IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 6, June 2017

Improvement in Prediction Accuracy using Data Mining Techniques

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

Student, Department of Computer Engineering, PICT Pune[1]

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

Abstract: In today's competitive world companies' aim is to maintain their customers. In the competitive 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 paper is to improve the accuracy of prediction by using Data mining algorithm with a Naive Bayes Classier for better results. The proposed system gives the accuracy of 72.62% by using data mining techniques on UCI machine learning Repository which is a bank data set.

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]. Churn occurs easily because of the strong and speedy growing competition environment in services which are providing in various sectors such as Telecommunication industry, Retail industry and Financial data analysis 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. In today's 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.

DOI10.17148/IJARCCE.2017.66104

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].

III. PROPOSED METHODOLOGY

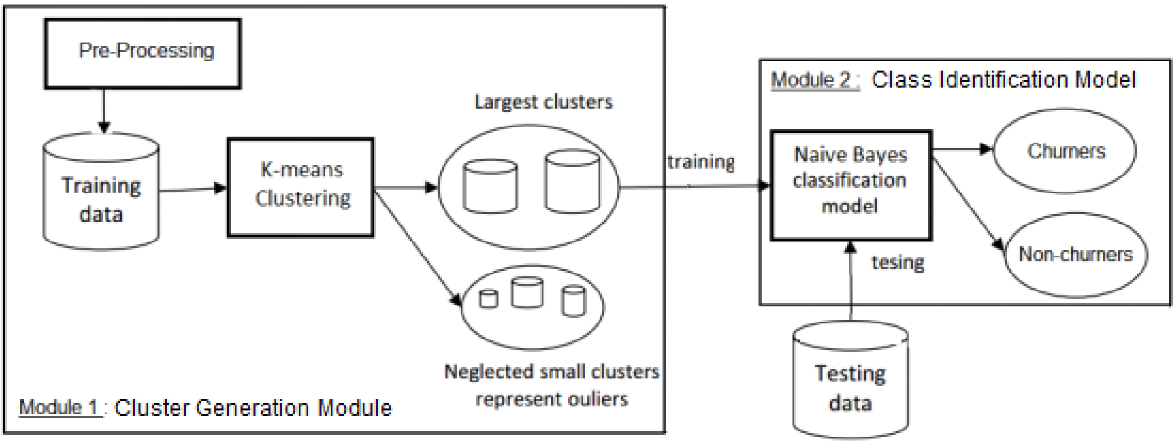

Pre-Processing
Training data
K-means Clustering
Largest clusters
Neglected small clusters represent ouliers
Module 1 : Cluster Generation Module
training
Module 2 : Class Identification Model
Naive Bayes classification model
Churners
Non-churners
tesing
Testing data

Figure 1: Architecture Diagram

Figure 1 shows architectural design of proposed system. Following are important components in the system :

- The proposed system has two main modules :
  1. Cluster Generation module
  2. Churn Class Identification module

The input to the system is the bank customer records and the file should be in .csv format. The modules are explained in detail as follows:

A. Cluster Generation module :

This module first performs the pre-processing of the given customer records. The detailed description of this module is as follows. This module contains following functions.

DOI10.17148/IJARCCE.2017.66104

Pre-processing: In this step, the input training data are first pre-processed. The loaded data set is the preprocessed to remove numbers as in select only numeric data and removing id column. The pre-processing of the data records include it may be necessary to convert the data set into the standard spreadsheet format and convert categorical attributes to binary. It may also be necessary to normalize the values of attributes that are measured on substantially different scales.

Clustering: Then Simple Data Mining algorithm uses Euclidean distance measure to compute distances between instances and clusters. Cluster centroids are the mean vectors for each cluster (so, each dimension value in the centroid represents the mean value for that dimension in the cluster). Thus, centroids can be used to characterize the clusters by using K-means Clustering.

B. Churn Class Identification module :

The new data set is generated by adding Cluster attribute to the original data set. In the data set, each instance now has its assigned cluster as the last attribute value. After that the new data set is as training data to the classification model then the model will classified into two categories as churner or non-churner. Now, we use this classification model to classify the test data into one of the stated two classes. Probability is obtained as follows:

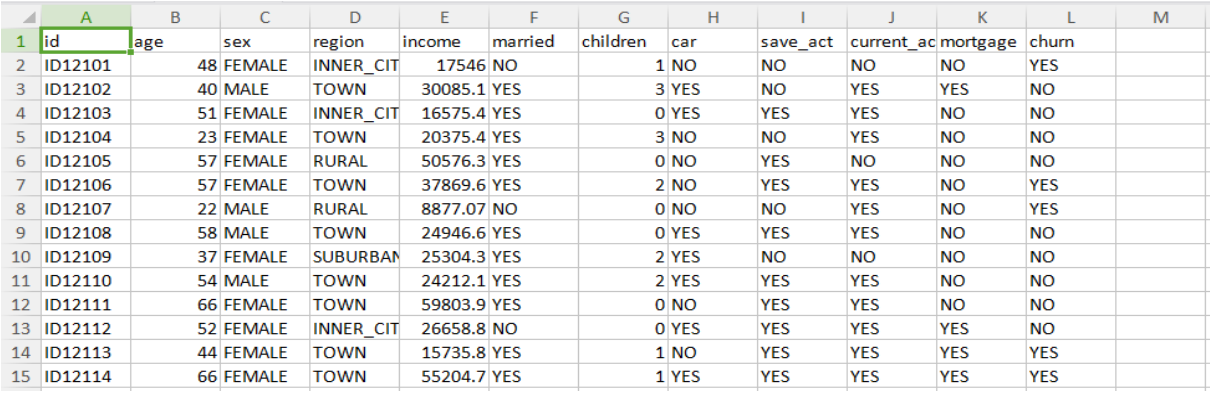
$$P(K_i|X) = P(x_1|K_i)P(x_2|K_i)\ldots P(x_n|K_i) / P(X)$$

Where, K_i represents the category of classes and X is the data record. X can be divided into pieces of instances, say x_1, x_2 ... x_n which are relative to the attributes X_1, X_2 ... X_n, respectively.

IV. EXPERIMENT AND RESULTS

A. Data set

The input given to the system is the bank customer records the business data set we used is provided by UCI Machine Learning Repository which is a bank data set[1]. The purpose to this work is to predict the churn prediction of customer and improve the accuracy of existing system based on the given banking information.



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	id	age	sex	region	income	married	children	car	save_act	current_ac	mortgage	churn	
2	ID12101	48	FEMALE	INNER_CIT	17546	NO	1	NO	NO	NO	NO	YES	
3	ID12102	40	MALE	TOWN	30085.1	YES	3	YES	NO	YES	YES	NO	
4	ID12103	51	FEMALE	INNER_CIT	16575.4	YES	0	YES	YES	YES	NO	NO	
5	ID12104	23	FEMALE	TOWN	20375.4	YES	3	NO	NO	YES	NO	NO	
6	ID12105	57	FEMALE	RURAL	50576.3	YES	0	NO	YES	NO	NO	NO	
7	ID12106	57	FEMALE	TOWN	37869.6	YES	2	NO	YES	YES	NO	YES	
8	ID12107	22	MALE	RURAL	8877.07	NO	0	NO	NO	YES	NO	YES	
9	ID12108	58	MALE	TOWN	24946.6	YES	0	YES	YES	YES	NO	NO	
10	ID12109	37	FEMALE	SUBURBAN	25304.3	YES	2	YES	NO	NO	NO	NO	
11	ID12110	54	MALE	TOWN	24212.1	YES	2	YES	YES	YES	NO	NO	
12	ID12111	66	FEMALE	TOWN	59803.9	YES	0	NO	YES	YES	NO	NO	
13	ID12112	52	FEMALE	INNER_CIT	26658.8	NO	0	YES	YES	YES	YES	NO	
14	ID12113	44	FEMALE	TOWN	15735.8	YES	1	NO	YES	YES	YES	YES	
15	ID12114	66	FEMALE	TOWN	55204.7	YES	1	YES	YES	YES	YES	YES	

Figure 2: Data set

The training data set contains 11 input attributes and 421 instances (i.e. 70% of original data), the testing data set contains 11 input attributes and 179 instances (i.e. 30% of original data).

B. Results and Discussion

The Prediction table of the proposed system is shown in following table 1:

Classes	Actual Value	Predicted Values
Churner	274	271
Non-churner	326	329

Table 1: Proposed System Prediction Table

The Accuracy of the proposed system is 72.62%.

The proposed system is compared with the following system using the performance evaluation parameters:

- Naive Bayes
- Decision tree

The Prediction table of the Naive Bayes Algorithm is shown in table 2.

Classes	Actual Values	Predicted Values
Churner	274	254
Non-churner	326	346

Table 2: Naive Bayes Prediction Table

The Accuracy of the Naive Bayes Algorithm is 58.65%.
The Prediction table of the Decision Tree Algorithm is shown in table 3.

Classes	Actual Values	Predicted Values
Churner	274	255
Non-churner	326	345

Table 3: Decision Tree Prediction Table

The Accuracy of the Decision Tree Algorithm 56.98%.

The clustering approach is use to generate the clusters of the given input data set into two clusters as shown in the following plot. The plot of the given test data set is predicted using the Naive Bayes algorithm which is shown in the figure below. First, we generate the clustering of similar customers based on the polarity using the Naive Bayes algorithm. From the total of 179 testing data set, the Naive Bayes has classified 100 as non-churner, 79 as churner.

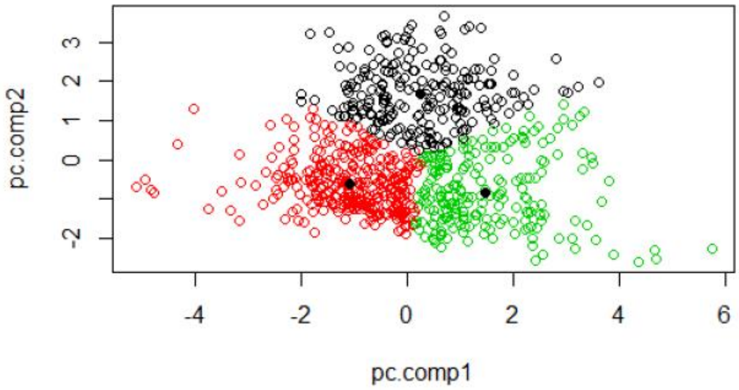


Figure 3: Clustering by Proposed Approach

The Naive Bayes algorithm is used for classifying the customers into two basic churn classes. The results obtained from the algorithm is shown in the following figure 4.

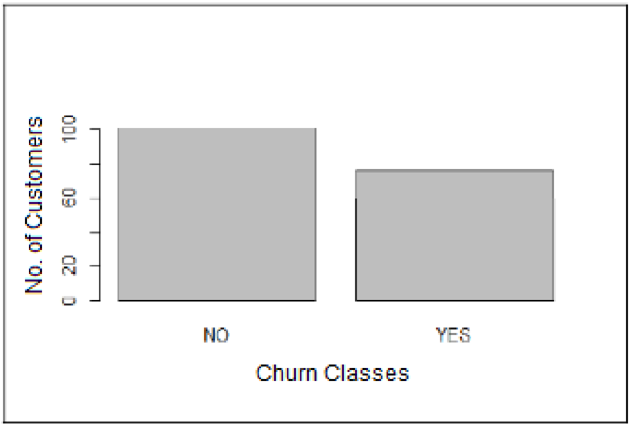
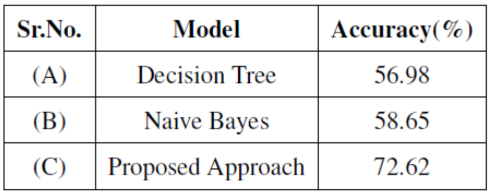


Figure 4: Classification by Proposed Approach

DOI10.17148/IJARCCE.2017.66104

C. Comparative Analysis

The proposed approach is compared with the other two machine learning algorithms based on the performance evaluation parameters. The Table below shows the classification of customer records of same data set into two churn classes using different algorithms along with the proposed approach.



Sr.No.	Model	Accuracy(%)
(A)	Decision Tree	56.98
(B)	Naive Bayes	58.65
(C)	Proposed Approach	72.62

The above table shows the performance measure with the different models. It is shown from the results that the proposed approach gives the better results than the other machine learning algorithm. This is because the proposed approach uses the hybrid approach of churn prediction approaches such as K-means clustering approach and then based on the maximum probability class out of the two defined churn classes the final result class is predicted.

IV. CONCLUSION

The result of the churn prediction model which is the combination of cluster generation and churn classification approach improve the prediction result as compared to Naive Bayes classifier.The proposed approach gives the accuracy of 72.62 % tested on the 179 instances of the testing data set. The classification result of proposed system gives high accuracy than the other two algorithms.

ACKNOWLEDGMENT

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 Naive Bayes Classifier with Association Rule Mining", IEEE 2nd International Conference on Big Data Security on Cloud, IEEE, 2016, pp. 129-133.

[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, August 2013, pp. 11-14.

[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.

[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, 2009, pp. 5445-5449.

[5] S. Balaji, S.K. Srinivasta, "Naive 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, Feburary 2012, pp. 99-102.

[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, April 2012, pp.19-27.

[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, 2004, pp. 508-523.

[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, December 2014, pp. 4-27.

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

[11] Aishwarya Churi, Mayuri Divekar and Sonal Dashpute, "Prediction Of Customer Churn In Mobile Industry Using Probablistic Classifiers", International Journal of Advance Foundation And Research In Science Engineering (IJAFRSE) Vol. 1, Issue 10 , March 2015, pp.41-49.

BIOGRAPHIES

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.