iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** VI **Month of publication:** June 2020

DOI: http://doi.org/10.22214/ijraset.2020.6344

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue VI June 2020- Available at www.ijraset.com

Customer Segmentation using Machine Learnin

Patel Monil[1], Patel Darshan[2], Rana Jecky[3], Chauhan Vimarsh[4], Prof. B. R. Bhatt[5]
[1, 2, 3, 4, 5]*Department of Computer Engineering, R.N.G. Patel Institute of Technology, Bardoli, India*

Abstract: In recent years, every e-commerce enterprise focuses on Customer Relationship Management(CRM) to provide the better services to the customer as compared to their competitors. Building a better relationship with customer help the enterprises in increasing profit and customers retention and satisfaction. It is necessary for enterprises to identify the potential customers in the market by mining the customer data to gain profitable insight. One of the efficient way to identify the different customer characteristics is by applying clustering analysis. In this paper, different clustering approach has been presented in order to segment the customer and apply the different marketing strategies accordingly. The possibility of hybrid combination of clustering algorithm can outperform individual model has also been discussed.
Keywords: Customer relationship management,clustering, customers retention,customer characteristics.

I. INTRODUCTION

In the face of product competition, enterprises should mine customer resources to achieve targeted measures for different customers and providing the services what the customer wants[4].The key to enterprise development is to start with the analysis of customers needs and use customer segmentation as the means to identify and analyze various consumer groups in the system, so as to provide different types of customers with distinctive marketing methods and improve their satisfaction.

One of the most useful techniques in business analytics for the analysis of consumer behavior and categorization is customer segmentation[5]. By using clustering techniques, customers with similar means, end and behavior are grouped together into homogeneous clusters[3].Cluster analysis is a kind of algorithm frequently used in data mining technology, which is mainly used in the analysis of enterprise data information to observe distribution characteristics existing in data sets, so as to achieve strategic goals[7].The K-means algorithm has a wide range of applications in helping telecom operators implement customer segmentation and accurately locate customers market needs[10,11].In addition to the K Means different others clustering approaches like Hierarchical clustering, Density based clustering, Affinity Propagation clustering has also been presented. Combining clustering algorithm can result into better clustering results than individual algorithm.

Customers vary in terms of behavior, needs, wants and characteristics and the main goal of clustering techniques is to identify different customer types and segment the customer base into clusters of similar profiles so that the process of target marketing can be executed more efficiently. The advantage and disadvantage of clustering technique has also been discussed for better clustering efficiency.

Certain parameters are considered while segmenting the customer. The clustering parameters can broadly be classified as geographic, demographic, psychographic and behavioural [1]. Predicting the future consumption trend of customers in
the way of segmentation of customer information and consumption behaviour, as well as the profit market planning of enterprises, so as to achieve the goal of reasonable allocation of service resources and the most profitable design of customer marketing programs [13].

II. METHOLOGY

A. Customer Relationship Management(CRM)

The modern marketing approach promotes the usage of CRM as part of the organizations business strategy for enhancing customer service satisfaction[15].CRM enables business enterprises in customer value analysis as well as the targeting of those customers that prove of greater value[3]. It also helps business organizations in developing high-quality and long-term customer-company relationships that increase loyalty and profits.

B. Customer Segmentation

Segmentation, also known as customer segmentation, refers to the process of dividing a market into different buyers with different behaviours, characteristics[5].Customer segmentation refers to a way of dividing according to different characteristics of consumer groups. This theory proposes to study and predict the future consumption trend of customers in the way of segmentation of customer information and consumption behaviour, as well as the profit market planning of enterprises.

C. Clustering

Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups. It is basically a collection of objects on the basis of similarity and dissimilarity between them. Different types of clustering algorithm available for efficiently classification.

D. Use of Customer Segmentation

Target Marketing and Customer Segmentation are so closely related that they are used interchangeably[8].Target marketing refers to the grouping of buyers based on certain characteristics which the firms intend to serve. It has been referred to as a personal branding strategy in the context of a specific customer[16]. Customers in the selected market are segmented into different groups based on their characteristics.

