A Systematic Review on Customer Segmentation Approach using Machine Learning

Nikita Jha(0225CS17MT08)
(Department of CSE, Global Nature Care Sangathan's Group of Institutions, Jabalpur)

Abstract— **Nowadays, maintaining customer loyalty and customer focus are the main challenges facing the retail industry. This raises the need to strengthen marketing strategies over time. This paper proposes a systematic approach to targeting customers and providing companies with maximum profits. An important starting point is to analyze the sales data obtained from the purchase history and determine the parameters with maximum correlation. Depending on the respective clusters, appropriate resources can be sent to profitable customers using machine learning algorithms like K-Means clustering,** hierarchical clustering, DBSCAN clustering **is used for customer segmentation used for providing appropriate recommendations to the customers. Following the successful clustering process, take the lead Organizations can make accurate decisions and organizations can delivers new products and services, and some can changes to current product services according to customer needs by properly identifying customers.**

Keywords— **Customer Segmentation, DBSCAN, Clustering, K-Means.**

I. Introduction

Customer relationship management (CRM) is a strategy in market that enables the retail industry to learn about customers' behaviors and needs that would help in developing strong relationships and customer loyalty. Advancements in technology have facilitated the above objective successfully in recent years as they help to solve business questions that in the past were too time-consuming to pursue because of manual computation. Particularly through data mining and the extraction of hidden patterns of customer purchases from large databases, organizations can identify valuable customers, predict their future behaviors. This enables firms to make proactive and knowledge-driven decisions.

Customer clustering and buyer targeting are the two intelligent components of Customer Relationship Management [1]. In segmentation, deciding upon the optimum number of clusters and the variables used for clustering is an important step. As there could be a large number of variables that can be used to differentiate customers, optimum number of variables have to be determined that can form the most distinct clusters [2].

The clustering parameters can broadly be classified as geographic, demographic, psychographic and behavioral.

Geographic clustering will group customers belonging to the similar area together with an assumption that the needs of the same area people will be similar.

Psychographic clustering includes grouping on the basis of personality, lifestyle or the social class. However, this kind of personalized recommendations might affect the privacy of the customer.

Demographic segmentation groups the market based on gender, age, education, income, occupation, religion, and nationality. This may result in ignoring the fact that customers may not act on the basis of these parameters.

Behavioral parameters include clustering on the basis of recency and frequency of purchases. Recency is how recent was the last purchase of customer and frequency is how often the purchase happens. Despite the simplicity of these param- eters, the clustering on this basis gives classes of customers which can be then handled differently leading to boost sales.

One of the most successful ways to target customers with marketing campaigns is through automated merchandising. This concept involves providing the customers with relevant and customer specific recommendations and this can be achieved through the use of recommender systems. Content based, Collaborative and Hybrid are three major types of recommender systems.

Literature Review

Clustering is a part of unsupervised learning. It has got various applications in numerous fields such as artificial in- telligence, bioinformatics, pattern recognition, segmentation and machine learning. The appropriate clustering algorithm needs to be decided on the basis of the scenarios based on the accuracy and efficiency [5]. Recommender systems have been used in various applica- tions today. Due to the ever-increasing data, these recom- mender systems face three common problems - cold start, sparsity and overspecialization.

Cold start problem occurs in at least one of the situations: A new user has to be categorized or a new product has to be recommended. It becomes difficult for the recommender systems to suggest products to new users as the system does not have the history of their purchases. Also, cold start problem for the new item is that the system does not have enough reviews or ranking related to that item which creates a problem to recommend the item to the appropriate user [7].

This results in inefficiency in user categorization and product recommendation. In order to improve the quality of recommendation, hybrid recommendation technique can be used which collectively makes use of the merits of content- based filtering for new products and collaborative filtering for new users. As per [8], it not only utilizes the customer purchase history but also incorporates contextual information of corresponding items through customer and item profiling.

