Customer Segmentation using K-means Clustering

[1]Tushar Kansal, [2]Suraj Bahuguna , [3]Vishal Singh, [4]Tanupriya Choudhury

[1]kansaltushar18@gmail.com;[2]bahugunasooraj@gmail.com;[3]vishalsinghpari.0816@gmail.com,[4]tanupriya1986@gmail.com

[1,2,3,4] University of Petroleum & Energy Studies (UPES),Dept. of Informatics,School of Computer Science,Dehradun

Abstract-:

The zeitgeist of modern era is innovation, where everyone is embroiled into competition to be better than others. Today's business run on the basis of such innovation having ability to enthral the customers with the products, but with such a large raft of products leave the customers confounded, what to buy and what to not and also the companies are nonplussed about what section of customers to target to sell their products. This is where machine learning comes into play, various algorithms are applied for unravelling the hidden patterns in the data for better decision making for the future. This elude concept of which segment to target is made unequivocal by applying segmentation. The process of segmenting the customers with similar behaviours into the same segment and with different patterns into different segments is called customer segmentation. In this paper, 3 different clustering algorithms (k-Means, Agglomerative, and Meanshift) are been implemented to segment the customers and finally compare the results of clusters obtained from the algorithms. A python program has been developed and the program is been trained by applying standard scaler onto a dataset having two features of 200 training sample taken from local retail shop. Both the features are the mean of the amount of shopping by customers and average of the customer's visit into the shop annually. By applying clustering, 5 segments of cluster have been formed labelled as Careless, Careful, Standard,Target and Sensible customers. However, two new clusters emerged on applying mean shift clustering labelled as High buyers and frequent visitors and High buyers and occasional visitors.

Keywords: Customer Segmentation, k-Means algorithm, Mean shift algorithm, Agglomerative algorithm, Machine learning, Python.

Introduction:

As more and more business being coming up every day, it has become significantly important for the old businesses to apply marketing strategies to stay in the market as the competition has been cut to throat. Change or die have become the simple rule of marketing in today's world. As the customer base is increasing day by day it has become challenging for the companies to cater to the needs of each and every customer, this is where Data mining serves a very important role to unravel hidden patterns stored in the company's database. Customer segmentation is one of the application of data mining which helps to segment the customers with similar patterns into similar clusters hence, making easier for the business to handle the large customer base. This segmentation can directly or indirectly influence the marketing strategy as it opens many new paths to discover like for which segment the product will be good, customising the marketing plans according to the each segment, providing discounts for a specific segment, and decipher the customer and object relationship which has been previously unknown to the company. Customer segmentation allows companies to visualise what actually the customers are buying which will prompt the companies to better serve their customers resulting in customer satisfaction, it also allows the companies to find who their target customers are and improvise their marketing tactics to generate more revenues from them.

Clustering has been proven effective to implement customer segmentation. Clustering comes under unsupervised learning, having ability to find clusters over unlabelled dataset. There are a number of clustering algorithm over which like k-means, hierarchical clustering, DBSCAN clustering etc. In this paper, three different clustering algorithms have been implemented over a dataset with two features with 200 records.

K-means Clustering:

It is the simplest algorithm of clustering based on partitioning principle. The algorithm is sensitive to the initialization of the centroids position, the number of K (centroids) is calculated by elbow method (discussed in later section), after calculation of K centroids by the terms of Euclidean distance data points are assigned to the closest centroid forming the cluster, after the cluster formation the barycentre's are once again calculated by the means of the cluster and this process is repeated until there is no change in centroid position.[10][11]

Agglomerative Clustering-:

Agglomerative Clustering is based on forming a hierarchy represented by dendrograms (discussed in later section). Dendrogram acts as memory for the algorithm to tell about how the clusters are being

978-1-5386-7709-4/18/$31.00 ©2018 IEEE

formed. The clustering starts with forming N clusters for N data points and then merging along the closest data points together in each step such that the current step contains one cluster less than the previous one.

Mean shift Clustering-:

This clustering algorithm is a non-parametric iterative algorithm functions by assuming the all the data points in the feature space as empirical probability density function. The algorithm clusters each data point by allowing data point converge to a region of local maxima which is achieved by fixing a window around each data point finding the mean and then shifting the window to the mean and repeat the steps until all the data point converges forming the clusters.

Elbow Method-:

Elbow method is used for finding optimal value of K for K-means clustering algorithm. This method works by finding the SSE of each data point with its nearest centroid with different values of K. As value of K increases the SSE will decrease and at a particular value of K where there is most decline in the SSE is the elbow, the point at which we should stop dividing data further.

