International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)
Web Site: www.ijettcs.org Email: editor@ijettcs.org
Volume 6, Issue 5, September- October 2017 ISSN 2278-6856

Novel clustering algorithm for moderating the risk of customer churn

K. Naga Dushyanth Reddy[1],Dr.N. Kasivishwanath[2]

[1]PG student, G. Pulla Reddy Engineering College, Dept. of computer science and Engineering, Kurnool, Andhra Pradesh - 518002, INDIA

[2]Professor& Head of the Dept., G. Pulla Reddy Engineering College, Dept. of computer science and Engineering, Kurnool, Andhra Pradesh - 518002, INDIA

Abstract

As market competition surging everyday in the telecom sector, customer churn management has become an imperative for the telecom organizations to enhance their profit levels and provide better services. The conventional churn prediction models in the telecom sector don't work well while dealing with the big data. Decision makers are consistently confronted with inaccurate operations management. As a solution to these adversities, several clustering methods are proposed such as Semantic Driven Subtractive Clustering Method (SDSCM) with K-means and K-median algorithms which are failed to address processing of numerous amount of data. The proposed system is a method of implementing Semantic Driven Subtractive Clustering Method (SDSCM) with K-medoid algorithm, which is capable of processing gigantic data sets and provides 3-D Trajectories that help in efficient decision making by simulating marketing strategies to ensure profit maximization.

Keywords: Customer churn, Clustering, K-medoids, SDSCM, Map Reduce.

1. INTRODUCTION

Big data is a name given to collection of different types of data sets that are collected from and stored on clusters. It is same as the normal data but huge in size which is increasing tremendously every second. Analyzing such enormous data and processing it poses a threat for database and data analytics research and creating a void in the speed and transparency level at which the information is refined. Big data plays a crucial role, not only in IT industries but also in business organizations like banking, telecom, and various marketing sectors. Organizations are using customer value analytics for expanding their marketing profits and discharge services over the channels.

Big Data is extension of Data Driven marketing, with the accessibility and availability of information enabling rational decision making. Big Data gives greater chance to comprehend the client, with more granular data accessible and usable, driving greater result for modeling and analytics. In marketing, data analytics includes Customer Analytics that help the organization to take rational decision. Customers are adapting new technologies, and service providers are trying to meet their customers' requirements to avoid "churn". A ***churn*** is a situation where a customer is ready to switch their service provider. To avoid this churn it is imperative for every service providers to constantly engage with the latest tools, technologies and trends. The volume and velocity of the data produced by these new technologies is what drives "big data".

2. RELATED WORK

With the increasing use of smart phones by majority population, the data usage of every user is enhancing day by day. This leads the telecom industries to monitor and maintain their customers' information and provide better services to prevent the customer churn.

Ms Nisha Saini[3] has reviewed various data mining techniques to avert customer churn in the telecom industry and she observed that Exhaustive CHAID technique proved to be more capable and precise than others to anticipate the consumers that are likely to churn in nearby future.

Before the arrival of big data, the churn estimation in telecom industry was done by using data mining techniques such decision trees, neural networks, which are failed to address the processing the large data sets and high response time.

Wenjie Bi [4] proposed a method called semantic-driven subtractive clustering method (SDSCM) for better clustering accuracy. SDSCM is a combination of axiomatic fuzzy set and subtractive clustering method, which sends accurate number of clusters and cluster cancroids as the initial parameters to K-means algorithm. Hadoop MapReduce technique is implemented in parallel to lower the time complexity.

M.Rohini[2] presented a comparison of SDSCM with K-means algorithm and SDSCM with K-median algorithm. This provides a fine scenario for the service providers to decide from the results obtained from this comparison.

Aruna Bhatt[1] has analysed a modernistic technique for face recognition by performing classification of the face images using unsupervised learning approach through K-Medoids clustering. The outcome indicates that the technique is effective when compared to other clustering methods.

Taking into consideration of the above techniques, it is to be noticed that K-means and K-median are inefficient to project the customer churn in the 3D trajectories. More over these algorithms are inappropriate for processing huge data and enhances the chance of sum of dissimilarities between the data points within cluster.

As a solution to these drawbacks, this document provide a novel clustering algorithm for moderating the risk of customer churn by using SDSCM, K-medoid algorithm with MapReduce programming model in parallel to process the large data sets. The proposed work try to minimize the dissimilarities of data points within cluster and it takes a data point within the cluster as centre of the cluster and provides graphical analysis for enhanced decision making.

