	Careful	Careful	Careful	Careful	Careful
20-40	Sensible	-	Careful/ Standard	Careful	Careful	Careful	Careful
40-60		Standard	Standard	Standard	-	-	-
60-80	Careless	Careless	-	Target	Target	Target	Target
80-100	Careless	Careless	Target	Target	Target	Target	Target

Table 4.1 gives overview of people's expenditure on bases of their age along with the clusters. As, customers having 0-20,000 annual income between age group of 0-20 spends sensibly. Same as age group of 0-20 having annual income about 20,000 to 40,000 spends money carefully. Age group of 40-60 years people having annual income about 20,000 to 40,000 spends money in a standard way.

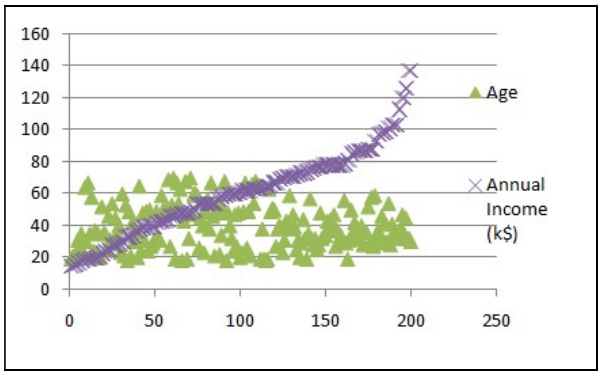

Figure 4.4 Correlation between centroids

Correlation between centroids of the cluster is shown in figure 4.4. Scatter diagram is used for calculating correlation between centroid. Here age and annual income is our two centroids. Clusters are formed based on annual income.

V. Conclusion

Data mining algorithms are commonly used to analyze data. We have used K means algorithm to identify customer spending patterns. The main aim of this paper is to extract information from large data and form clusters for simplification. The proposed method is implemented in python. The results of the study are shown as clusters on the basis of their age group and annual income. The results support business organizations in promoting their products and will result in better sale.

Reference

[1] Blume, Matthias, et al. "Predictive modeling of consumer financial behavior using supervised segmentation and nearest-neighbor matching." U.S. Patent No. 6,839,682. 4 Jan. 2005.

[2] Singh, Vivek K., et al. "Predicting spending behavior using socio-mobile features." 2013 International Conference on Social Computing. IEEE, 2013.

[3] Khajvand, Mahboubeh, et al. "Estimating customer lifetime value based on RFM analysis of customer purchase behavior: Case study." Procedia Computer Science 3 (2011): 57-63.

[4] Farajian, Mohammad Ali, and Shahriar Mohammadi. "Mining the banking customer behavior using clustering and association rules methods." (2010): 239-245.

[5] Hung, Jui-Long, and Ke Zhang. "Revealing online learning behaviors and activity patterns and making predictions with data mining techniques in online teaching." MERLOT Journal of Online Learning and Teaching (2008).

[6] L. Ertoz, M. Steinbach, and V. Kumar, "Finding Clusters of Different Sizes, Shapes, and Densities in Noisy," SIAM International Conference on Data Mining, Feburary 20, 2003.

[7]. L. Kaufman and P. J. Rousseeuw, "Finding Groups in Data: An Introduction to Cluster Analysis," Wiley Series in Probability and Statistics. John Wiley and Sons, New York, Novemeber 1990.

[8] Ngai, Eric WT, Li Xiu, and Dorothy CK Chau. "Application of data mining techniques in customer relationship management: A literature review and classification." Expert systems with applications 36.2 (2009): 2592-2602.

[9] Yadav, Mahendra Pratap, Mhd Feeroz, and Vinod Kumar Yadav. "Mining the customer behavior using web usage mining in e-commerce." 2012 Third International Conference on Computing, Communication and Networking Technologies (ICCCNT'12). IEEE, 2012.

[10] Quilumba, Franklin L., et al. "Using smart meter data to improve the accuracy of intraday load forecasting considering customer behavior similarities." IEEE Transactions on Smart Grid 6.2 (2014): 911-918.

 DOI : 10.35291/2454-9150.2019.0204