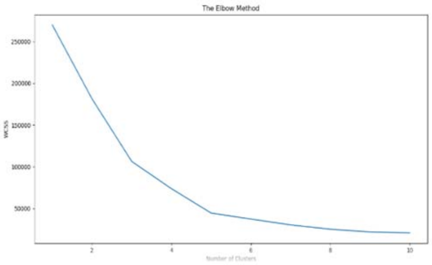

Fig. 1: Graph for Values of K VS WCSS(Within Cluster Sum of)

From the above graph it is clearly be seen that from number of cluster = 4 to number of cluster = 5 there has been substantial decrease hence, we choose the K value for our dataset as 5.

Dendrogram-:

Dendrogram is the hierarchical representation of object, it is used to determine the output of the hierarchical clustering. The way Dendrogram is interpreted is by checking the height of each clade (horizontal line), the lower the height the more associated data points are and greater the height more less associated data points.

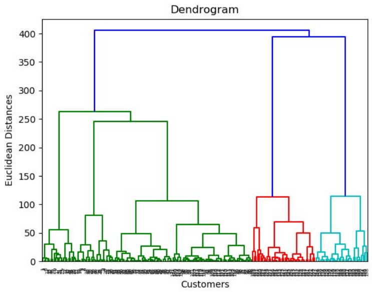

Fig2: Dendrogram Structure of the dataset.

Fig (2) shows Dendrogram of our dataset built using [9].The figure displays how the clusters are being formed to eventually converge to a single cluster. Dendrogram is being used used for finding the numberof clusters that is optimal to apply for Agglomerative clustering, in the Fig (2) we look for the longest vertical line which is not being cut by any of the clades (horizontal line) extended virtually over complete width of the graph, the second last clade of green colour have it's right leg bearing the longest vertical line which is not been cut by any clade. Now, by drawing a hypothetical horizontal line cutting through the longest vertical line, we get the horizontal line cutting total 5 vertical lines providing us the optimal number of clusters for our dataset.

Bandwidth:

Bandwidth can be considered as the radius of the circle (kernel) describing how much the data points should be in the cluster. It is the only input requisite for the Mean shift algorithm, calculated by the help of K Nearest Neighbours. Mean shift algorithm is very much sensitive to theinitialization of bandwidth, a small value can slow down the converging process while a large value can speed up convergence.

Methodology:

Data Collection:

The dataset has been taken from a local retail shop consisting of two features, average number of visits to the shop and average amount of shopping done on yearly basis.

Feature Scaling:

The data has been scaled using Standard Scaler [9], by applying standard scaler the data gets centred around 0 with standard deviation of 1.

$$\frac{x - mean(X)}{stdev(X)}$$

x = entry in a feature set xi ϵ X

mean (X) = mean of feature set X

stdev (X) = standard deviation of X

K means Clustering:

Choosing the optimal number of clusters:

Elbow method is applied to calculate value of K for the dataset.

Step-1: Run the algorithm for various values of k i.e making the k vary from 1 to 10.

Step-2: Calculate the within cluster squared error.

Step-3: Plot the calculated error, where a bent elbow like structure will form, will give the optimal value of clusters.

SSE is calculated by -:

$$\sum_{i=1}^{k} \sum_{Xj \, \epsilon \, Si} \|Xj - \mu i\|^2$$

X_j = data point in S_i cluster

μ_i = centroid of the cluster

Algorithm:

Step-1: Initialize the K (= 5) clusters.

Step-2: Assign the data point that is closest to any particular cluster.

Step-3: Recalculate the centroid position based on the mean of the cluster formed

Step-4: Repeat step 2 and 3 until the centroid position remains unchanged in the previous and current iteration.

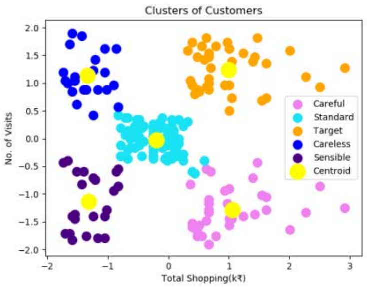

Fig. 3: Clusters formed by K means

The figure above shows the 5 final clusters where the cluster in orange colour gives the target customer.

Agglomerative Clustering:

Choosing the optimal number of clusters:

Cluster value for this algorithm have been calculated by the Dendrogram as described in Dendrogram section which also gave the value of K = 5.

