e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:01/January-2022 Impact Factor- 6.752 www.irjmets.com

CUSTOMER SEGMENTATION ALGORITHMS

Prince Nathan S*[1], Onkar Saudagar*[2], Rutika Shinde*[3]

*[1]PG Student, Department Of Data Science And Analytics, National Institute Of Electronics And Information Technology, Chennai, India.

*[2]Student, Department Of Information Technology, Pune Institute Of Computer Technology, Pune, Maharashtra, India.

*[3]Student, Department Of Information Technology, Sinhgad Institute Of Technology And Science, Pune, Maharashtra, India.

ABSTRACT

We live in a world where massive amounts of data are collected on a regular basis. The necessity to analyze such data is critical. In this modern era of innovation, when everyone is competing to be better than everyone else, the company plan must be tailored to the current circumstances. Because there are so many prospective clients who are unsure of what to purchase and what not to buy, today's business is based on new concepts. The businesses themselves are unable to diagnose their target potential clients. This is where machine learning comes in; various algorithms are used to detect hidden patterns in the data in order to make better decisions. The customer segmentation method uses the clustering approach to determine which consumer segment to target. The clustering approach employed in this work is the K-means algorithm, which is a partitioning strategy for segmenting clients based on comparable criteria. The elbow approach is used to find the best clusters.

Keywords: Machine Learning, K-Means, Elbow Method.

I. INTRODUCTION

The widespread use of data mining techniques in extracting meaningful and strategic information from an organization's database has resulted from the increased competition among businesses over the years, and the large historical data that is available has resulted from the widespread use of data mining techniques in extracting meaningful and strategic information from an organization's database. Data mining is a process in which technologies are employed to extract data patterns and display them in a human-readable manner that may be utilized for decision-making. Clustering algorithms use data tuples as objects, according to. They divide the data items into groups or clusters, with objects inside a cluster being similar to one another and things in other clusters being distinct. Client segmentation is the process of dividing a customer base into various groups known as customer segments, each of which includes consumers with comparable characteristics. The segmentation is based on similarities in a variety of ways important to marketing, such as gender, age, hobbies, and other purchasing behaviours. The ability to modify market programmes so that they are suitable for each of the customer segments, support in business decisions; identification of products associated with each customer segment and to manage the demand and supply of that product; identifying and targeting the potential customer base, and predicting customer defection, and providing directions in finding solutions are all important aspects of customer segmentation. The goal of this study is to use a data mining strategy to discover customer segments utilising the K-means clustering algorithm as a partitioning tool. The elbow approach is used to find the best clusters.

II. LITERATURE REVIEW

Customer Segmentation

Because of the intense rivalry in the business sector, businesses have had to improve their profitability and business throughout time by meeting client requests and attracting new customers based on their wants. Customer identification and meeting each customer's needs is a difficult and time-consuming process. This is because clients differ in terms of their wants, tastes, and preferences, among other things. Customer segmentation, as opposed to a "one-size-fits-all" strategy, separates consumers into groups with similar features or behavioural traits. Customer segmentation, according to [5,] is a strategy for splitting the market into homogeneous segments. The information employed in the customer segmentation approach, which separates consumers into groups, is based on a variety of elements, including data geographical circumstances, economic conditions, demographical conditions, and behavioural tendencies. The customer segmentation strategy enables a company to make better use of its marketing resources.

Clustering and K-Means Algorithm

Clustering techniques create clusters that are similar within themselves depending on specific features. The distance between the items in space is used to define similarity.

One of the most common centroid-based algorithms is the K-means method. Assume D is a data collection with n items in space. Partitioning techniques divide objects in D into k clusters, C1,..., Ck, i.e., Ci D and Ci Cj = for (1 I j k). The centroid of a cluster, Ci, is used to represent that cluster in a centroid-based partitioning approach. The centroid of a cluster is its centre point in terms of concept. Dist(p,ci) measures the difference between an item p Ci and ci, the cluster's representative, where dist(x,y) is the Euclidean distance between two points x and y.

Algorithm:

The k-means algorithm for partitioning, where each cluster’s center is represented by the mean value of the objects in the cluster.