The second challenge is of sparsity. Every store has a huge number of customers as well as products. Every user purchases or ranks a limited number of items. When a consumer-product matrix is created based on the past purchases of the consumers, there are only a few elements who have a value other than 0, making it a very sparse matrix. Sparsity is an issue as when the similarity of two customers is calculated based on the consumer-product matrix, the probability to get the similarity very low or even zero, increases [9]. This results in difficulties in finding appropriate matches and consequently produces recommendations with poor accuracy. In numerous recent models, Latent Feature Indexing (LSI) is applied to reduce the dimensionality of the user-item utility matrix. It utilizes Singular Value Decomposition (SVD) as its underlying matrix factorization algorithm. Dimensionality reduction as explained in [10], not only deals with sparse values by increasing the density of the matrix but also improves performance by boosting computation speed.

The third challenge that is encountered is overspecialization. All recommender systems try to suggest items the user is already familiar with. In doing this, there is no surprise factor in the recommendations as it obstructs the users from a totally new and different product being recommended to them. This challenge can be overcome using cosine similarity which is a collaborative filtering technique based on neighborhood [10]. It considers the levels of similarity between the user and the items purchased by them and their candidate neighbors.

SEGMENTATION TYPES

Behavioural Segmentation

Behavioural segmentation is one of the efficient segmentation methods. It is the most used segmenting method because it is easy to be collecting the data. The segmentationis done on the basis of the customer's behaviour. The behavioural pattern can be identified through the buying pattern, the quantity of the product, the quality of the product,usage, brand of the product etc. This will help the companiesto identify the behaviour of the customers and they will be ableto provide the products according to this data.[6]

Psychographic Segmentation

Psychographic segmentation is the segmentation based purely on the customer lifestyle, beliefs, opinion, activities, interest, jobs etc. This will help the company to identify the customer needs according to their lifestyle, attitude, job etc.[5]

Geographic Segmentation

Geographic segmentation is the segmentation based on areas such as country, states, urban, rural, coastal etc. This will help the company to focus on the customers more in the region with fewer sales and also helps to identify the priority of orders in each region.[15]

Demographic Segmentation

The demographic segmentation is the segmentation basedon life stage information or socio-demographic information such as age, gender, education, occupation, marital status, income etc. This segmentation helps to focus more on products for each life stage.[5]

Value-based Segmentation

In value-based segmentation, the customers are segmented according to the value. This helps to identify the most valuable customer and the values of each customer and the changes ofvalues by the change in time.[6]

Current Value= (Average amount asked to pay for acustomer - Cumulative amount in arrears for the customer/total period of use).[7]

Propensity based Segmentation

Propensity based segmentation is segmentation based on some scores such as churn scores, propensity scores, etc. This segmentation contains computation and the binning of customers into groups according to the score [6].

CLUSTERING TECHNIQUES

K- Means Clustering Algorithm

K-Means is one of the most used clustering algorithms for segmentation. It is easy to use and it is very efficient. K-Means algorithm is proposed by J.B.MacQueen. K-Means clusteringalgorithm aims to minimise the cluster performance index, square error term, and the error criterion. In K-Means the M- points in N-dimension are divided into K-cluster assuming that k as their centroids. Here we try to optimise the result byplacing the dots in a wise manner as the distance between centroids of sample points is as far as possible and calculate the distance and then we will place the sample points with thecriterion of minimum distance with centroids and the iterativeprocess continues until there is no change in distance occur in the further iteration process.[10],[11]

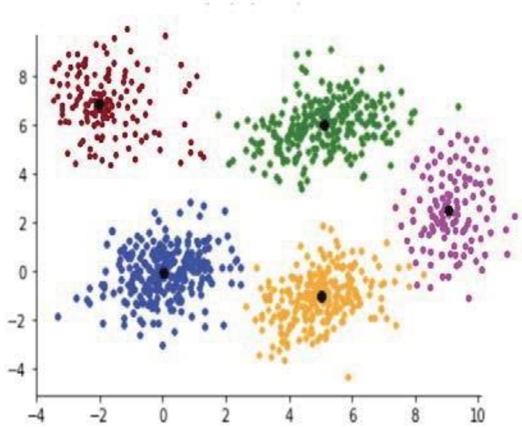


Fig.1. Clusters Due to K-Mean Clustering

Hierarchal Clustering Algorithm

Hierarchal Clustering is the method of clustering which builds a hierarchy model of data points in the cluster as the move into the cluster or move out of the cluster. There are two categories for this clustering. [9],[12].