Algorithm:

1) Each data point is taken as to be a cluster.
2) Merge the two closest cluster.
3) Step 2 needs to be repeated until all the data points are merged together to form a single cluster. However, as we have defined the value of K as 5, the algorithm will stop when all the data points are part of any of the 5 clusters.

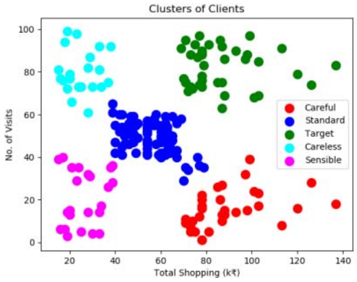

Fig.4: Clusters formed by Agglomerative Clustering

Since, the number of clusters for K-means and Agglomerative are equivalent they gave same pattern for final clusters. However, if you look closely a point on the top left corner in Standard cluster has changed its cluster and similar case is with 2-3 points on lower right corner in Standard cluster.

Mean Shift Clustering:

This non-parametric clustering method is being applied to see some different pattern in a dataset as K-means and Agglomerative gave almost the same result. There is no need of choosing the number of clusters. However, it needs one input parameter, bandwidth (radius) which is calculated using K-nearest neighbour algorithm. This algorithm follows an iterative approach where a point of local maxima is found around each data point defined by probability density function, and iterates until when all the data point converges up the hill (created by PDF), also known as 'hill climbing algorithm'.

PDF can be estimated by-:

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

h = bandwidth

K = kernel

Some examples of Kernel:

1) *Rectangular:*

$$f(x) = \begin{cases} 1, & a \leq x \leq b \\ 0, & else \end{cases}$$

2) *Gaussian:*

$$f(x) = e^{\frac{-x^2}{2\sigma^2}}$$

Algorithm:

1) A window is associated around each data point created by PDF
2) Mean around the window is calculated
3) Window is moved towards the newly calculated mean.
4) Step 2 and 3 are repeated until when all the data points converge to a local maxima resulting in clusters.

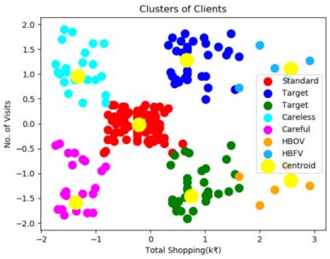

Fig. 5: Clusters formed by Mean Shift

The final outcome gave two new clusters labelled as High Buyer Frequent Visitors and High Buyer Occasional Visitors, these two new clusters can help the retail shop to treat the customers lying their segment as their VIP customers providing them accolade on purchase of products or giving them extra discounts.

Literature Review:

Customer Segmentation:

Jayant et al. [1] states that in customer segmentation, the customers are divided into different groups where customers of the same group are similar to each other in terms of marketing. Customers are divided into different clusters based on various attributes such as age ,interests, age, spending habits etc.

Sulekha et al. [7]provides the four popular bases for segmentation

1) Geographic Segmentation: segmentation on the basis geographic region, population density or climate.
2) Demographic Segmentation: market segment on the basis of age, size and family type, etc.
3) Psychographic Segmentation: segmentation based on customer's life style variables like interests, opinions, attitudes etc.
4) Behavioural Segmentation**:** segmentation is based on actual customer behaviourtowards products like brand loyalty, user status, readiness to buy etc.

Customer segmentation is based on based on the strategy called divide and conquer by utilising the advantage of segmentation the marketers can gain advantage over a particular segment and slowly can prevail over other marketers. Using market segmentation the marketers can focus more on customer relationship management which was not earlier possible with existing mass marketing tactics.

Clustering-:

Clustering is the process of grouping the information in the dataset based on some similarities. There are a number of algorithms which can be chosen to be applied on a dataset based on the situation provided. However, no universal clustering algorithm exists that's why it becomes important to opt for appropriate clustering techniques. Vaishali et al. [8]. In this paper, we have implemented three clustering algorithms using python scikit learn library [9].

K-Means Clustering:

K- means algorithm is one of the most popular partitioning clustering algorithm. This clustering algorithm depends on the centroid where each data point is placed on one of the K non overlapping clusters which are selected before running of the algorithm, Chinedu et al. [2]. The clusters formed corresponds to the hidden pattern in the data which gives the required information to help in decision making process.