Input: k: the number of clusters, D: a data set containing n objects.

Output: A set of k clusters.

Method:

(1) arbitrarily choose k objects from D as theinitial cluster centres;

(2) repeat

(3) (re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;

(4) update the cluster means, that is, calculatethe mean value of the objects for each cluster;

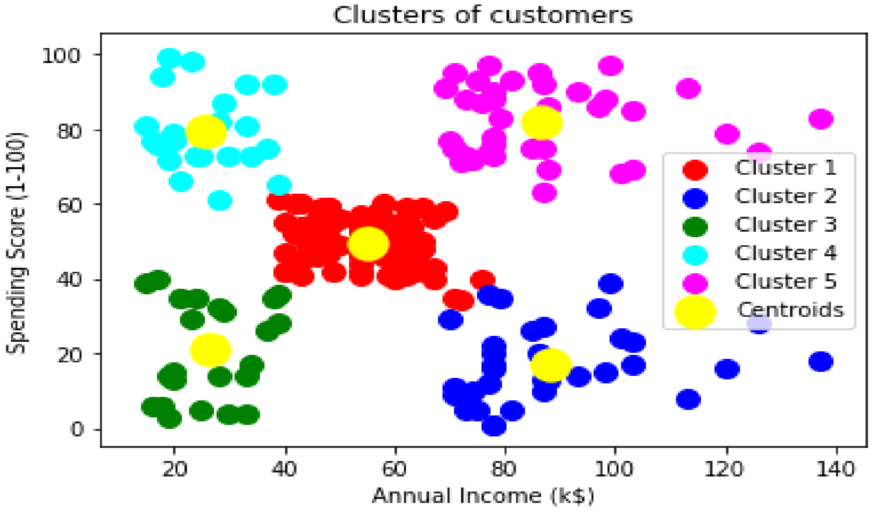
(5) until no change.

III. METHODOLOGY

The data set for clustering and the K-means technique was obtained from a shopping centre store. The data collection comprises five properties and 200 tuples, which represent the information of 200 customers. CustomerId, gender, age, yearly income (k$), and expenditure score on a scale of 1-10 are among the attributes in the data collection (1-100). budgets, get a competitive advantage over their competitors by exhibiting a greater understanding of the customer's wants. It also aids a company in improving marketing efficiency, recognising new market opportunities, developing a stronger brand strategy, and assessing client retention.

Visualize the clusters

```
from sklearn.cluster import KMeans
kmeans = KMeans(n_clusters = 5, init = 'k-means++', random_state = 42)
y_kmeans = kmeans.fit_predict(X)
# Visualising the clusters
plt.scatter(X[y_kmeans == 0, 0], X[y_kmeans == 0, 1], s = 100, c = 'red', label = 'Cluster 1')
plt.scatter(X[y_kmeans == 1, 0], X[y_kmeans == 1, 1], s = 100, c = 'blue', label = 'Cluster 2')
plt.scatter(X[y_kmeans == 2, 0], X[y_kmeans == 2, 1], s = 100, c = 'green', label = 'Cluster 3')
plt.scatter(X[y_kmeans == 3, 0], X[y_kmeans == 3, 1], s = 100, c = 'cyan', label = 'Cluster 4')
plt.scatter(X[y_kmeans == 4, 0], X[y_kmeans == 4, 1], s = 100, c = 'magenta', label = 'Cluster 5')
plt.scatter(kmeans.cluster_centers_[:, 0], kmeans.cluster_centers_[:, 1], s = 300, c = 'yellow', label = 'Centroids')
plt.title('Clusters of customers')
plt.xlabel('Annual Income (k$)')
plt.ylabel('Spending Score (1-100)')
plt.legend()
plt.show()
```