III. CLUSTERING TECHNIQUES

A. K-Means Clustering

K-means clustering algorithm is one of the clustering algorithms based on division. It adopts a heuristic iterative process to re-divide data objects and re-update cluster centres. The basic idea of the algorithm is: suppose a set with element
objects and the number of clusters to be generated[2]. In the first round, a sample element is randomly selected as the initial cluster centre[6], and the distance between other sample elements and the centre point is analysed the clusters are respectively divided according to the distance. In each of the following rounds, the iterative operation of the above steps is continuously performed, and the average value of the element objects obtained this time is taken as the centre point of the next round of clustering until the condition that the clustering centre point no longer changes in the iteration process is met.The specific processing steps are as follows:

```
randomly chose k examples as initial centroids
while true:
   create k clusters by assigning each
       example to closest centroid
   compute k new centroids by averaging
       examples in each cluster
   if centroids don't change:
       break
```

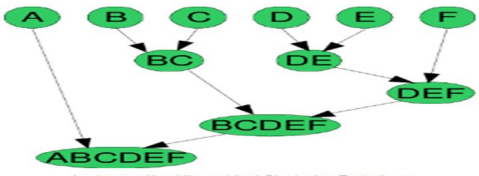
Fig. 1 K-Means Algorithm

B. Hierarchical Clustering

Hierarchical clustering is a method of cluster analysis which builds a hierarchy of data points as they move into a cluster or out of it [14]. Strategies for this algorithm generally fall into two categories:

1) *Agglomerative* - This clustering algorithm does not require us to prespecify the number of clusters. Bottom-up algorithms treat each data as a singleton cluster at the outset and then successively agglomerates pairs of clusters until all clusters have been merged into a single cluster that contains all data.

- Step- 1: In the initial step, we calculate the proximity of individual points and consider all the six data points as individual clusters as shown in the image below.



Agglomerative Hierarchical Clustering Technique

- Step- 2: In step two, similar clusters are merged together and formed as a single cluster. Let's consider B,C, and D,E are similar clusters that are merged in step two. Now, we're left with four clusters which are A, BC, DE, F.
- Step- 3: We again calculate the proximity of new clusters and merge the similar clusters to form new clusters A, BC, DEF.
- Step- 4: Calculate the proximity of the new clusters. The clusters DEF and BC are similar and merged together to form a new cluster. We're now left with two clusters A, BCDEF.
- Step- 5: Finally, all the clusters are merged together and form a single cluster.

Fig. 2 Hierarchical Agglomerative clustering

2) *Divisive* - Also known as top-down approach. This algorithm also does not require to prespecify the number of clusters. Top-down clustering requires a method for splitting a cluster that contains the whole data and proceeds by splitting clusters recursively until individual data have been splitted into singleton cluster.

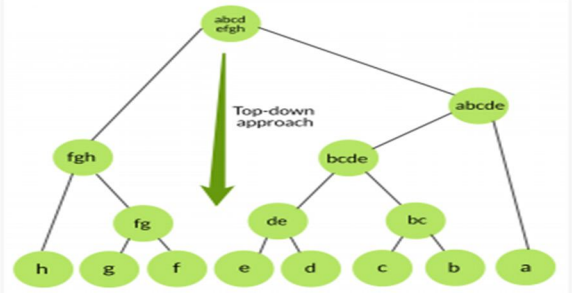


Fig. 3 Hierarchical Divisive clustering

Divisive algorithm is also more accurate. Agglomerative clustering makes decisions by considering the local patterns or neighbor points without initially taking into account the global distribution of data. These early decisions cannot be undone. whereas divisive clustering takes into consideration the global distribution of data when making top-level partitioning decisions.

C. Density Based CLustering

The DBSCAN algorithm is based on this intuitive notion of clusters and noise[12]. DBSCAN is a clustering method that is used in machine learning to separate clusters of high density from clusters of low density.The key idea is that for each point of a cluster, the neighborhood of a given radius has to contain at least a minimum number of points[9]. DBSCAN algorithm requires two parameters:

1) *eps :* It defines the neighborhood around a data point i.e. if the distance between two points is lower or equal to 'eps' then they are considered as neighbors. If the eps value is chosen too small then large part of the data will be considered as outliers. If it is chosen very large then the clusters will merge and majority of the data points will be in the same clusters.
2) *MinPts:* Minimum number of neighbors (data points) within eps radius. Larger the dataset, the larger value of MinPts must be chosen. As a general rule, the minimum MinPts can be derived from the number of dimensions D in the dataset as, MinPts >= D+1.

