ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

Analysis of Customer Churn by Big Data Clustering

M.Rohini[1], P.Devaki[2]

PG Scholar, Department of Computer Science and Engineering, Kumaraguru College of Technology, Coimbatore, Tamilnadu, India[1]

Assistant Professor(SRG), Department of Computer Science and Engineering, Kumaraguru College of Technology, Coimbatore, Tamilnadu, India[2]

ABSTRACT: The interest for big data mining techniques has increased tremendously in the recent researches. Numerous classification and clustering techniques based on both supervised and unsupervised learning models were proposed. And these models had been applied in a wide range of business applications related to customer management to determine the optimal cluster for dynamic data. However, when dealing with big data in the industry, existing churn prediction models cannot work very well. In addition, decision makers are always faced with imprecise operation management. Hence the need for prediction mechanism for various strategies has become more important in large scale data source and streams. Hence, Big Data clustering algorithm called semantic-driven subtractive clustering method (SDSCM) [1] which is a combination of Axiomatic Fuzzy Sets (AFS) and Subtractive clustering method (SCM) is used to classify the customer churn rate in Telecom industry. In this work, we compareSDSCM with parallel k-mean and parallel k-median classifiers to cluster the records of the customer information or log details for maximizing the customer retention in the Telco service. The experimental result aims to predict which combination of algorithm works good based on accuracy of classifying customer in terms of loyalty.

KEYWORDS: SDSCM, AFS, SCM.

I. INTRODUCTION

Analysis of the Customer churn in the large data space for customer retention is the open research in big data classification and clustering. Customer churn refers to the loss of customers who switch from one company to another competitor within a given period. Customer churn misclassification using the supervised clustering or unsupervised clustering can lead to huge economic losses and even hurt the company's growth. Customer churn management is extremely important especially for the service industry using the parallel algorithms in the data mining for customer data clustering. Large data Analysis, large number of redundant data and noisy data has to be eliminated. Big data clustering technique has been a solution to generate the prediction of the outcome using data structures and ranking techniques which can generate as candidate solutions.

Customer Relationship Management Principles
Customer relationship management refers to strategies and technologies that is used to manage and analyze customer interactions and data throughout lifecycle of the customer in an organization, with the intension of improvising the relationships with customers in terms of business and driving sales growth. Analyzed data of the customer was collected through multiple sources, and present it to the business managers so that it will be helpful to make appropriate decisions.
Analytical Model in Customer Relationship Management

- Statistical Models

In this data is collected and scrutinize the data samples in terms of defined sets based on relation of the data. The set is formed from the different variety of data population. Model can be validated under different actions. In Statistical model the prediction result is accurate.

DOI: 10.15680/IJIRCCE.2017. 0503381

- Frequency Analysis

The analysis is carried out based on the frequency of occurrence. The occurrence is distributed into new clusters with varying frequency. The purpose of this analysis is to count the frequency of occurrence of similar events and associates the similar occurrence.

- Data mining

It is process of identifying hidden *data* and analyzing that in all means and summarizing it into useful information that can be used to increase customer retention. The data mining process caries out following step to analyze the customer data

- Data Cleaning - In this process the noise and data not in harmony with other are removed.
- Data Integration - In Data Integration multiple data sources are merged for better comparison and mining.
- Data Selection – The process is relevant to the analysis task and data are retrieved from the data sources.
- Data Transformation - In this step data are transformed into forms so that mining can be done by performing aggregation operations.

Supervised Classification

- The set of possible classes is known in advance. The input data, also called as training dataset, consists of records having multiple attributes or features.
- Each record is tagged with a class label.
- The objective of classification is to analyze the input data and to establisha precise model for each class using the features present in the data.

This model is used to classify test data for which the class descriptions are not known

Unsupervised Classification

In this data analysis process, Set of possible classes is not known. After classification name to that class is assigned. Unsupervised classification is called clustering. Unsupervised learning is closely related to the problem of density estimation in data grouping and organization of the large data in data warehouse. Unsupervised learning also comprises of many other techniques that tends to summarize and describes the key features of data. Many methods employed in unsupervised learning are based on data mining methods used to pre-process data.

- Pattern Recognition

The analysis focuses on the recognition of patterns and regularities in data. It produces the possible capabilities based on similarity in terms of expression.

- Data Correlation

It is linear Analysis of sets of data using dependent phenomena in predicting the relationship.

