International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 06 | June 2019 www.irjet.net p-ISSN: 2395-0072

CUSTOMER SEGMENTATION FROM MASSIVE CUSTOMER TRANSACTION DATA

Neethu CM[1], Anitha Abraham[2]

[1]*Post Graduate Student*
[2]*Assistant Professor, College of Engineering Kidangoor*

Abstract - *In this internet era, more and more people use online shopping. Analysing massive customer transaction data about these online activities can be used to improve the business and to satisfy customer demands in a better way. In this research paper we try to study different methods employed to analyse the customer transaction data. In our study we have studied methods like K-Means clustering, PAM clustering, Agglomerative, Divisive and Density Based clustering methods. Based on our study we have identified that K-Means is the widely used clustering method.*

Key Words: Clustering, Partitional clustering, Hierarchical Clustering

1. INTRODUCTION

Customer Segmentation is also known as clustering of customers. Clustering can be considered the most important unsupervised learning problem. It deals with finding a structure in a collection of unlabelled data [1] et al says the definition of clustering could be the process of organizing objects into groups whose members are similar in some way. A cluster is therefore a collection of objects which are similar between them and are dissimilar to the objects belonging to other clusters. [2] Clustering is most important technique used in data mining. The most commonly used algorithms in Clustering are Hierarchical and Partitioning. [3] Pramod Gupta says that applications of clustering in various field like In Biology, clustering has been used to find groups of genes that have similar functions. In Information Retrieval, clustering can be used to group search results of a query into a small number of clusters, each of which capturing a particular aspect of the query. In Geology, cluster analysis has been applied to find patterns in the atmospheric pressure of polar regions and areas of the ocean that have a significant impact on land climate. In Medicine, cluster analysis can also be used to detect patterns in the spatial or temporal distribution of diseases like cancer, autism, etc.

2. RELATED WORK

Customer Segmentation is also called clustering of customers. It achieving successful modern marketing and customer relationship management. There are different clustering methods available. Data objects are divided into non overlapping clusters so that every and each object is in one subset in partitional method. [12] explained K-Medoids, PAM, CLARA (Clustering Large Applications) and CLARANS. K-means is the most easy and wide used partition method which describes details in [14], [15],[16],[17],[18],[19]. L.Rokach et al [19] says that the K-means algorithm may be viewed as a gradient-decent procedure. PAM is similar to K-means and find median instead of doing mean. PAM is to find the representative object for each cluster, this representative object called medoids which means that the most centrally located point in the cluster. Consider Oj is data object and Om is selected medoids [24] we can say that the Oj belongs to the cluster Om. The quality of clustering is determined by average dissimilarity between an object and the medoids of its cluster. RuiXu et al proposed algorithm which is in [21]. J.Han et al [15] proposed CLARA algorithm is used to handle large data set and relies on sampling. CLARA drawn multiple samples and gives best clustering results. Quality of clustering is determined with average dissimilarity of all objects in the entire data set in [19] describes the algorithm. Han et al [15] proposed CLARANS method for clustering polygonal objects. CLARANS is main memory clustering techniques and it uses K-medoids method for clustering. Because K-medoids is very robust and to avoid outliers. Application is that CLARANS is used to cluster spatial coordinates. Since the parameter maxneighbour is set to high, it is very effectively same as the quality of the clustering produced by PAM. Same as for lower value of maxneighbour produces a lower clustering quality.

Hierarchical method is one of the clustering method. It can be categorized into two, they are Agglomerative method and Divisive method. Cen Li et al [22] proposed an agglomerative method. It starts with many small clusters so it is called bottom up approach. Each and every iteration each smaller clusters combine to form large clusters. And finally we got a large cluster. [15] The hierarchical clustering methods could be further divided according to the measure is calculated. Single-link cluster, Complete-link cluster, Average-link cluster. [23] et al proposed Balanced Iterative Reducing and Clustering using Hierarchies (ie) BIRCH it is an agglomerate technique, it is used for big databases. This method clusters the incoming multi-dimensional metric data points incrementally and dynamically to provide quality clusters. Guha et al [4] developed another agglomerative hierarchical clustering algorithm, ROCK. Grouping data with attributes or distinct non-metric attributes. This method is a measurement of link used to reveal the relation between combine of objects and their common behaviour. George Karypis et al [24] developed a technique known as