In this algorithm, we have 3 types of data points.A point is a core point if it has more than MinPts points within eps. A point which has fewer than MinPts within eps but it is in the neighborhood of a core point is border points. A point which is not a core point or border point is noise.

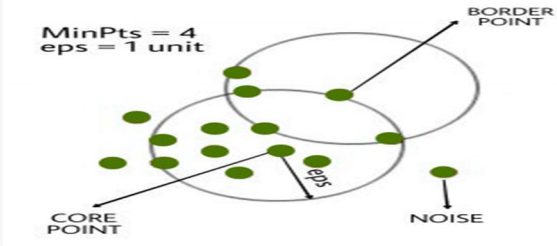


Fig. 4 Density based clustering

The algorithm is as follows:

1. Find all the neighbor points within eps and identify the core points or visited with more than MinPts neighbors.
2. For each core point if it is not already assigned to a cluster, create a new cluster.
3. Find recursively all its density connected points and assign them to the same cluster as the core point.
   A point *a* and *b* are said to be density connected if there exist a point *c* which has a sufficient number of points in its neighbors and both the points *a* and *b* are within the *eps distance*. This is a chaining process. So, if *b* is neighbor of *c*, *c* is neighbor of *d*, *d* is neighbor of *e*, which in turn is neighbor of *a* implies that *b* is neighbor of *a*.
4. Iterate through the remaining unvisited points in the dataset. Those points that do not belong to any cluster are noise.

Fig. 5 Density based clustering algorithm

D. Affinity Propagation Algorithm

AP algorithm is a clustering algorithm based on the similarity between N data samples. The AP algorithm doesn't need to give the initial cluster centre or the number of clusters first, but treats all samples as potential cluster centres, called exemplar; it also establishes attractiveness information (that is, the similarity between any two data samples) for each data sample with other data samples with the help of euclidean distance and stores in similarity matrix which describe similarity between datapoints[4]. In the AP algorithm, two important parameters are the preference, which controls how many exemplars (or prototypes) are used, and the damping factor which damps the responsibility and availability of messages to avoid numerical oscillations when updating these messages. Different forms of affinity propagation are available like adaptive affinity propagation, partition affinity propagation etc in order to deal with the clustering time and clustering efficiency of affinity propagation .The flowchart of affinity propagation is as follows:

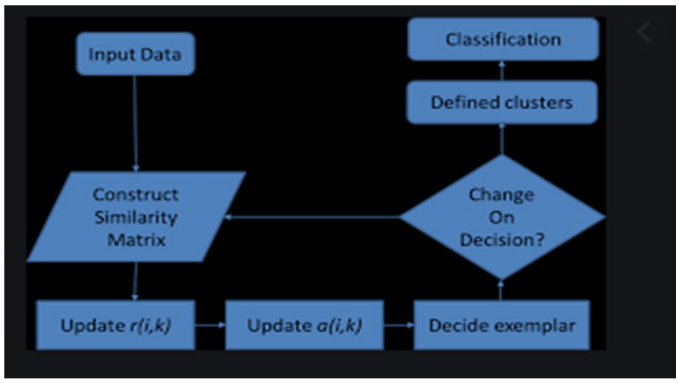


Fig. 6 Flowchart of affinity propagation

The responsibility matrix R has values r(i, k) that quantify how well-suited x_k is to serve as the exemplar for x_i, relative to other candidate exemplars for x_i [4]. The availability matrix A contains values a(i, k) that represent how appropriate it would be for x_i to pick x_k as its exemplar, taking into account other points' preference for x_k as an exemplar.