Big Data Processing

Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture and process the data within a tolerable elapsed time. Big data is a set of techniques to discover hidden values from large datasets that are diverse and complex in nature and of massive scale. Data growth challenges and opportunities of the rainfall data is expressed in three-dimensional. Big data requires new form of processing to enable enhanced decision making, insight discovery and process optimization.

II. LITERATURE SURVEY

Wenjie Bi, et al [1] proposed method called semantic-driven subtractive clustering method (SDSCM) for improvising the clustering accuracy. The SDSCM is implemented by Hadoop MapReduce technique in parallel, to reduce the time complexity. To get more precise operational result, SDSCM uses Axiomatic fuzzy sets (AFSs) for clustering the customer based on the usage of the services provided by the Telecom Industry.

Dheeraj Kumar, et al [2] proposed clusiVAT technique for determining the number of cluster in the given data set. This is because existing clustering algorithm suffers with the initialization problem. clusiVAT is compared with four other clustering algorithm such as *k*-means, single pass *k*-means(spkm), online *k*-means, and clusteringusing representatives(CURE). The algorithm is applied to dimension ranging from 2D to high-dimensional data set and also to real data set. Friedman test is used to compare the algorithms. The study proves that clusiVAT is fastest of other four and also found to be accurate as well. In Online k-Means, the data is partitioned into s clusters so that s centroids are obtained and hence forth is combined to form one cluster centre by weighted k-means. Whereas in SPKM the data is partitioned into s clusters as same as Online k-Means according to the number of data points that can be loaded in to the memory. CURE uses representative points to determine the cluster center. It has two data structures, heap for storing the representativepoints and k-d tree for finding the nearest neighbor. The clusters that are nearby are combined to form a big cluster till the required number of cluster is obtained. In clusiVAT, data set is partitioned into almost equal size and VAT algorithm is applied to the clusters.

Yi Wang, Qixin Chen et al. [3] proposed the clustering of electricity consumption behavior dynamics approach for analyzing electricity consumption of individual customer to enhance the load serving. Symbolic aggregate approximation is used to minimize the storage space by reducing the dimension of the data set. This is done by transforming the load curves to symbolic strings using piecewise aggregate approximation (PAA). Then to model the dynamic of electricity consumption, time-based Markov model is used. Markov model is used to predict the future consumption of electricity from the current state of use of electricity.The customers with similar behavior are classified into groups byclustering technique by Fast Search and Find of Density Peaks (CFSFDP).Kullback–Liebler distance is used to measure the distance between any two random consumption of a particular customer based on that the customer’s cluster group is decided. Big data is handled with ease as CFSFDP **uses** divide-and-conquers technique.

Jin Xiao, Xiaoyi Jiang et al. [4] proposed an approach multiple classifiers ensemble selection model based on the group method of data handling (MCES-GMDH) for analysis of the behaviors of customers about changes in their business relationships. To predict the probability of shifting from one competitor to another is great concern to the company. In the fields of telecommunication and finance, the number of customer who are likely to churn is comparatively high than in any other sectors. In past, churn prediction was handled as both supervised as well as unsupervised learning methods. Churn prediction can be defined as a supervised problem in pattern recognition perspective, if predefined forecast horizon is given. Multiple classifiers ensemble (MCE)

Q. He, K. Chang et al. [5] proposed, almost all text corpora, such as blogs, emails and RSS feeds, are a collection of text streams. The traditional vector space model (VSM) cannot capture the temporal aspect of these text streams. So far, only a few bursty features have been proposed to create text representations with temporal modeling for the text streams. Author proposed bursty feature representations that perform better than VSM on various text mining tasks, such as document retrieval, topic modelling and text categorization. For text clustering, author proposed a novel framework to generate bursty distance measure. Author evaluated it on UPGMA, Star and k-medoids clustering algorithms. The performance of bursty distance measure did not only equally well on various text collections, but it was also able to cluster the news articles related to specific events much better than other models.

Lun-Wei Ku, Yu et al. [6] proposes that discussion on the method for identifying an opinion with its holder and topic, given a sentence from online news media texts. The analysis of an approach of exploiting the semantic structure of a sentence, anchored to an opinion bearing verb or adjective. This method uses semantic role labeling as an intermediate step to label an opinion holder and topic using data from Frame Net. The utilization of the following three phases for opinion analysis: identifying an opinion-bearing word, labeling semantic roles related to the word in the sentence, and then finding the holder and the topic of the opinion word among the labeled semantic roles. For a broader coverage, the use of clustering technique is to predict the most probable frame for a word, which is not defined in earlier works.