3. PROPOSED METHOD

AFS is an effective way for knowledge representation of fuzzy data sets using member functions, by which complex concepts can be expressed by using simple attributes of ASF. ***SCM*** is a clustering method for quickly computing number of clusters and their centroids of unprocessed data. This AFS and SCM are integrated to a single method called SDSCM collectively to provide accurate input values and parameters to the K-medoid algorithm. A serial ***SDSCM (***Semantic Driven Subtractive Clustering Method) is an integrated method of AFS and SCM, designed to express the semantic signification of fuzzy sets and for increasing the cluster accuracy. K-medoid algorithm receives the input values from the SDSCM to perform the further steps of clustering as illustrated in the Fig.1.

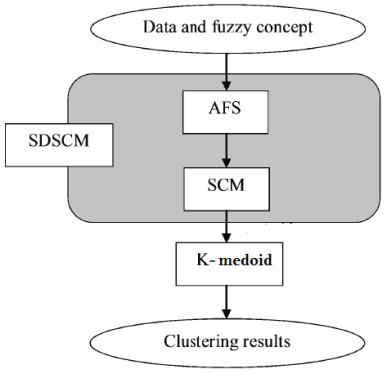


Fig. 1: Flow chart of k-medoid with serial SDSCM

Axiomatic Fuzzy Set*:* AFS uses a set of attributes based on the requirement and determines the fuzzy value. A membership function is defined to signify relationship between the elements of fuzzy set and universe of discourse.

Let X, M are two sets, where X is group of elements and M is group of attributes on X.
For example *X = {x1, x2, x3, x4, x5}* be a set of five customers and
M = {m 1, m2, m3, m4, m5} is group of attributes,

Where m1 = Peak Calls, m2 = Off-peak Calls, m3 = Weekend Calls, m4 = National Calls, m5 = International Calls. These attributes are described with sample values in the following Table.1.

Ta**ble.1:** Describing the attributes of set M

	Peak Calls	Off-peak Calls	Weekend Calls	National Calls	International calls
x_1	73	31	1	105	0
x_2	54	9	14	77	2
x_3	57	32	6	95	1
x_4	25	21	1	47	9
x_5	57	6	20	83	0

For a fuzzy concept ŋ= {m1, m2, m3} + {m4, m5}, the semantic significance of n is “persons with many calls during peak time and off-peak time and weekend” or “persons with many national and international calls” which represents “customers who are less prone to churn”. Based on the membership values of the fuzzy concept ŋ (x_i), the biggest membership value that is near to semantic concept will have less possibility to churn and a smallest value may have high possibility to churn

Subtractive Clustering Method (SCM): The supervised SCM approach is used to carry out the initialization, with utilization of purely text content. The main difference between a supervised classification and an unsupervised classification is that the class membership functions of the records in each cluster are more accurate for the case of supervised classification

K-medoid algorithm: K-medoid algorithm is a partitioned clustering algorithm which is slightly altered from K-means algorithm. These two attempts to reduce the squared-error but K-medoid algorithm is more effiecient than K-means algorithm. The algorithm proceeds in two steps:

- *BUILD-step: k* "centrally located" objects are chronologically selected, which are used as initial medics
- *SWAP-step:* If the objective function minimized by interchanging (swapping) a selected object with an unselected object, then swap is carried out. This is continued untill the objective function is no longer being reduced.

The algorithm is as follows:

1. firstly k random points are selected as the medics from n data points of the data set.
2. Associate each data point to nearest medoid by using any of the most common distance metrics.
3. For each pair of non-selected object *h* and selected object *i*, calculate the total swapping cost TCih.
4. If TCih < 0, *i* is replaced by *h*
5. Repeat the steps 2-3 until there is no change of the medics.

Four situations are considered in this process:

i. *Shift-out membership*: an object pi need to be shifted from currently considered cluster of oj to another cluster;
ii. *Update the current medoid*: New medoid oc is found to replace current medoid oj ;
iii. *No change*: Objects in the current cluster outcome have same or even smaller square error criterion(SEC) measure for all possible redistributions considered;
iv. *Shift-in membership*: Outside object pi is assigned to current cluster with the new (replaced) medoid oc.

The difference between k-means and k-medics is equivalent to the difference between mean and median: mean gives the average value of all data items collected, while median gives the value around that which all data items are uniformly distributed around it.

k-medoid clustering algorithm also supports to get the results in 3D trajectories which are used in advanced versions and gives a clear scenario for the service providers to take a favorable decision to prevent is the of churn of the customers.