CHAMELEON is type of clustering algorithm based on agglomerative HC with k-nearest neighbour graph. In which an edge is eliminated if both vertices are not within the k-nearest closest points related to each other. In this method, authors ignored the issue of scaling to large data sets that cannot fit in the main memory. CLICK is another agglomerative algorithm based on the calculation of the minimum weight cut to form clusters [30]. Here, we use the weighted graph and the edge weights are assigned a new interpretation. By combining probability and graph theory, the edge weight between two node is calculated. CLICK assumes the similarity values within clusters and between the clusters and follow Gaussian distributions. CLICK recursively checks the current sub graph and generates a kernel list, which consists of the components satisfying some criterion function. Using kernels as basic set of clusters and CLICK carried out singleton clusters. These clusters contain only one node and merge to generate resulting clusters. Additional heuristics are provided to accelerate the algorithm performance. CAST is another agglomerative algorithm used probabilistic model for graph based theory clustering algorithms [34]. Cast is the heuristics original theoretical version and creates clusters sequentially and each cluster begin with unassigned data point randomly. SudiptoGuha et al proposed CURE algorithm [26]. CURE is a novel hierarchical clustering Algorithm. In CURE, a constant number c of scattered points in a cluster are first chosen. The scattered points capture the shape and extent of the cluster and these points are next shrunk towards the centroid of the cluster by a fraction.

[5] TengkeXiong et al proposed an DHCC (Divisive Hierarchical Clustering of Categorical Data), it is a divisive hierarchical algorithm for categorical data. This method uses MCA (Multiple Correspondence Analysis) is carried out by performing a standard correspondence analysis (CA) on an indicator matrix. DHCC starts with an all clusters containing all the categorical objects, and repeatedly chooses one cluster to split into two sub clusters. A binary tree is employed to represent the hierarchical structure of the clustering results. In DHCC, divides the cluster Cp involves finding a sub optimal solution to the optimization problem on the data set Cp with K =2. DHCC consists of two phases, initial splitting phase and refinement phase. [29] R.Datta et al and [36] M s Lew et al proposed one of the current techniques for image retrieval is content-based image retrieval. In content based image retrieval approach visual features such as color feature, texture feature, shape feature and local features are automatically extracted from the image objects and organized as feature vectors. Then at search phase, after selecting the query image by user, retrieval engine retrieves the most similar images to the query image by performing similarity comparison between query feature vector and all the feature vectors in database.

Jorg Sander et.al proposed [33] Density-Based Clustering for spatial databases. In which DBSCAN relies on a density based notion of clusters and is designed to discover clusters of arbitrary shape as well as to distinguish noise. The generalized algorithm called GDBSCAN can cluster point objects as well as spatially extended objects according to both their spatial and their non-spatial attributes. [35] Markus M. Breunig et.al proposed LOF: Identifying Density-Based Local Outliers. It introduce a new method for finding outliers in a multidimensional data set. We introduce a local outlier (LOF) for each object in the data set, indicating its degree of outlierness. This is, to the best of our knowledge, the first concept of an outlier which also quantifies how outlying an object is. The outlier factor is local in the sense that only a restricted neighbourhood of each object is taken into account.

3. PROPOSED FRAMEWORK

In this experiment created an online shopping site. For that used product images and its descriptions from the internet. Download the images of each category such as Electronics, Clothes, Home Appliance etc.

3.1 Problem Statement

Let S = (S_1, S_2-------S_n) be n transaction record and (i_1, i_2-----i_m) be m items in S. Transaction record S_i is an item set represented as (x_1, x_2------x_z) where x_j € I for1≤j≤z. Efficiently create a product clusters (c_1, c_2-----c_m) from the product I based on the transaction set S.