Agglomerative Clustering:

This clustering comes under hierarchical clusters are formed based on some hierarchy. Hierarchical clustering is It is based on the concept that objects that are closer are more related to each other in comparison of the objects that are far from each other., T.Nelson et al. [3]. The main challenge of Hierarchical method is that once it undergoes split or merge operation it can never be undone. This challenge is profitable as it leads to smaller computation costs by not worrying about a combinatorial number of different choices. Yogita et al. [4]. There are two strategy in hierarchical clustering, first is top-down strategy also known as divisive clustering and second is bottom – up strategy also known as agglomerative clustering.

Agglomerative clustering process is generally slower than divisive clustering but allows more flexibility because it permits the user to supply any arbitrary similarity function defining what constitutes a similar cluster pair to merge together, Omar et al. [5].

Mean shift Clustering:

Sulekha et al. [7] defines this algorithm as a gradient ascent technique. In mean shift the local maxima of a density function is found from the given data samples that are discrete. It works with a search window that is positioned over a section of the distribution. The mean shift technique is used for real data analysis which is an application dependent tool and initially shape of data cluster is not assumed. This algorithm has wide applications in object detection, image segmentation.

Results:

We have taken two internal clustering measure, silhouette score and Calinski-Harabasz index.

Silhouette Score:

It is a way of measuring how well the data point has been clustered into the correct cluster.

First Step-:

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

Second Step:

b = Average distance between the data point and the closest cluster data points.

Third Step:

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

For the data point to be well grounded in its cluster 'b' needs to be large and 'a' needs to be small so that difference between the two is as large as possible.

'max (b,a)' is added to normalized the silhouette score. Higher the score better the data point belongs to that cluster.

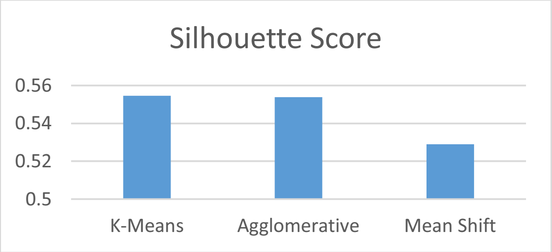

Fig. 6: Comparison of Silhouette Score

Figure above displays the silhouette score for the three algorithms applied in this paper, the graph shows there is not much significant difference in K-means and Agglomerative clustering. Hence, these two algorithms were able to cluster our data well than Mean shift algorithm as displayed by the low value of silhouette score.

Conclusion:

As our dataset was unlabelled, in this paper we have opted for internal clustering validation rather than external clustering validation, which depends on some external data like labels. Internal cluster validation can be used for choosing clustering algorithm which best suits the dataset and can correctly cluster data into its opposite cluster.

References:

[1] Jayant Tikmani, Sudhanshu Tiwari, Sujata Khedkar "Telecom customer segmentation based on cluster analysisAn Approach to Customer Classification using k-means", IJIRCCE,Year: 2015.

[2] Chinedu Pascal Ezenkwu, Simeon Ozuomba, Constance kalu Electrical/Electronics & Computer Engineering Department, University of Uyo, Uyo, Akwa Ibom State, Nigeria "Application of K-Means Algorithm for Efficient Customer Segmentation: A Strategy for Targeted Customer Services", IJARAI,Year: 2015.

[3] T.NelsonGnanarajDr.K.Ramesh Kumar N.Monica"Survey on mining clusters using new k-mean algorithm from structured and unstructured data", IJACST,Year: 2014.

[4] Yogita Rani and Dr. Harish Rohil"A Study of Hierarchical Clustering Algorithm", IJICT,Year: 2013.

[5] Omar Kettani, FaycalRamdani, BenaissaTadili"An Agglomerative Clustering Method for Large Data Sets", IJCA,Year: 2014.

[6] Snekha, ChetnaSachdeva, Rajesh Birok"Real Time Object Tracking Using Different Mean Shift Techniques–a Review", IJSCE,Year: 2013.

[7] SulekhaGoyat"The basis of market segmentation: a critical review of literature", EJBM,Year: 2011.

[8] Vaishali R. Patel and Rupa G. Mehta "Impact of Outlier Removal and Normalization Approach in Modified k-Means Clustering Algorithm", IJCSI,Year: 2011.

[9]Scikit-learn: https://scikit-learn.org

[10] Tanupriya Choudhury, Vivek Kumar, Darshika Nigam, Intelligent Classification & Clustering Of Lung & Oral Cancer through Decision Tree & Genetic Algorithm, International Journal of Advanced Research in Computer Science and Software Engineering,2015

[11] Tanupriya Choudhury, Vivek Kumar, Darshika Nigam, An Innovative and Automatic Lung and Oral Cancer Classification Using Soft Computing Techniques, International Journal of Computer Science & Mobile Computing,2015