III. PROPOSED WORK

Parallel K Median classifier is used to cluster the records of the customer information or log details for maximizing the customer retention in the particular service and organization. The feature Selection is carried out using the k Means clustering algorithm from which the training labels to the parallel k median classifier is assigned as candidate solution. The Classifier explores the relationship against each strategy which can be conditioned as criteria to classify the records of the customer log using the candidate solution provided as training labels. Churn prediction model is built on each cluster using periodic boosting technique to enhance a customer churn prediction model. It provides the solutions with respect to cost benefit analysis. In addition, it integrates the compactness within classes with the separation between classes simultaneously. Furthermore, it is possible to derive generalization bounds for these algorithms by using Eigen value analysis of the kernel matrices**.**

These algorithms allow computation of the entire path of solutions for every value of regularization parameter using the fact that their solution paths have piecewise linear form. To solve these rational equations, the rational approximation technique with quadratic convergence rate, and thus, our algorithm can follow the nonlinear path much more precisely.

FLOW DIAGRAM:

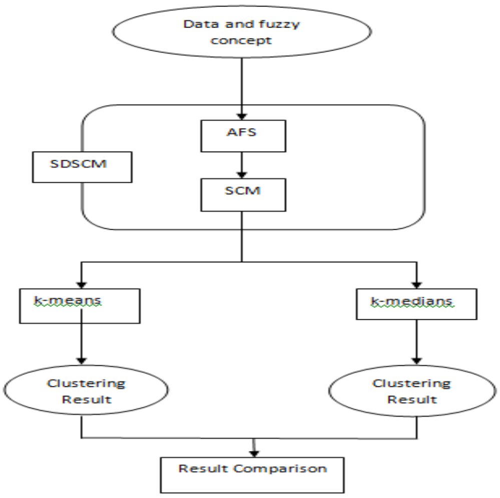

Fig 1: Flow Diagram

IV. CHURN PREDICTION MODELING

A.Computing Loyalty of Customer

Loyalty of the customer is derived from the attribute such as number, duration and charges for day, evening and night calls. Axiomatic Fuzzy Set (AFS) is an effective way to describe the fuzzy concept. ASF is used to group the customer based on loyalty. Customer with largest membership value is considered as high loyal person with less prone to churn. Thus, the assignment is always performed to a cluster with the same label, and each cluster maintains homogeneity of class distribution. Specifically, the cosine similarity function is used for assignment purposes. In each auxiliary phase, a probabilistic model is created, which relates the attribute probabilities to the cluster-membership probabilities, based on the clusters. A prediction method is applied to each sample to get base model. A bagged ensemble predicts a new sample by its base models that classifies. The final prediction of the class is normally obtained by majority voting. Each bootstrap sample is drawn by randomly generating subsets of samples where each sample is selected with replacement and equal probability.

B.Determining Initial Cluster Centroids

Subtractive clustering method is used for computing the initial cluster centroids, which belongs to unsupervised learning. It can quickly determine the number of clusters and cluster centroids based on the raw data.

The supervised SCM approach is used in order to perform the initialization, with the use of purely text content. The main difference between a supervised initialization and an unsupervised initialization is that the class memberships function of the records in each cluster are more accurate for the case of supervised initialization. Thus, the *k*-means clustering algorithm is modified, so that each cluster only contains records of a particular class. The data which provides a good starting point for the clustering process based on text content. Since the key techniques is classification for noisy attributes. For content and auxiliary information integration are in the second phase, the most of the subsequent discussion will be on the second phase of the algorithm.

C.Parallel k-median Clustering

K-median uses initial cluster centroids from SCM and update its centroids repetitively to form precise cluster through parallel processing in Hadoop. K-median minimizes the dissimilarities between points labeled in a cluster.

In other words, the probability of misclassifying samples from the negative class will be lower than the probability of error for the positive class. However, the samples are retrieved from the positive class, the impact of the priors has to be reversed. Suppose the dataset has to be taken for manipulation to the extreme and inverse the imbalance between the two classes.

The parallel k-median clustering is known to be efficient in clustering large datasets. To solve the well-known clustering problem, k-median is one of the best far-famed unsupervised learning algorithms. The Parallel k-median algorithm aims to partition the group of objects supported their attributes/features, into k clusters, where k is predefined or user-defined constant, into k clusters, where k is a predefined or user-defined constant.