3.2 System Architecture

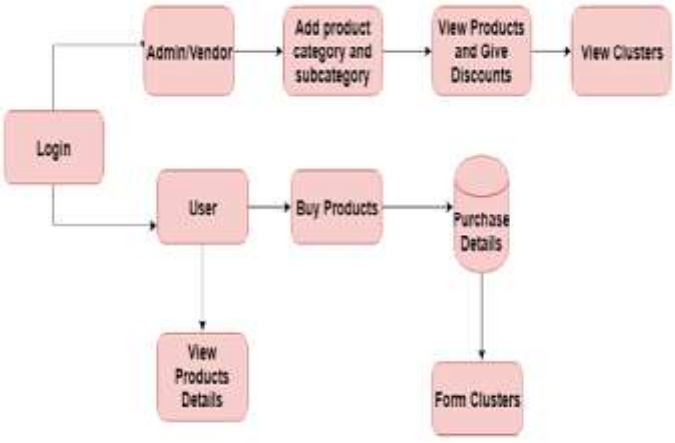

Figure 1: System Architecture for online shopping activities

3.3 Experimental Details

In this experiment we formed clusters based on product wise, season wise, and the combination of season and gender wise. We formed clusters from the transaction details of customers. The main modules are User, Admin and Vendor. In the new method we adopt k-means algorithm and build a product tree where the leaf nodes usually denote the provided products and the internal nodes are multiple

product categories. A product tree often consists of several levels and thousands of nodes, and the number of nodes increases rapidly from the top level to the bottom level. In transaction data, each product (item) bought by a customer corresponds to a leaf node in the product tree. To facilitate the analysis and visualization of customer's behaviour, we build a "personalized product tree" for each customer, called purchase tree. The purchase tree can be built by aggregating all products in a customer's transactions, and pruning the product tree by only retaining the corresponding leaf nodes and all paths from root to leaf node.

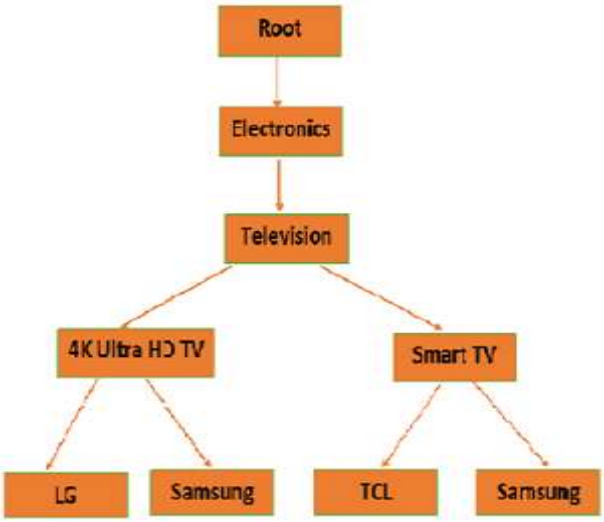

Figure 2: Product Tree

The products are often organized according to multiple types of categories. In this paper, we propose the product tree to systematically organize the products in a company.

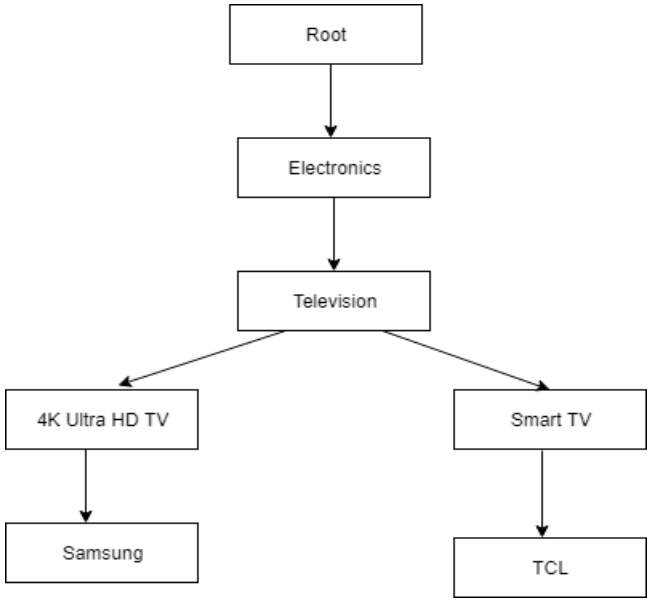

Figure 3: Purchase Tree for the figure 2.