a. *Agglomerative*

This is a bottom-up algorithm treat where each singleton cluster merges up to most similar ones and at last merging up into a single cluster containing all the singleton clusters

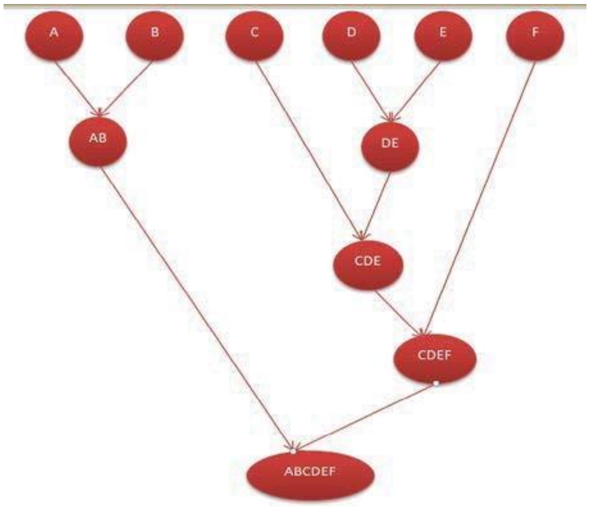


Fig.2.Image of clusters formed by agglomerative.

a. *Divisive*

This is a top-down treat where the one single cluster divides into smaller groups until each singleton clusters are found.

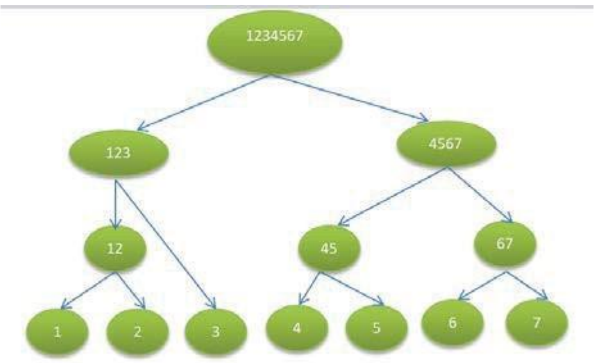


Fig.4.Image of clusters formed by divisive.

B. *Density-Based Clustering*

Density-based clustering also known as DBSCAN. DBSCAN is based upon the notion of clusters and noise whichis used to get a cluster of arbitrary shapes. DBSCAN is used to differentiate the high-density cluster from the low-density cluster. DBSCAN is efficient for the large data set. In this, themain idea used is the neighbourhood of a given radius of eachpoint at least a minimum number of points in it. From the graph, we can easily identify the cluster points and the noisepoints. Cluster points are the points which are clusteredtogether and it contains a high density than the points outsidethe clusters and noise points are points which are not belonging to any of the clusters. [9]

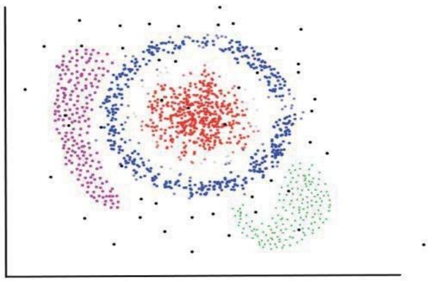


Fig.3 Image of clusters formed by DBSCAN

Affinity Propagation Clustering

Affinity Propagation Clustering also called APC algorithm which is based upon the similarity of N data set points. APC algorithm treats all sample points as exemplar orcluster centres. APC finds out the similarity between two datasample with the help of Euclidean distances and stores in a matrix called the similarity matrix. There are two matrices involved in this method they are responsible matrix (R) and availability matrix (A).R(i,j) matrix measures the accumulated evidence of how well-suited sample x_j serves as the exemplar. A(i,k) measures the accumulated evidence of how appropriatexi chooses x_k as its exemplar.[9]

