International Journal of Engineering & Technology, 7 (2.32) (2018) 35-38

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Customer Data Clustering using Density based algorithm

B.Sekhar Babu, P. Lakshmi Prasanna , P.Vidyullatha

Department of Computer Science, K L University, Guntur, Andhra Pradesh, India

Abstract

This paper is about Clustering different segments of customers and their patterns of behaviour over different time intervals which are a very important application for business to maintain Business to Customer (B2C) Relationship. For clustering different segments of customers the input data will be taken from various business organizations like smart retail stores, and other stores. We take the input data from a particular amount of time like a year's data. All this data will be taken from the organization's databases. It has been observed that maintaining the old customers generate more profit when compared to attracting new ones. So, Customer retention is the important factor in our project. The main objective is to identify the elements with high profit and value to group them into different clusters. This will help to identify the high-value low-risk customers. From the results obtained, we can be able to propose some strategies to the organization that helps in retaining the old customers and improve the profits.

Keywords: *Data mining, customer clustering, density varying, high-value low-risk customers.*

1. Introduction

For powerful and beneficial business, reasonable confirmation for high benefits and generally safe clients and holding those clients and bring the accompanying level clients to above group is a key task for business visionaries and showcasing individuals [1,2]. In the base of clients, promoting individuals firstly should distinguish client bunch using a numerical strategy and afterward actualize a productive battle plan to focus on the clients that are gainful. This procedure meets with numerous issues that are extensive. The imperative part of the past studies utilized diverse logical models to section customers without considering the association between customer group and a campaign programs [3]. Due to progresses in preparing and information stock stockpiling ranges, huge associations are loading up the data of gigantic measure and the standard logical models are difficult to suspect the partitions [4]. A while later, the information that is useful is more often than not left untouched, and the more noteworthy points of interest of enhanced computational and data gathering capacities are simply recognized in halfway entirety [5]. Along the lines, utilizing information mining techniques, it is believable to separate illustrations that are significant and relationship from the customer data of the affiliation [6]. Information mining procedures like bunching and affiliation can be used to find critical cases for future expectation designs. Information mining is the methodology of isolating supportive information from given unrefined data.it is generally called learning revelation in databases. Cluster [7] is a critical information examination that tries to isolate an accumulation of articles into homogeneous gatherings which are called groups. In information mining [8] clusters are arrangement of information with the end goal that the between group closeness is minimized and intra-group similitude is augmented for the given information set.

2. Existing System

In the existing framework the grouping is performed utilizing demographic clustering calculation. Demographic grouping is dissemination based clustering calculation. clusters [9] are arranged by the esteem appropriations of their individuals. Demographic information contains numerous clear cut factors. The mining capacities functions admirably with information sets that involves this kind of factors. It is an iterative procedure over the info information [10]. Every information record is perused in a steady progression. The comparable property of every record with that of their as of now existing groups is computed. In the event that the most astounding ascertained similitude is over the given limit, the record is added to the applicable cluster. These attributes of the clusters are changed likewise. On the off chance that the assessed comparability is not over the predefined limit, or if there is no cluster (which is of the initial case) a group is made that contains the record alone. Demographic Clustering utilizes [11] the factual Condorcet measure to keep up the records task to groups and for making new clusters. The Condorcet basis valuates how each found bunch is homogeneous (in that the records it contains are comparative) and how they found groups are heterogeneous among one other. The iterative technique of discovering groups stops after two or above goes among the info information if the change of bunching result as per the predefined foundation does not demonstrate another pass [12]. The client information that needs to isolate into fragments construct shareholder esteem in light of factors as

1. Current client productivity
2. Measure of hazard
3. Measure of the lifetime esteem
4. Maintenance likelihood

Demographic clustering algorithm is utilized to group the clients in the past case. There are two stages in grouping process. In first stage, the information is gathered from the association retail bril-