D. Experimental Result

The accuracy of SDSCM with k-median is compared against SDSCM with k-means clustering. The accuracy is calculated by comparing the predicted result against the original data set. For the given dataset, the prediction accuracy of k-mean is found to be 87% and for k-median is 94%.

ARCHITECTURE DIAGRAM

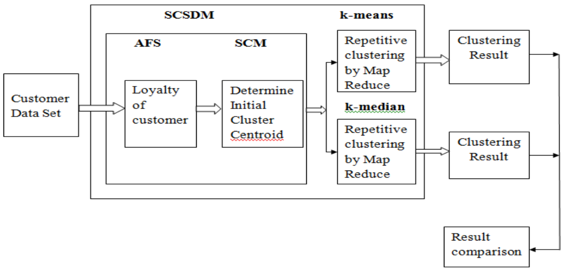

Fig 2: Architecture Diagram

V.CONCLUSION

The design and implementation of clustering paradigm using unsupervised model to cluster the significant information for inadequate data as most of the data finds to be imbalanced in the existing clusters of data records. Hence the need for prediction mechanism for various strategies has become more important in large scale data source and streams. In existing system a clustering algorithm called semantic-driven subtractive clustering method (SDSCM) is used. Experimental results indicate that SDSCM has less efficiency. Then a parallel SDSCM algorithm is implemented through a Hadoop Map Reduce framework. In the case study, the proposed parallel SDSCM algorithm enjoys a fast running speed when compared with the other methods. Parallel SDSCM along with k-mean and k-median helps by predicting the customer who are prone to churn in beforehand. Experimental analysis proves that proposed system performs the state of art approaches in terms of class distribution for each strategy against cluster accuracy and efficiency.

As a future work, marketing suggestion can be given along with the predicted result to improvise the effectiveness of the system. Results show that the marketing simulation is essential to gain maximizing profits for enterprises and to retain the valuable customers.

REFERENCES

[1] Wenjie Bi, Meili Cai, Mengqi Liu, and Guo Li," A Big Data Clustering Algorithm for Mitigating the Risk of Customer Churn", IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, VOL. 12, NO. 3, JUNE 2016.

[2].Dheeraj Kumar, James C. Bezdek , SutharshanRajasegarar, Christopher Leckie, and Timothy Craig Havens "A Hybrid Approach to Clustering in Big Data" in IEEE TRANSACTIONS ON CYBERNETICS, VOL. 46, NO. 10, OCTOBER 2016.

[3] Yi Wang, Qixin Chen, Chongqing Kang, Qing Xia “Clustering of Electricity Consumption Behavior Dynamics Toward Big Data Applications “ in IEEE TRANSACTIONS ON SMART GRID, VOL. 7, NO. 5, SEPTEMBER 2016

[4] H. Xue, S. Chen, Q. Yang, “Structural regularized support vector machine: aframework for structural large margin classifier”, IEEE Trans. Neural Netw. 22(April (4)) (2011) 573–587.

[5] G. Guo, J. Zhang, and D. Thalmann, “A simple but effective method to incorporate trusted neighbors in recommender systems,” in Proceedings of the 20th International Conference on User Modeling, Adaptation and Personalization (UMAP), 2012, pp. 114– 125.

[6] H. Ma, H. Yang, M. Lyu, and I. King, “SoRec: social recommendation using probabilistic matrix factorization,” in Proceedings of the 31st International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), 2008, pp. 931–940.

[7] A.G. Ivakhnenko, “Heuristic Self-Organization in Problems of Engineering Cybernetics,” *Automatica*, vol. 6, no. 2, 1970, pp. 207–219.

[8] A. Marqués, V. García, and J. Sánchez, “On the Suitability of Resampling Techniques for the Class Imbalance Problem in Credit Scoring,” *J. Operational Research Soc*., vol. 64, no. 7, 2013, pp. 1060–1070.

[9] Y. Huang *et al.*, “Telco churn prediction with big data,” in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, San Francisco, CA, USA, 2015, pp. 607–618.

[10]C. L. Chen and CY. Zhang, “Data-intensive applications, challenges, techniques and technologies: A survey on big data,” *Inf. Sci.*, vol. 275, pp. 314–347, Aug. 2014.

[11] H. Li, D. Wu, and G. X. Li, “Enhancing telco service quality with big data enabled churn analysis: Infrastructure, model, and deployment,” *J. Comput. Sci. Technol.*, vol. 30, no. 6, pp. 1201–1214, Nov. 2015.