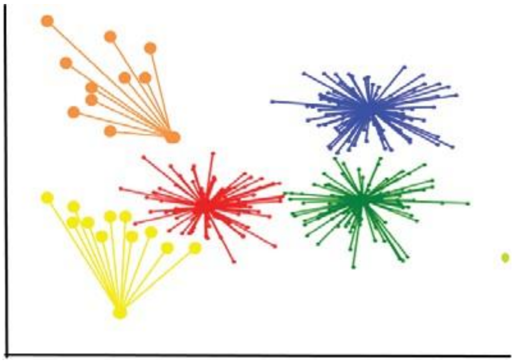


Fig.5.Image of clusters formed by APC

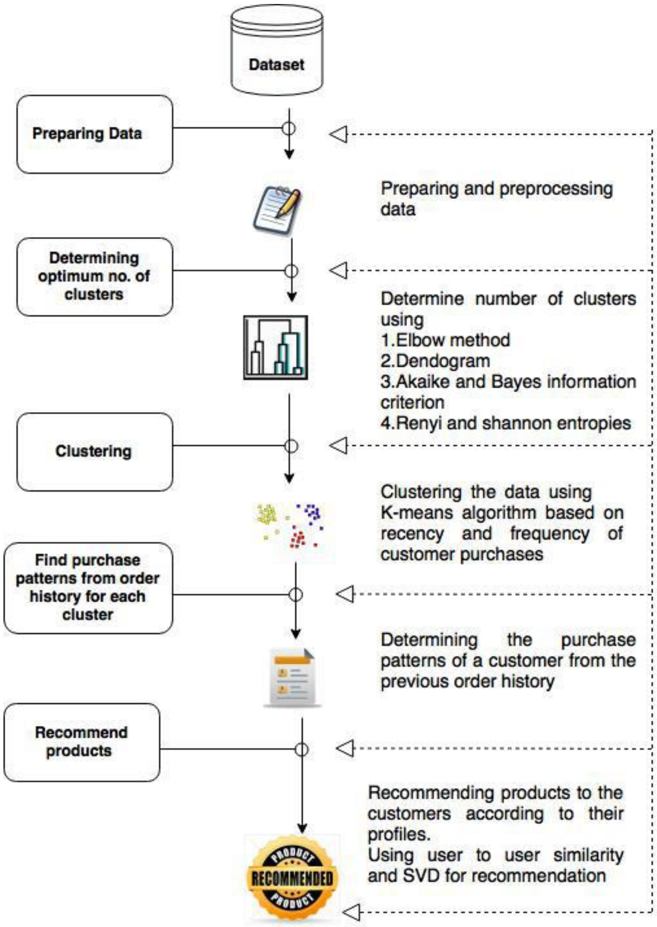
Mean shift Clustering

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 work 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

Proposed Methodology Architecture


Dataset
Preparing Data
Determining optimum no. of clusters
Clustering
Find purchase patterns from order history for each cluster
Recommend products
PRODUCT RECOMMENDED PRODUCT

Preparing and preprocessing data

Determine number of clusters using
1.Elbow method
2.Dendogram
3.Akaike and Bayes information criterion
4.Renyi and shannon entropies

Clustering the data using K-means algorithm based on recency and frequency of customer purchases

Determining the purchase patterns of a customer from the previous order history

Recommending products to the customers according to their profiles.
Using user to user similarity and SVD for recommendation

CONCLUSION

The competition among e-commerce business is increasing by each day the importance of customer segmentation is also increasing. Maintaining a customer is a crucial task for the company. Without understanding who is your best customer, what your customer needs etc. the business cannot be able to focus on the customers and the services. Customer segmentation is the best solution to identify this problem. It helps the business to focus more on marketing. The paper concludes that customer segmentation helps to improve the e-commerce business and the best technique that can be used for customer segmentation.