IV. COMPARISION CLUSTERING TECHNIQUE

It is necessary to compare the different clustering technique discussed in order to identify which clustering algorithm should use in which situation. Comparision table is as follows:

Table 1. Comparision clustering algorithm

Criteria	K-Means	Hierarchical Clustering	Density based Clustering	Affinity Propagation
Computation Speed	High	Low	Low	Low
Clustering time	Less	More	More	More
Granularity	Yes	No	Yes	No
Effect on size of Data	Good	Not good	Not good	Not good
Handle Dynamic Data	Yes	No	Yes	Yes
Clustering result efficiency	Medium	Low	Medium	High

Each clustering algorithm have their own advantages as well as disadvantage with respect to specific situation. K Means is most widely used clustering algorithm for customer segmentation. The K Means require initial number of cluster which is difficult to predict which can affect clustering result. Hierarchical clustering does not require initial number of cluster condition and the time complexity of hierarchal clustering is high and suited for small to medium size dataset. Density based clustering algorithm can be used to find arbitrarily shaped clusters but it not suited for more density difference in datapoints as well as cluster result efficiency is not good. Affinity Propagation algorithm does not require initial cluster as well as clustering result efficiency is high and clustering time is high which somehow guarantee high clustering efficiency and applicable from small to medium size dataset.

V. CONCLUSION

When dealing with large magnitude of data, organizations need to make use of more efficient clustering algorithms for customer segmentation. These clustering model need to possess the capability to process this enormous data effectively. Each of the above discussed clustering algorithms come with their own set of merits and demerits. With different technique pointed above, it came to light that hybrid approach of combining the algorithm can be useful depending upon different situation and the requirement and apply the strategy accordingly. The selection process of clustering technique would require considerable time for studing and implementing as well as processing of data with adequate understanding of goals and apply the algorithm on requirement basis.
Hence, it would be helpful for the organization for identifying the distinct group of customers that increases their profit. It also help them in maintaining customer relationship and customer retention by executing different marketing strategies.

REFERENCES

[1] Bhade, Kalyani, et al. "A Systematic Approach to Customer Segmentation and Buyer Targeting for Profit Maximization." 2018 9th International Conference on Computing,Communication and Networking Technologies (ICCCNT). IEEE, 2018.

[2] Kansal, Tushar, et al. "Customer Segmentation using K-means Clustering." 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS). IEEE, 2018.

[3] Tripathi, S., A. Bhardwaj, and E. Poovammal. "Approaches to clustering in customer segmentation." International Journal of Engineering & Technology 7.3.12 (2018): 802-807.

[4] Deng, Yulin, and Qianying Gao. "A study on e-commerce customer segmentation management based on improved K-means algorithm." *Information Systems and e-Business Management* (2018): 1-14.

[5] Hoegele D, Schmidt SL, Torgler B (2016) The importance of key celebrity characteristics for customer segmentation by age and gender: does beauty matter in professional football? RMS 10(3):601–627.

[6] Arora P, Deepali, Varshney S (2016) Analysis of K-means and K-medoids algorithm for big data. Procedia Comput Sci 78:507–512.

[7] de Oña, Juan, Rocío de Oña, and Griselda López. "Transit service quality analysis using cluster analysis and decision trees: a step forward to personalized marketing in public transportation." Transportation 43.5 (2016): 725-747.

[8] Huang S, Wang Q, School B (2014) Method for customer segmentation based on three-way decisions theory. J Comput Appl 34(1):244–248.

[9] Loh, Woong-Kee, and Young-Ho Park. "A survey on density-based clustering algorithms." Ubiquitous information technologies and applications. Springer, Berlin, Heidelberg, 2014. 775-780.

[10] Luo Y, Cai Q, Xi H et al (2013) Customer segmentation for telecom with the k-means clustering method. Inf Technol J 12(3):409–413

[11] Qiuru, Cai, et al. "Telecom customer segmentation based on cluster analysis." 2012 International Conference on Computer Science and Information Processing (CSIP). IEEE, 2012.