Parallel K-medoid:

In K-medoids, data points in cluster are selected in the medoid and average dissimilarity to all objects within cluster is negligible. Data points and cluster centroids are taken as input and the result will be the member of the each cluster.

A serial SDSCM raises a problem of processing huge volume of information as is is only suitable for limit amount of data. For this, a parallel processing system must be appended to handle the situations in the case of processing large data sets.

Serial SDSCM is very effective for up to 100k data but as the data increases, the time taken for the Serial SDSCM increases.

Implementation of sdscm and k-medoid with Map Reduce:

Almost in every business organizations, identifying the customer churn is foremost problem of the service providers. While dealing among big data sets, it has become more a necessity than a problem. Due to the pressing need of processing and analyzing these large data sets, big data technologies provides various of processing engines and APIs that can process the data faster than the conventional system. ***Map Reduce*** is the programming framework that consists of two major functionalities called *map and reduce.* Map phase maps the input data on the given map functions and give the results to reducer. Reducer receives the values and applies the reduce function to merge the outcome of mappers. As the Map Reduce engine is composed of JobTracker and TaskTracker, job scheduling across different clusters could be efficiently done.

Psuedocode of parallel SDSCM:

Step1: calculate distance matrix
Step2: The parameter has to be determined and we determine the neighbor radius.
Step3: mountain function are initialized
Step4: mountain functions are updated
Step5: Cluster centriods are determined

In above parallel SDSCM algorithm, each step in algorithm internally performs map and reduce functions. The output values of parallel SDSCM with MapReduce are sent to K-means algorithm as the initial parameters. These clustering methods of parallel SDSCM and K-medoid are illustrated in the flow chart shown in Fig.2.

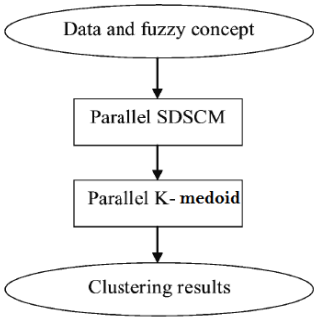
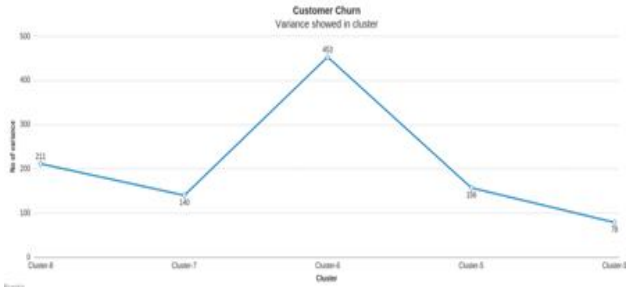


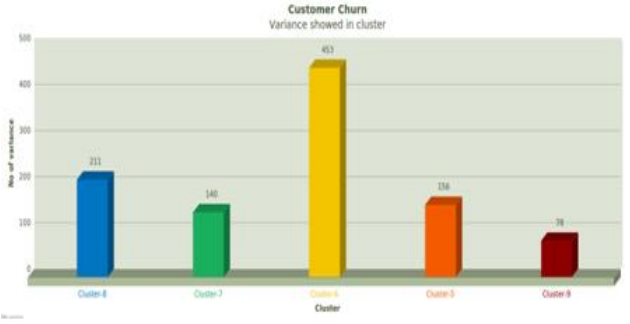
Fig. 2: Flow chart of clustering methods

4. RESULTS AND ANALYSIS

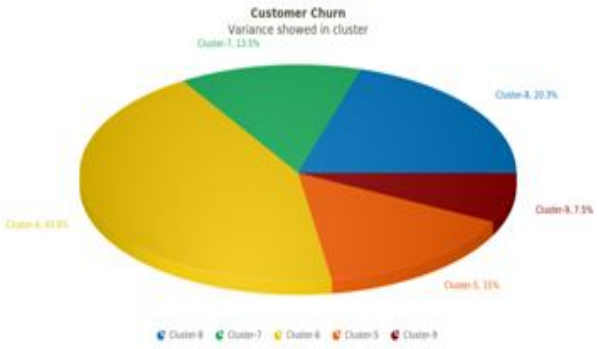
Being able to signify the semantic values from the cluster method, K-medoid defines the 3D trajectories and provides a graphical or the statistical analysis of data, which makes easy for service providers to make decisions. The following Fig.3 (a, b, c, d) shows various kinds of graphical representation of cluster data.



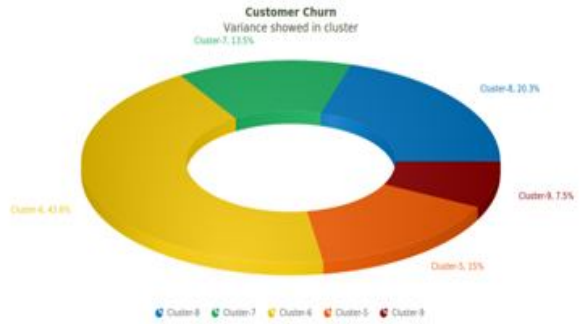
(a) Line Graph



(b) Bar Graph



(c) Pie chart



(d) D'hunt

The statistical measures shows the bench marks of serial SDSCM and parallel SDSCM as illustrated in the following Fig.4. The following analysis shows the implementation of serial SDSCM in comparison with parallel SDSCM.

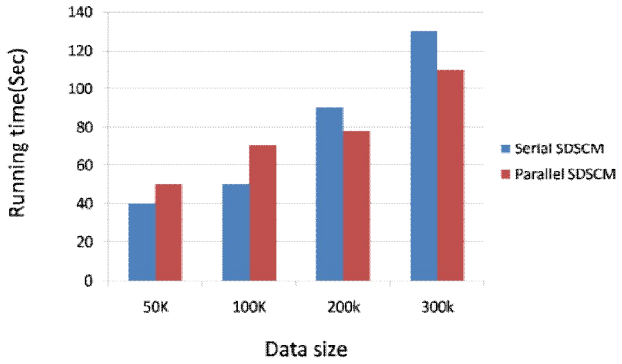


Fig.4 Comparison of serial and parallel SDSCM

5. CONCLUSION

The ever growing market simulation needs the real time analysis of all the trade and the business transactions of all the organizations. The entire study and proposed system aims at processing massive data and provides the results in form of various clusters. In the era of modernization it is essential to predict various mechanisms and strategies that are important to retain the existing customers and increase the profits. An existing clustering method called semantic-driven subtractive clustering method (SDSCM) is used along with the K-Medoid algorithm aim to provide service providers with efficient methods to avert churning customers. As many applications now a day's running big data applications using the big analytics and its technologies, the traditional techniques are no longer suitable while dealing with huge data. Therefore, implementing the efficient systems to analyze real time and distributive data sets across the different clusters by coordinating traditional clustering systems with various technologies of big data opens new doors in the era of clustering analytics.

REFERENCES

[1].Aruna Bhatt; "k-medoids clustering using partitioning around medoids for performing face recognition" International Journal of Soft Computing, Mathematics and Control (IJSCMC), Vol. 3, No. 3, August 2014.
[2].M. Rohini, P.Devaki; "Analysis Of Customer Churn By Big Data Clustering" International Journal Of Innovative Research in Computer And Communication Engineering, Vol.5 , Issue 3, March 2017
[3].Nisha Saini; "Churn Prediction In Telecommunication Industry Using Decision Tree", Streamed Info-Ocean, Vol-1, Issue-1, January-June 2016.
[4].Wenjie Bi, Meili cai, Memgqui Liu, Guo Li; "A Big Data Clusterin Algorithm For Mitigating The Risk Of Customer Churn", IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, VOL. 12, NO. 3, JUNE 2016.
[5].K-Medoid Clustering Algorithm: https://en.wikipedia.org/wiki/K-medoids
[6].Decision Tree learning Technique: https://en.wikipedia.org/wiki/Decision_tree_learning
[7].Manhattan norm https://en.wikipedia.org/wiki/Norm_(mathematics)#Taxicab_norm_or_Manhattan_norm
[8].Silhouette Clustering: https://en.wikipedia.org/wiki/Silhouette_(clustering)

K.Naga Dushyanth Reddy received the bachelor's. Degree in Computer Science and Engineering and currently pursuing Master's in Computer Science from G.Pulla Reddy Engineering College, Kurnool, A.P. His research interests include big data analytics.

Dr. N. Kasivishwanath received bachelor's Degree from Marathwada University in Computer Science & Engineering, Master's degree from BITS, Pilani in Information Systems and has received Ph.D from Rayalaseema University, Kurnool in Wireless Networks. He is currently working as Professor and Head of the Dept.of Computer Science& Engineering in G.Pulla Reddy Engineering College, Kurnool, A.P.