REFERENCES

[1] R. Siva Subramanian, Dr. D. Prabha, ‘A Survey on Customer Relationship Management’, *International Conference on Advanced Computing and Communication Systems*, Coimbatore, India, January 2017.

[2] Witschel H.F., Loo S., Riesen K.‘How to Support Customer Segmentation with Useful Cluster Descriptions’ *Advances in Data Mining:Applications and Theoretical Aspects*, vol. 9165, Springer, 2015.

[3] Ina Maryani, Dwiza Riana, ‘Clustering and profiling of customers using RFM for customer relationship management recommendations’ *5th International Conference on Cyber and IT Service Management*, Denpasar, Indonesia, August 2017.

[4] Anshu Sang, Santosh K. Vishwakarma, ‘A Ranking based Recommender System for Cold Start & Data Sparsity Problem’*Tenth International Conference on Contemporary Computing*, , Noida, India, August 2017.

[5] Ilung Pranata, Geoff Skinner, ‘Segmenting and targeting customers through clusters selection & analysis’, under review for *International Conference on Advanced Computer Science and Information Systems*, Depok, Indonesia, October 2015.

[6] Raj Bala, Sunil Sikka, Juhi Singh, ‘A Comparative Analysis of Clustering Algorithms’, *International Journal of Computer Applications*, pp. 35-39, August 2014.

[7] Dr. Sarika Jain, Anjali Grover, Praveen Singh Thakur, Sourabh Kumar Choudhary ‘Trends, Problems And Solutions of Recommender System’ *International Conference on Computing, Communication and Automation*, Noida, India, May 2015.

[8] Pratik Ghanwat, Anu Chacko, ’Improved Personalized Recommenda- tion System with Better User Experience’,*International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, Udupi, India, September 2017.

[9] Marius Kaminskas, Derek Bridge, Franclin Foping, Donogh Roche, ‘Product Recommendation for Small-Scale Retailers’, *16th International Conference on Electronic Commerce and Web Technologies*, vol. 239, Springer, Spain, September 2015.

[10] Zan Huang, Daniel Zeng, Hsinchun Chen, ‘A Comparative Study of Recommendation Algorithms in E- Commerce Applications’, *Proceedings of the IEEE Region 10 Conference*, The Pennsylvania State University, pp.1-23. Available: http://citeseerx.ist.psu.edu. [Accessed: March 08, 2018]

[11] The Instacart Online Grocery Shopping Dataset 2017.[Online]. Available: https://www.instacart.com/datasets/grocery-shopping-2017. [Accessed: September 25, 2017]

[12] Trupti M. Kodinariya, Prashant R. Makwana, ‘Review on determining number of Cluster in K-Means Clustering’, *International Journal of Advance Research in Computer Science and Management Studies*, vol. 1, pp. 90-95, November 2013.

[13] P. Praveen, B. Rama, ‘An Empirical Comparison of Clustering using Hierarchical methods and K-means’, *2nd International Conference on Advances in Electrical, Electronics, Information, Communication and Bioinformatics*, Chennai, India, February 2016.

[14] Xiaojun Chen, Yixiang Fang, Min Yang, Feiping Nie, Zhou Zhao, Joshua Zhexue Huang, ‘PurTreeClust: A Clustering Algorithm for Customer Seg- mentation from Massive Customer Transaction Data’, *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, pp. 559-572, March 2018.

[15] Naznin Akter, A. H. M. Sajedul Hoque, Rashed Mustafa, Mohammad Sanaullah Chowdhury, ‘Accuracy Analysis of Recommendation System Using Singular Value Decomposition’, *19th International Conference on Computer and Information Technology*, North South University, Dhaka, Bangladesh, December 2016.