We often created a transactional databases for this experiment. User module contains the registration, sign in, purchase products as their own needs. And Suggestions to the particular user based on the items purchased first time. For Example, If a user purchased a mobile phone, then later login he will be see the suggestions about different mobile phones. Clusters are formed based on the purchasing history details. Partitional clustering method is used for cluster formation. Mainly used K-means method for cluster formation. It formed clusters like product wise, season wise, and the combination of season and gender wise. Admin module contains some extra privilege including all privilege of visitor and user. Add products, edit product information and remove product. Ship order to user based on order placed by sending confirmation mail. See the clusters and analyze those clusters. He can understand which category of product most sold. Vendor module contains registration and login to the system. If he will add their products only his registration accepted by administrator. Vendors can also see the clusters.

K-Means Algorithm

Input: *S* (instance set), *K* (number of cluster)

Output: clusters

1. Select a point randomly from the database

   as initial centroids.

2. Choose clusters as k=3;

3. Compute similarity distance between center point and each data points.

4. Form k clusters by assigning all data points to the closest centroids.

3.4 Results and Analysis

Partition method is recommended for large data set and hierarchical method is for small data set. Performance of partition is better than hierarchical algorithms. Easy to implement. Comparatively more efficient than hierarchical method. Finally product wise, month wise, season wise and region wise clusters were formed.

4. CONCLUSION

Customer segmentation is effectively implemented by using K-means partitioning method. Segmentation implemented like product wise, season wise, region wise, combination of season wise and gender wise. Partition method is work well large databases. Segmentation provides loyalty between marketer and customers and provide customer relationship management, improves business and get more profits.

REFERENCES

[1] Saroj, T. C., and Chaudhary, T., 2015. "Study on various clustering techniques". International Journal of Computer Science and Information Technologies, 6(3),pp. 3031–3033.

[2] Tan, P.-N., Steinbach, M., and Kumar, V., 2013. "Data mining cluster analysis: basic concepts and algorithms". Introduction to data mining.

[3] Gupta, P., 2011. "Robust clustering algorithms". PhD thesis, Georgia Institute of Technology.

[4] Gupta, P., 2011. "Robust clustering algorithms". PhD thesis, Georgia Institute of Technology.

[5] Xiong, T., Wang, S., Mayers, A., and Monga, E., 2012. "Dhcc: Divisive hierarchical clustering of categorical data". Data Mining and Knowledge Discovery, 24(1), pp. 103–135.

[6] Berkhin, P., 2006. "A survey of clustering data mining techniques". In Grouping multidimensional data. Springer, pp. 25–71

[7] Everitt, B., Landau, S., Leese, M., and Stahl, D., 2001."Cluster analysis. 4th". Arnold, London.

[8] Hansen, P., and Jaumard, B., 1997. "Cluster analysis and mathematical programming". Mathematical programming, 79(1-3), pp. 191–215.

[9] Jain, A. K., and Dubes, R. C., 1988. "Algorithms for clustering data".

[10] Jain, A. K., Murty, M. N., and Flynn, P. J., 1999. "Data clustering: a review". ACM computing surveys (CSUR), 31(3), pp. 264–323.

[11] Kolatch, E., et al., 2001. "Clustering algorithms for spatial databases: A survey". PDF is available on the Web, pp. 1–22.

[12] Wilks, D. S., 2011. "Cluster analysis". In International geophysics, Vol. 100. Elsevier, pp. 603–616.

[13] Grabmeier, J., and Rudolph, A., 2002. "Techniques of cluster algorithms in data mining". Data Mining and knowledge discovery, 6(4), pp. 303–360.

[14] Migu´eis, V. L., Camanho, A. S., and e Cunha, 2012. "Customer data mining for lifestyle segmentation". Expert Systems with Applications, 39(10), pp. 9359– 9366.

[15] Ng, R. T., and Han, J., 2002. "Clarans: A method for clustering objects for spatial data mining". IEEE transactions on knowledge and data engineering, 14(5), pp. 1003–1016.

[16] Huang, J. Z., Ng, M. K., Rong, H., and Li, Z., 2005."Automated variable weighting in k-means type clustering".IEEE Transactions on Pattern Analysis andMachine Intelligence, 27(5), pp. 657–668.