liant store and after that the information purging is performed. It includes expelling the clamor. The deficient information, missing information and immaterial information are expelled and designed by required configuration [13]. In second stage, the groups are created and profiled to recognize most ideal clusters. The era of cluster is performed utilizing demographic grouping strategy with info parameters as: Regency, Add up to client benefit , Add up to client income , Best income Department. The following stride in the clustering procedure is to profile the groups by executing SQL inquiries [14]. The reason for profiling is to assess the potential business estimation of every group quantitatively. By profiling the total estimations of the shareholder are clustered by factors. The outcome got in the wake of profiling will decide the best group among all clusters shaped. Distinctive techniques are actualized in view of got estimation of the groups. The information is initially removed from the databases and level documents and changed over into level records. In this manner, the records are handled. The whole yield information set would have client data attached to the end of every record. From the got result, it is conceivable to determine some abnormal state business methodologies. The group containing, clients who have higher income per individual than of different bunches is delegated High esteem cluster. The bunch which has high income and high cost can be delegated Medium esteem. The clusters which have low income, minimal effort are named Low esteem. The clusters which have low income, high cost are delegated Negative esteem. Some conceivable systems can be incorporated are:

1. A retention or maintenance system for best clients.
2. Cross offering system can be installed between high esteem and medium esteem groups since they are close in value.
3. The cluster with low esteem all in all seems, by all accounts, to be a gathering of new clients for which the information gathered is not satisfactory to choose the practices they may appear.
4. Group with negative esteem gives off an impression of being the recognizably awful group, with a low income rate.

Limitation: The limitation with demographic clustering is more categorical variables are to be defined.

3. Proposed System

In the present paper, rather than utilizing demographic clustering algorithm we will utilize thickness based nearby thickness contrast dbscan (LDD-DBSCAN) grouping calculation. It can discover self-assertive molded bunches and it likewise decreases single connection effect.it can likewise recognize groups in differing thickness locales of high dimensional information. Dissimilar to demographic bunching the proposed calculation does not require all out factors. It requires just two parameters and for the most part cold hearted to the requesting of the databases [15]. It is impervious to commotion and can deal with groups of different shapes and sizes not at all like dbscan clustering algorithm. So we are attempting to perform client cluster utilizing LDD-DBSCAN calculation. For executing the grouping method, we are taking the guide of 'Fast digger' information mining instrument. With the assistance of this instrument we are attempting to prepare the given information set and group into different clusters as for ldd-dbscan algorithm [16].

4. Description

In the present paper, rather than utilizing demographic clustering algorithm we will utilize thickness based neighborhood thickness contrast dbscan (LDD-DBSCAN) clustering algorithm. It can discover discretionary molded groups and it additionally lessens single connection effect.it can likewise recognize bunches in changing thickness areas of high dimensional information. Not at all like demographic grouping does the proposed calculation not require absolute factors. It requires just two parameters and for the most part uncaring to the requesting of the databases. It is impervious to clamor and can deal with groups of different shapes and sizes. So we are attempting to perform client grouping utilizing LDD-DBSCAN calculation. For executing the clustering methodology [17], we are taking the guide of "Rapid Miner datamining tool". With the assistance of this tool we are attempting to handle the given information set and cluster into different groups as for ldd-dbscan calculation. The technique we take after is made out of three stages. These stages are executed in three modules that are Pre-Processing (Cleaning) [18] of information, Clustering of information and Profiling information.

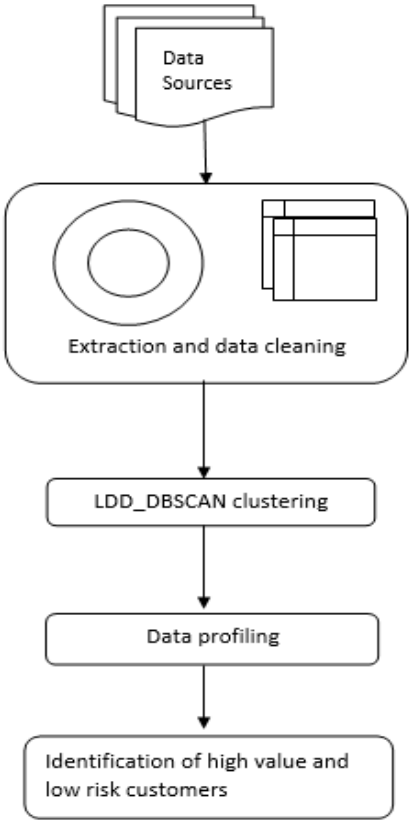


Figure 1: Clustering the Customer Data

4.1. Data Cleaning

From the data that is collected from the stores we need to perform some data processing operations on it like invalid data cleaning. Data cleaning removes the unnecessary records like outliers, missing values, and redundant data. Sometimes this irrelevant data occurs due to human typographical errors or system errors, i.e. entering the same data twice etc. So, in order to remove this irrelevant data either we remove the data or replace the missing data [19]. After cleaning the data, we need to perform the clustering to find the hidden information about the high profit low risk customers from the data.

Data cleaning majorly consists of three different tasks:

1. Fill in missing values:

1. Ignore the tuple: it is usually done when class label missing.
2. Use the attribute mean to fill in the missing value.

2. Identify outliers and smooth out noisy data:

1. Binning
2. Clustering: group values in clusters and then detect and remove outliers.
3. Regression: It is done by smoothing and fitting the data into regression functions.

3. Correct inconsistent data:

1. Use of domain knowledge or seeking expert decision.

4.2. Clustering

After the data undergoes cleaning it then is sent for the clustering operation. The data is clustered by ldd-dbscan algorithm:

Let us consider a sample data set D on which we have to perform clustering.

Algorithm:

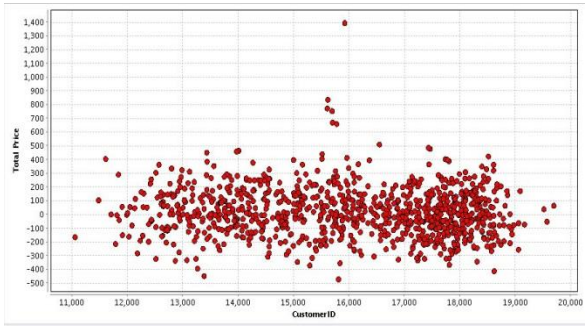
Step1: First all the objects in the dataset are unclustered.
For every unclustered object or element i.e. Ob €D
step3: If this unclustered object has minimum points in its ε radius, then it is a core point
step4: Assign a Cluster ID to this new core object Ob
step5: Put this specified object into a Queue
While Queue is not empty
Get the top element of the Queue i.e. b
step8: Calculate Relative Density,
R= {a € D | dist (b, a) <= ε}
Step9: For every object a € R
If 'a' is an unclustered object and its relative density is similar to that of core object.
Then insert the unclustered object into the Queue
If a is not known or is named as noise
Step10: Then generate a new Cluster ID to the object a
Step11: End for loop
End while loop
Else name Ob as noise
End for loop

4.3. Data Profiling

Data profiling is the process of examining the data available in an existing data source (e.g. a database or a file) and collecting statistics and information about that data. The insight gained by data profiling can be utilized to determine how difficult it will be to use existing data for different purposes. It can likewise be utilized to give measurements to evaluate data quality and figure out if or not metadata precisely portrays the source information. After the clustering process the clusters are profiled and accordingly the strategies are given [20].

5. Results and Discussions:

Java is used as a language to implement the algorithm with the help of rapid miner. The performance of the above algorithm is evaluated by using a 2-Dimensional synthetic dataset of a online retail data set. The dataset contains 1000 objects in 2-Dimensional plane. The experiments are done using different values of parameters. The below figures shows cluster formed by LDD- DBSCAN algorithm *by for the values of ε = 0.5, μ =20*. It is able to handle the density variations that exist within the cluster.



The Fig.1: data set before clustering

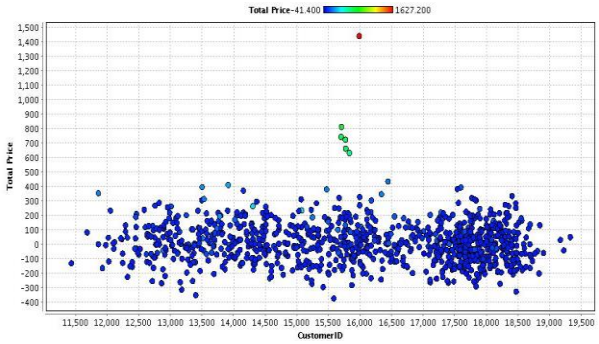


Fig.2: LDD-DBSCAN algorithm for cluster generation

6. Conclusion

We used Rapid Miner tool to segment the high profit and low risk customers who contribute for maximum profits of the company using one of the Data Mining technique called Clustering. We used retail store data in order to cluster the customers. The solution we proposed was able to cluster the high values customers even for varying densities. By finding out the high value and low risk customers we can be able to suggest different strategies to maintain the high value customers and as well move the customers of low value to high value which helps in improving the company profits.

References

[1] Dr. Sankar Rajagopal, "Clustomer data clustering using data Mining technique", International Journal of Database Management Systems. 2011 Nov 3(4).
[2] Richa Sharma, Bhawna Malik, Anant Ram, "Local Density Differ Spatial Clustering in Data Mining", International Journal of Advanced Research in Computer Science and Software Engineering.2013 Mar, 3(3).
[3] Narges Delafrooz and Elham Farzanfar, "Determining the customer lifetime value based on the benefit clustering in the Insurance industry", Indian journal of Science and Technology.2016 Jan 9(1).
[4] Chatti Subbalakshmi, G.Rama Krishna and S.Krishna Mohan Rao, "Classification of dynamic clustering based on Soft Computing. Indian journal of Science and Technology. 2015 Dec ,8(36).
[5] M. Parimala, Daphne Lopez, "K-Neighbourhood Structural Similarity Approach for Spatial Clustering", . Indian journal of Science and Technology. 2015 Sep, 8(23).
[6] K. Selvakumar, L.Sai Ramesh, A. Kannan, "Enhanced K-Means Clustering Algorithm for Evolving User Groups", Indian journal of Science and Technology. 2015 Sep ,8(24).
[7] D.Muruga Radha Devi, P.Thambidurai, "Similarity Measurement in Recent Biased Time Series Databases using Different Clustering Methods",Indian journal of Science and Technology.2014 Feb, 7(2).
[8] Zeying Li, "Research on customer segmentation in retailing based on clustering model", IEEE International Conference on Computer Science and Service system(CSSS). 2011 June 27-29.

[9] Shaily, G.L,Mehul, P. Bharot, and Darshak B.Mehta, "Web Usage Mining to discover visitor group with common behaviour using DBSCAN clustering algorithm", International Journal of Engineering and Innovative Technology.2013 Jan.
[10] Anant Ram, Sunita Jalal, Anand Jalal, and Manoj kumar, "A density based algorithm for discovering density varied clusters in spatial databases", International journal of Computer Applications.2010 Jun 3().
[11] Gan Teck Wei, Shirly K, Wahidah Hussain, and Zurinahni Zainol, "A study of customer behavior through Web Mining", Journal of Information Sciences and Computing Technologies. 2015 Feb ,2(1).
[12] I.Krishna Murthy, Data Mining- Statistics Applications, "A Key to Managerial Decision Making", indiastat.com, April-May 2010.
[13] Vidyullatha pellakuri, 2 D Rajeswara Rao, p lakshmi Prasanna "a conceptual framework for approaching predictive modeling using multivariate regression analysis vs artificial neural network", journal of theoretical and applied information technology, ISSN: 1992-8645,20th july 2015. vol.77. no.2
[14] Kim, Yong Seog, & Street, W. Nick, "An intelligent system for customer targeting: A data mining approach. Decision Support Systems", 37, 215–228: 2004.
[15] vidyullatha pellakuri, D. Rajeswara rao, "progressive decision making in the department of cardiology by optimized rough set model", int j pharm bio sci 2016 april; 7(2): (b) 658 – 665, ISSN 0975-6299
[16] A.K. Jain, M.N.Murthy, and P.J.Flyn, "Data clustering- a review" ACM Computing Surveys (CSUR).
[17] Han Jia -Wei and Micheline Kamber, "Data Mining concepts and Techniques", Higher education Press: 2001.
[18] Anusha M,V Srikanth, "Enhancement of Wireless Mesh Network using Cognitive Radio's," European Journal of Journal of Applied Sciences, Vol.7, No.3,pp.108-113,2015.
[19] P. Lakshmi Prasanna, D.Rajeswara Rao, B.Sekhar Babu,Vidyullatha Pellakuri, "Big Data for Mobile Applications In Retail Market", ARPN Journal of Engineering and Applied Sciences, ISSN 1819-6608, VOL.10, NO. 18, OCTOBER 2015.
[20] Anusha M,V Srikanth, "An Efficient MAC Protocol for reducing Channel Interference and Access Delay in Cognitive Radio Wireless Mesh Networks", International Journal on Communications Antenna and Propagation (IRECAP), ol.6, No.1, 2016.