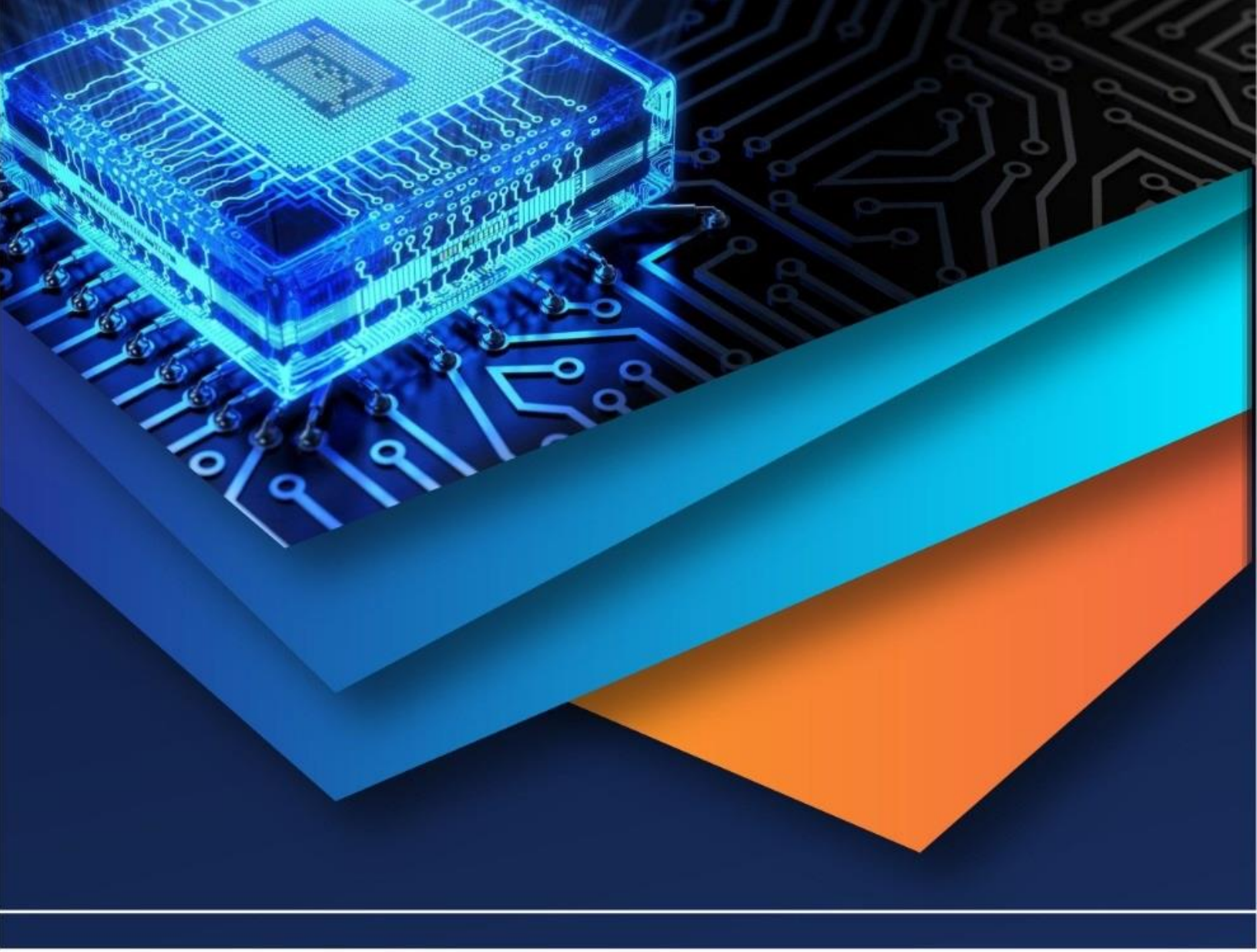
[12] Xu, Huajie, and Guohui Li. "Density-based probabilistic clustering of uncertain data." 2008 International Conference on Computer Science and Software Engineering. Vol. 4. IEEE, 2008.

[13] Lee, Jang Hee, and Sang Chan Park. "Intelligent profitable customers segmentation system based on business intelligence tools." Expert systems with applications 29.1 (2005): 145-152.

[14] Salvador, Stan, and Philip Chan. "Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms." 16th IEEE international conference on tools with artificial intelligence. IEEE, 2004.

[15] Rygielski, Chris, Jyun-Cheng Wang, and David C. Yen. "Data mining techniques for customer relationship management." Technology in society 24.4 (2002): 483-502.

[16] Aaker, Jennifer L., Anne M. Brumbaugh, and Sonya A. Grier. "Nontarget markets and viewer distinctiveness: The impact of target marketing on advertising attitudes." Journal of Consumer Psychology 9.3 (2000): 127-140.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY