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

A Survey of Data Mining Techniques for Improvement of Prediction Accuracy

Pallavi D. Bagul[1], Prof. K. C. Waghmare[2]

Student, Department of Computer Engineering, PICTPune, India[1]

Assistant Professor, Department of Computer Engineering, PICT Pune,India[2]

ABSTRACT: In today's competitive world companies' aim is to maintain their customers. The competition environment companies need to build their predictive models to identify their potential customer behaviors. Data mining techniques can be used to build the prediction model for companies because it can extract the predictive information from large databases. The accurate prediction helps in the growth of the industry. The prediction model is build by using Naive Bayes algorithm. But it is based on the independent assumptions between features. The objective of this research is to improve the accuracy of prediction by using Data mining algorithm with a Naive Bayes Classifier for better results. The proposed system is to improve the churn prediction accuracy by using data mining techniques.

KEYWORDS: Data Mining, Classification, Prediction, etc.

I. INTRODUCTION

The industry is dynamic and vibrant with large base customers. It faces a number of challenges in Data Mining because of the huge volume of data belonging to the companies. Companies use their data to make business decisions for the growth of industry and to analyze customer behavior. Before making any business decisions, business intelligence is necessary and important. Business intelligence (BI) is the set of techniques used for analyzing data and presenting actionable information. BI can predict trends, so companies need feasible BI to process their data and make decisions [1].

Churn is a term used in many companies which is mean loss of customers of the company for many reasons one of them the dissatisfaction of customer. In many companies churn term refers to customer's decision to leave the current service provider and move to other service provider [9].

- *Telecommunication industry:* Telecommunication market is rapidly growing and highly competitive. It creates a demand for data mining in order to understand the business, identify the telecommunications patterns and improve the quality of service.
- *Retail Industry:* It is the major application area of data mining because to identify customer buying behavior, achieve good customer retention. Retails data mining can help identify customer buying behavior and achieve better customer retention and satisfaction.
- *Financial data analysis:* Financial data collected in the banking sector and financial industry are often relatively complete, reliable and of high quality which facilities systematic data analysis and data mining. Loan payment prediction and customer credit policy analysis [10].

Churn occurs easily because of the strong and speedy growing competition environment in services which are providing in various sector, also churn can be happen for another reasons for examples customer's dissatisfaction with services provided by company and high cost of these services which can be in another service provider with best quality and lower cost. So churn become a concern issue in that sector because retaining of existing customer is costly than acquiring new one. Hence the need of predicting such customer we have to build a model which will give accurate prediction about customer behavior. To improve the prediction accuracy some powerful Data mining techniques are used with the Naive Bayes classifier.

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

In the competitive world in many companies, customer churn is concern issue that for the retaining existing customers is higher cost than acquiring new customers, so the company predict the customers those are having high probability to churn and understand the reasons behind this churn and try to solve it.

II. RELATED WORK

Data mining techniques are used to build the prediction model. Different algorithm is needed for prediction of the different data set. That is for different application we need to consider different algorithm.

Naive Bayes Classifier is used to classify a business data set but result is not satisfactory. And then, they used association rule to decrease features by combining related attributes. The purpose to use Apriori Algorithm is to combine related attributes by its frequent item sets. Based on two result tables and the calculated proportions, the result met our expectations. The main purpose for our research is to reduce attributes by combining related attribute to fit the independent assumption in Naive Bayes Classifier [1].

K-means method is used to develop a model to find the relationship in a customer database. Cluster analysis (K-means) find the group of persons belongs which criteria. The customer data of LIC have taken for the experiment purpose. Only the age and few premium policy such as three policies are used for the analysis. Cluster analysis using K-means to find the distance between the three customers. K-means is suitable technique for cluster analysis. It may make a good bond between the customer and insurance policy organization. This method is to find the cluster (C1) have the 3 customers (S1,S2,S10) which satisfied with all the conditions of cluster same as the S1,S2,S10 then allocated the cluster C1. Cluster C2, C3 allocated as the cluster C1.It will leads to increase the profit percentage of an organization. Some Clustering optimization method is used to find the appropriate or local optimal solution [2].

Naive Bayes Classification algorithm for customer classification and prediction on Life Insurance of customers and used Naive Bayes classification for classifying the customers from the huge data set. It also examines the challenges of using data mining technology for predicting the customer behavior. They experimented with classification technique namely Naive Bayes Classification and Data collected from IRDA Data set of Life Insurance Corporation of India. In this paper, posterior classification process applied for the data. It clearly proved that the Naive Bayes classifier is much better than other classifier to perform the policy preferences towards the customers. This technique helps us to increase the revenue of the organization [5].

To improve the customer segmentation clustering algorithm RFM (Recency, Frequency, Monetary) values are used. Then the performance of the algorithm compared with other traditional techniques such as K-means, single link and complete link.RFM is one of the very effective method for customer segmentation. For segmenting the customers, the attribute R, F and M are used as three in clustering techniques. To find the distance between from each object to all other object, here Manhattan distance used and store it in distance matrix. The parallel merging of clusters pairs improves the quality of clustering algorithm. The performance of the clustering algorithms were measured in term of four criteria (MSE, Intra cluster distance, Inter Cluster distance, Intra cluster distance divided by the inter cluster distance [6].

A method to design retail promotions, informed by product associations observed in the same groups of customers. It used the Clustering and association rule find to identify customer behavior. It can be easily predict the sales. The customer with similar purchasing behavior are first grouped by means of clustering techniques such as K-means method and for each cluster an association rule such as Apriori algorithm to identify the products that are brought together by the customers. Analysis of customer behavior aims to improve the overall performance of the enterprise [7].

The two-stage hybrid models to combine unsupervised learning technique with supervised learning technique. It developed a model for the prediction of customer churn. The important decision is the separation of churners from non-churners in customer churn management. Decision tree model are very popular in prediction of churn. It used multiple variables for clustering and examines different hybrid approaches for utilizing the results of clustering in order to build supervised learning models for prediction of churn. In the hybrid method, clustering used as a first stage and decision tree used as a second stage. C5.0 decision tree models with boosting improved the performance of models in term of top decides lift. Three customers churn data set used in this paper [3].

The standard random forests approach is developed in effectively for predicting the customer churn. In this study, proposed a improved balanced random forests method (IBRF). The experimented were conducted with the help

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

of real bank customer churn data set. The data set extracted from the banks data warehouse. IBRF proved that better prediction results among the random forests such as weighted random forests and balanced random forests. The proposed method is the combination of the two random forests such as balanced random forest and weighted random forest. IBRF is better than that of SVM, ANN and DT. This method to be proven that better accuracy, faster training speeds [4].

The standard measure to improves marketing decision making. Bayesian network classifier used for slope estimation problem of customer life cycle. In this study, they tried to acknowledge the heterogeneity in the long life customer and it is proved that possible to predict that the slope of customer life cycle of long life customers. (TAN) Tree Augmented Naive Bayes Classifiers were presented extension of Naive Bayesian classifier.To measure the performance of classifier, the performance of correctly classified used [8].

III. APPROACHES OF CLUSTERING AND CLASSIFICATION

a) Clustering Approach : The clustering is the unsupervised learning algorithms.

i. K-means : K-means clustering is an algorithm to classify or to group the different objects based on attributes or features into K number of group. Kcentroids are defined for K clusters which are generally far away from each other. Then the elements aregrouped into clusters which are nearer to the centroid of that cluster. After this first step, again the newcentroid for each cluster is calculated based on the elements of that cluster. The same method is followed,and the element is grouped based on new centroid. In every step, the centroid changes and elementsmove from one cluster to another. The same process is followed till no element is moving from onecluster to another. i.e. until we get two consecutive same steps with the same centroid andthe same elements [13].

b) Classification Approach : The classification is the supervised learning algorithms.

i. Support Vector Machines : In machine learning, support vector machines are supervised learning models with associated learningalgorithms that analyze data and recognize patterns, used for classification and regression analysis. Givena set of training examples, each marked as belonging to one of two categories, an SVM training algorithmbuilds a model that assigns new examples into one category or the other, making it a nonprobabilisticbinary linear classifier. An SVM model is a representation of the examples as points in space,mapped so that the examples of the separate categories are divided by a clear gap that is as wide aspossible. New examples are then mapped into that same space and predicted to belong to a categorybased on which side of the gap they fall on. In addition to performing linear classification, SVMs canefficiently perform a non-linear classification using what is called the kernel trick, implicitly mappingtheir inputs into high-dimensional feature spaces.

ii. **Decision Tree :** A decision tree is a decision support tool that has a flowchart like tree structure. It serves as a basis ofseveral commercial rule induction systems. The decision tree is a structure that has an internal nodewhich is also called as non-leaf node. It represents a test on an attribute. Each branch represents anoutcome of the test. Each leaf node or terminal node holds a class label. The topmost node in the tree is theroot node. Decision trees can produce binary trees as well as non-binary trees. Decision trees are widelyused because they can easily be converted to classification rules.

iii. **Naive Bayes :** The Naive Bayes' classifier is based on Bayesian theorem with independence assumptions betweenpredictors. This model is easy to build, with no complicated iterative parameter estimation which makesit particularly useful for very large datasets. Despite its simplicity, the Naive Bayesian classifier oftendoes surprisingly well and is widely used because it often outperforms more sophisticated classificationmethods. Bayes theorem provides a way of calculating the posterior probability, P(c|x), from P(c), P(x),and P(x|c). Naive Bayes classifier assume that the effect of the value of a predictor (x) on a given class (c)is independent of the values of other predictors. This assumption is called class conditionalindependence.

$$P(C = c|X = x) = \frac{P(C = c)\ \Pi_i P\ (X_i\ |C = c)}{P(X = x)}$$

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

P(c|x) is the posterior probability of class (target) given predictor (attribute).

P(c) is the prior probability of class.

P(x|c) is the likelihood which is the probability of predictor given class.

P(x) is the prior probability of predictor.

The posterior probability can be calculated by first, constructing a frequency table for each attributeagainst the target. Then, transforming the frequency tables to likelihood tables and finally use the NaiveBayesian equation to calculate the posterior probability for each class. The class with the highestposterior probability is the outcome of prediction [13].

IV. CONCLUSION

Customer churn is a big issue in many companies because it happen easily under light of strong competition in this business area, so the company need to build a churn prediction model that identify churner and non-churner customer and avoid the churn. The churn prediction accuracy is improved by using Data mining algorithm with Naive Bayes classifier algorithm. The result of the churn prediction model is more accurate than the result of Naive Bayes classifier.

ACKNOWLEDGEMENT

We sincerely thank our Dissertation Coordinator Dr. A. S. Ghotkar and Head of Department Dr. R. B. Ingle for their support providing all the help, motivation and encouragement for completion of this work.

REFERENCES

[1] Tianda Yang, Kai Qian, Dan Chia-Tien Lo, Ying Xie and Yong Shi, Lixin Tao, "Improve the Prediction Accuracy of Nave Bayes Classifier with Association Rule Mining", IEEE 2nd International Conference on Big Data Security on Cloud, IEEE, pp. 129-133, 2016.

[2] Narander Kumar, Vishal Verma, Vipin Saxena, Cluster Analysis in Data Mining using K-Means Method, International Journal of Computer Applications , Vol. 76, No. 12, pp. 11-14, August 2013.

[3] Indranil Bose, Xi Chen, Hybrid models using Unsupervised Clustering for Prediction of Customer Churn, Proceedings of the International MultiConference of Engineers and Computer Scientists, Vol. 1, March 18-20, 2009, Hong Kong.

[4] Yaya Xie, Xiu Li, E.W.T. Ngai, Weiyan Ying, Customer churn prediction using improved balanced random forests, An International Journal of Expert System with Applications , Vol . 36, pp. 5445-5449, 2009.

[5] S. Balaji, S.K. Srinivasta, Nave Bayes Classification approach for Mining Life insurance Databases for Effective Prediction of Customer Preferences over Life Insurance Products, International Journal of Computer Applications, Vol.51, No. 3,2012.

[6] Prabha Dhandayudam, Dr.Illango Krishnamurthi , An improved Clustering Algorithm for customer segmentation, International Journal of Engineering Science and Technology, Vol. 4, No. 2, pp.99-102, Feburary 2012.

[7] P. Issakki Alias Devi, S.P. Rajagopalan, Analysis of Customer Behavior using Clustering and Association Rules, International Journal of Computer Applications, Vol. 43, No.23, pp.19-27, April 2012.

[8] Bart Baesens, Geert Verstraeten, Dirk Van den Poel, Michael Egmont Petersen, Patrick Van Kenhove, Jan Vanthienen, Bayesian network classifiers for identifying the slope of the customer lifecyele of long life customers, European Journal of Operational Research, Vol. 156, pp. 508-523, 2004.

[9] Lina Ahmed Mohammed Nour Ali, Dr. Atif Ali, "Implementation of Naive Bayes algorithm for building churn prediction model for telecommunication company", University of Science and Technology, pp. 4-27, December 2014.

[10] S. Janakiraman, K. Umamaheswari, "A Survey on Data Mining Techniques for Customer Relationship Management", International Association of Scientific Innovation and Research (IASIR), pp. 55-61, 2014.

[11] Amjad Hudaib, Reham Dannoun, Osama Harfoushi, Ruba Obiedat, Hossam Faris, "Hybrid Data Mining Models for Predicting Customer Churn", February 2015.

[12] R. Margaret, Business Intelligence (BI) Definition. http://searchdatamanagement.techtarget.com/Definition/ business-intelligence, Nov. 04, 2015.

[13] Aishwarya Churi, Mayuri Divekar,Sonal Dashpute,Prajakta Kamble,Reena Mahe,"Prediction Of Customer Churn In Mobile Industry Using Probabilistic Classifiers",International Journal of Advance Foundation And Research In Science & Engineering (IJAFRSE)Volume 1, Issue 10 , March 2015.

BIOGRAPHY

Pallavi D. Bagul is a Student, Department of Computer Engineering, Pune Iinstitute of Computer Technology Pune. Her research interest in Data Mining.

Prof. K. C. Waghmare is a Assistant Professor of Department of Computer Engineering, Pune Iinstitute of Computer Technology Pune, Her research interest in Data Mining, Data Structure and Design and analysis of algorithm.