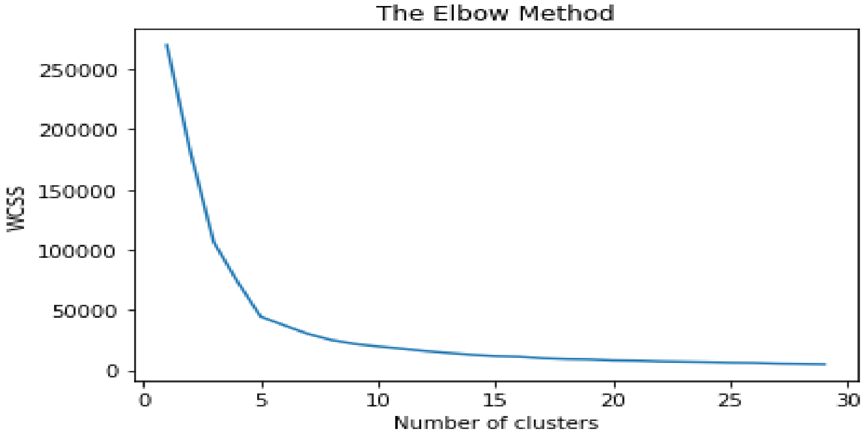


Elbow Method:

The elbow technique is based on the finding that increasing the number of clusters can assist lower each cluster's total within-cluster variation. This is due to the fact that having more clusters enables for the capturing of smaller groupings of data items that are more comparable to one another.

To find the best clusters, we first employ the clustering method with different values of k. This is accomplished by varying k from one to ten clusters. The total intra-cluster sum of squares is then calculated. Then, based on the number of clusters, we depict the intra-cluster sum of squares. In our model, the graphic shows the approximate number of clusters necessary. The optimal clusters may be discovered by looking at the graph where it bends.

```
# ELBOW METHOD TO DECIDE THE NO. OF CLUSTERS
from sklearn.cluster import KMeans
wcss = [ ]
for i in range(1, 30):
  kmeans = KMeans(n_clusters = i, init = 'k-means++', random_state = 42)
  kmeans.fit(X)
  wcss.append(kmeans.inertia_)
plt.plot(range(1, 30), wcss)
plt.title('The Elbow Method')
plt.xlabel('Number of clusters')
plt.ylabel('WCSS')
plt.show()
```

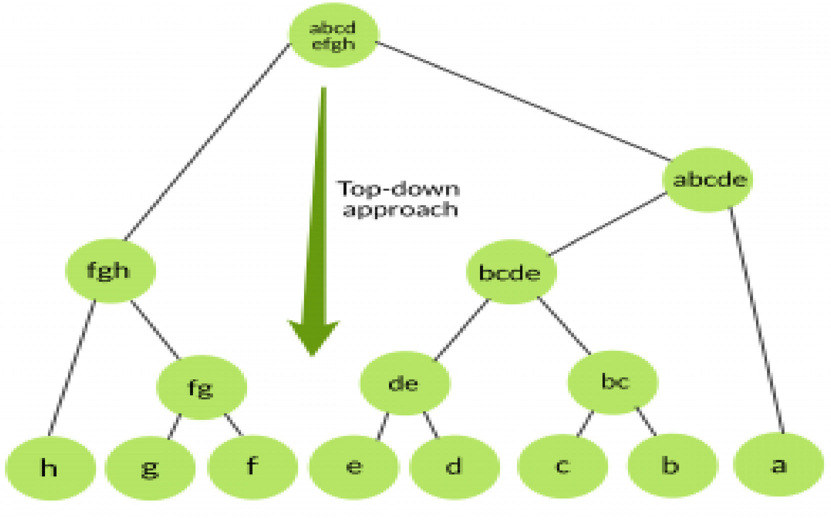


Here the optimal number of clusters are 5.

Agglomerative Clustering:

Also known as hierarchical agglomerative clustering or bottom-up technique (HAC). Flat clustering yields an unstructured set of clusters; this structure is more revealing. We don't have to define the number of clusters in advance with this clustering procedure. Bottom-up algorithms start by treating each piece of data as a singleton cluster, then agglomerate pairs of clusters until all of them are merged into a single cluster that includes all of the data.