[16] H.Y. Ma, “A Study on Customer Segmentation for E-Commerce Using the Generalized Association Rules and Decision Tree,” American Journal of Industrial and Business Management, vol. 5, pp. 813-818. , 2015 .http://dx.doi.org/10.4236/ajibm.2015.512078.

[17] B. Kaur, and P.K. Sharma, “Implementation of Customer Segmentation Using Integrated Approach,” .International Journal of Innovative Technology and Exploring Engineering (IJITEE), vol. 8, pp. 770-772. 2019.

[18] R.-S. Wu, and P.-H. Chou, “Customer Segmentation of Multiple Category Data in E-Commerce Using a Soft-Clustering Approach,” Electronic Commerce Research and Application, vol. 10, pp. 331- 341. 2011. doi:10.1016/j.elerap.2010.11.002.

[19] R. Kaur, and K. Kaur, “Data Mining on Customer Segmentation: A Review,” International Journal of Advanced Research In Computer Science, vol.8, no. 5,pp. 857-861. 2017.

[20] J. N. Sari, L. E. Nugroho, R. Ferdiana, and P.I. Santosa , “Review onCustomer Segmentation Technique on E-Commerce,” American Scientific Publishers, vol.4,no.2,pp.400-407. 2011. doi:10.1166/asl.2011.1261.

[21] H. Ziafat and M. Shakeri , “Using Data Mining Techniques in Customer Segmentation,” Int. Journal of Engineering Research and Applications,vol. 4,no.9,pp.70-79. 2014.

[22] S.-Y. Kim, T.-S. Jung, E.-H. Suhand and H.-S. Hwang, “Customer Segmentation and Strategy Development Based on Customer Lifetime Value: A Case Study,” Expert Systems with Applications, vol. 31, pp. 101-107. 2006. doi:10.1016/j.eswa.2005.09.004.

[23]]. H. Valecha, A. Varma, I. Khate, A. Sachdeva and , M. Goyal “Prediction of Consumer Behaviour using Random Forest,” 5th IEEEUttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON).IEEE. 2018.

[24] P. Monil, P. Darshan, R.Jecky, C. Vimarsh, and, B.R. Bhatt, “Customer Segmentation Using Machine Learning,” International Journal for Research in Applied Science and Engineering Technology (IJRASET), vol. 8, no. 6, pp.2104-2108. 2020. http://doi.org/10.22214/ijraset.2020.6344.

[25] S. Tripathi, A. Bhardwaj, and E, P, “Approaches to Clustering in Customer Segmentation,” International Journal of Engineering and Technology, vol. 7,no. 3.12, pp. 802-807. 2018.

[26] K.R Kashwan and C.M .Velu, “Customer Segmentation Using Clustering and Data Mining Techniques,” International Journal of Computer Theory and Engineering, vol.5,no. 6,pp. 856-861. DOI: 10.7763/IJCTE.2013.V5.811. 2013

[27] Y. Rani, and H.Rohil, “ A Study of Hierarchical Clustering Algorithm,” International Journal of Information and Computation Technology, vol. 3,no. 11,pp. 1225-1232. 2013.

[28] G. Dongari, “How to Create New Feature Using Clustering”. 2017 Retrieved 26, September, 2020. https://towardsdatascience.com/how- to-create-new-features-using-clustering-4ae772387290.

[29] Retrieved 25 September, 2020. https://www.shopify.com/encyclopedia/what-is-ecommerce.

[30] Martin, G, “The Importance of Market Segmentation”, American Journal of Business Education, vol.4,no. 6 pp.15-18. (2011).

[31] M.Namvar, , M. R Gholamian, and S. K Abi, “A Two Clustering Method for Intelligent Customer Segmentation”.2010 International Conference on Intelligent Systems, Modelling and simulation.(IEEE).pp.215-219. 2010.DOI: 10.1109/ISMS.2010.48