[17] Tsai, C.-Y., and Chiu, C.-C., 2008. "Developing a feature weight self-adjustment mechanism for a k-means clustering algorithm". Computational statistics & data analysis, 52(10), pp. 4658 4672.

[18] Chen, X., Xu, X., Huang, J. Z., and Ye, Y., 2013. "Tw-k-means: automated two-level variable weighting clustering algorithm formultiview data". IEEE Transactions on Knowledge and Data Engineering, 25(4), pp. 932–944.

[19] Rokach, L., and Maimon, O., 2005. "Clustering methods".In Data mining and knowledge discovery handbook.Springer, pp. 321–352.

[20] Green, P. E., Kim, J., and Carmone, F. J., 1990. "Apreliminary study of optimal variable weighting in k-meansclustering". Journal of Classification, 7(2),pp. 271–285.

[21] Xu, R., and Wunsch, D., 2005. "Survey of clusteringalgorithms". IEEE Transactions on neural networks,16(3), pp. 645–678.

[22] Li, C., and Biswas, G., 2002. "Unsupervised learningwith mixed numeric and nominal data". IEEE Transactionson Knowledge & Data Engineering(4), pp. 673–690

[23] Zhang, T., Ramakrishnan, R., and Livny, M., 1996. "Birch: an efficient data clustering method for very large databases". In ACM Sigmod Record, Vol. 25, ACM, pp. 103–114.

[24] Karypis, G., Han, E.-H., and Kumar, V., 1999. "Chameleon: Hierarchical clustering using dynamicmodeling". Computer, 32(8), pp. 68–75.

[25] Karypis, G., and Kumar, V., 1998. "Multilevel k-way partitioning scheme for irregular graphs". Journal of Parallel and Distributed computing, 48(1), pp. 96–129.

[26] Guha, S., Rastogi, R., and Shim, K., 1998. "Cure: an efficient clustering algorithm for large databases". In ACM Sigmod Record, Vol. 27, ACM, pp. 73–84.

[27] Roux, M., 2015. "A comparative study of divisive hierarchical clustering algorithms". arXiv preprint arXiv:1506.08977.

[28] Izadpanah, N., 2015. "A divisive hierarchical clustering-based method for indexing image information".arXiv preprint arXiv:1503.03607.

[29] Datta, R., Joshi, D., Li, J., and Wang, J. Z., 2008. "Image retrieval: Ideas, influences, and trends of the new age". ACM Computing Surveys (Csur), 40(2), p. 5.

[30] Lew, M. S., Sebe, N., Djeraba, C., and Jain, R., 2006. "Content-based multimedia information retrieval: State of the art and challenges". ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 2(1), pp. 1–19.

[31] MacQueen, J., et al., 1967. "Some methods for classification and analysis of multivariate observations". In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, Vol. 1, Oakland, CA, USA, pp. 281–297.

[32] Reddy, M. V., Vivekananda, M., and Satish, R. "Divisive hierarchical clustering with k-means and agglomerative hierarchical clustering".

[33] Sander, J., Ester, M., Kriegel, H.-P., and Xu, X., 1998. "Density-based clustering in spatial databases: The algorithm gdbscan and its applications". Data mining and knowledge discovery, 2(2), pp. 169–194.

[34] Ben-Dor, A., and Yakhini, Z., 1999. "Clustering gene expression patterns". In Proceedings of the third annual international conference on Computational molecularbiology, ACM, pp. 33–42.

[35] Breunig, M. M., Kriegel, H.-P., Ng, R. T., and Sander, J., 2000. "Lof: identifying density-based local outliers". In ACM sigmod record, Vol. 29, ACM, pp. 93–104.

[36] M. S. Lew, N. Sebe, C. Djeraba, R. Jain, Content-based multimedia information retrieval: State of the art and challenges, ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 2 (1) (2006) 1{19.

Author Profile

Neethu C M ,post graduation student in college of engineering kidangoor. Specialization in computer and information science. Received the graduation in computer science and Engineering from Chennai University.

Anitha Abraham, She is working as an Asst.Professor at college of engineering kidangoor. She received graduation in computer science and engineering from college of engineering kidangoor and post graduation in communication and network technology from MG University. Her area of interest includes Cryptography,Image Processing, Artificial Intelligence.