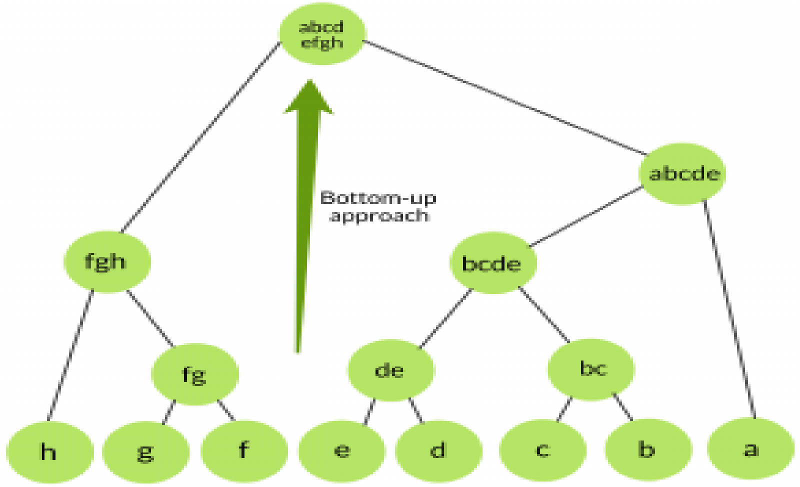
```
from sklearn.cluster import AgglomerativeClustering
hc= AgglomerativeClustering(n_clusters=5, affinity='euclidean', linkage='ward')
y_pred= hc.fit_predict(x)
```



Hierarchical divisive clustering

Divisive Clustering

A top-down strategy is also known as a top-down approach. This approach also eliminates the need to define the number of clusters ahead of time. Top-down clustering necessitates a method for breaking a cluster that contains all of the data and then recursively splitting clusters until all of the data has been split into singletons.



Hierarchical agglomerative clustering

IV. CONCLUSION

Cluster I, as seen in the diagram above, represents a client with a high annual income as well as a high annual spend. Cluster II denotes a group with a high yearly income but a low annual expenditure score. Customers in Cluster III have a low yearly income and a low annual expenditure. Cluster IV represents a modest annual income but a large annual expenditure. Cluster V customers have a medium-income and a medium Spending Score.

V. REFERENCES

[1] XIANG-BIN YAN, YI-JUN LI, Customer Segmentation based on Neural Network with Clustering Technique, 5th WSEAS Int. Conf. on Artificial Intelligence, Knowledge Engineering and Data Bases, Madrid, Spain, February 15-17, 2006 (pp265-268).

[2] Ling Luo, Bin Li et. al. "Tracking the Evolution of Customer Purchase Behaviour Segmentation via a Fragmentation-Coagulation Process", Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17).

[3] Mark K.Y.Mak, George T.S.Ho, S.L.Ting,"A Financial Data Mining Model for extracting Customer Behaviour", INTECH open access publisher, 23 July 2011.

[4] Juni Nurma San, Lukito Nugroho, Ridi Ferdiana, P.InsapSantosa, "A Review on Customer Segmentation Technique on E-Commerce", Advance Science Letters, Vol.4, 400-407,2011.

[5] Kishana R. Kashwan, Member, IACSIT, C.M.Velu, "Customer Segmentation using clustering and Data Mining Techniques", International Journal of Computer Theory and Engineering, Vol.5, No.6, December 2013.

[6] LuoYe, CaiQiuru, XiHaixu,LiuYijun and Zhu Ghuangping, "Customer Segmentation for Telecom with the k-means Clustering Method", Information Technology Journal 12(3):409-413,2013.

[7] Yi Zuo,A B M Shawkat Ali, KatsutoshiYada, "Customer purchasing behaviour using statistical learning theory",18th International Conference on Knowledge–Based and Intelligent Information & Engineering Systems-KES2014.

[8] Haiying Ma, "A study on Customer Segmentation for E-Commerce using the Generalized Association Rules and Decision Tree", American Journal of Industrial and Business Management, 2015,5,813-818

[9] Jyoti, Savita Bisnoi " A Predictive Analytics of Cluster using Associative Techniques Tool", IJRDET, Vo - 5,Issue 6, June 2016

[10] R.Kaur,K.Kaur, "Data Mining on Customer Segmentation: A Review", International Journal of Advanced Research in Computer Science, Volume No-5,2017.