Performance Evaluation of Churn Customer Behavior based on Hybrid Algorithm

Riddhima Rikhi Sharma
Research Scholar, Guru Gobind singh college of modern technology, Kharar

Rajan Sachdeva
Assistant Professor, CSE Dept., Guru Gobind Singh college of modern technology, Kharar

ABSTRACT

Various algorithms of Data Mining have been used for making distinguish between customers into loyal and churn. Boosting algorithms are iterative studying process that will combines poor classifiers as a way to create a powerful a classifiers. SVM is utilized for segmentation associated with churn clients. This paper represents the proposed Hybrid approach is an integration of two techniques named random forest and Support Vector Machine(SVM) that have feature of Artificial bee colony (ABC), provides better and accurate results in the prediction of churn customers.

Keywords

SVM; Customer churn behavior; artificial bee colony algorithm; Churn customers .

1. INTRODUCTION

Data Mining helps to create model by extracting useful information from Database. The model helps to define patterns and relationships. DM activities are divided into three general categories.

Discovery- It is process to know about hidden patterns without any detailed information about the patterns.

Predictive Modeling- this type of modeling is used to make patterns from the large database in order to make future predictions.**Forensic Analysis-** this type of DM helps to extract patterns to find unknown data elements.

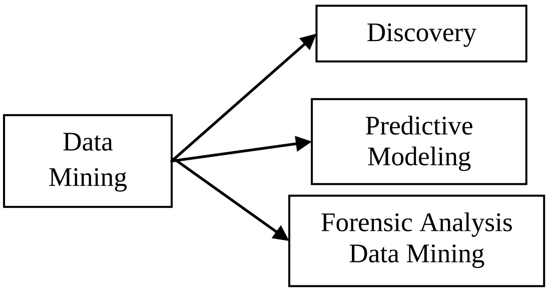

Fig 1: DM Categorization

1.1 Data mining in CRM

1.1.1 Association

It helps to make relationship among data, taken from various records. Statics and apriority algorithms are the tools for the association discovery. Market Basket analysis is best one application used by association rules. These rules are basically related to if-else statement to uncover relationship in information system. It is used to analyze the data with support and confidence.

1.1.2 Classification

it is used to classify the customers according to the group or priority to build a prediction model for data mining. NN, DT is the type of classifications. Classification is used to predict the group membership. Classification is process of identifying the new observations.

1.1.3 Clustering

clustering refers to the make clusters or segmentation of a task into number of similar types or clusters. Clusters are built by user and are not pre-defined. NN and Discrimination analysis are clustering types. Clustering is often a process regarding partitioning some data (or objects) into some meaningful sub-classes, named clusters. Help consumers understand the natural group or structure in a very data fixed. Used either as a stand-alone tool for getting insight straight into data submission or as a preprocessing phase for other algorithms

1.1.4 Forecasting

Forecasting is related to results or future values of data. It belongs to logical and modeling relationships, for example Demand forecasting. Forecasting may be the process of earning predictions for the future based about past along with present data and analysis of trends. Forecasting would be the process of getting predictions of the future based with past along with present facts and examination of styles.

1.1.5 Regression

It is basically related to mapping a data object to real value to provide prediction. Linear Regression and Logistic Regression are the types of regression analysis. Regression investigation is popular for prediction and forecasting, where it is use provides substantial overlap with the field associated with machine finding out. Regression analysis is usually used to understand which one of many independent parameters is associated with the centered variable, in order to explore the types of these associations. In confined circumstances, regression analysis can be used to infer causal relationships involving the independent and dependent parameters.

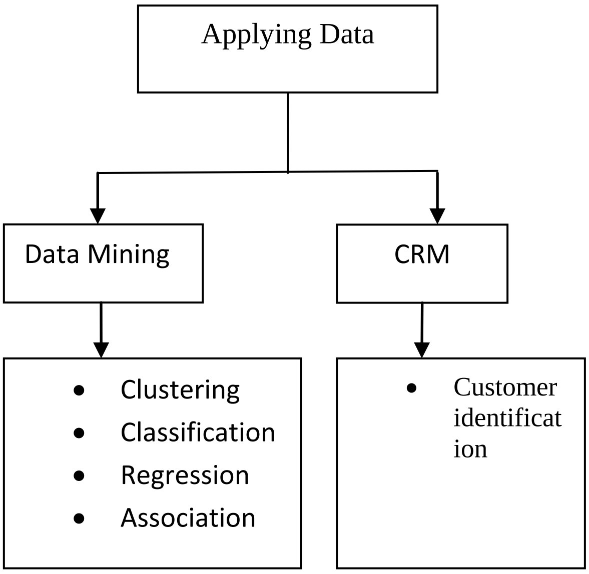

Fig 2: CRM in DM

2. CHURN CUSTOMERS

Customers are the root of any organization to change the level of company at another level. Customer who are tending to move from one organization to another, are known as churn customers. Churn is loss or detection of clients from one company to another. Churn management makes useful strategy in order to retain customers in the company as churn customers are to predicted, for examples by providing offers and better services. Customers can be grouped into two categories –loyal and churn customers.

Loyal customers are the customers who visit to company at regular interval. Churn customer's leads to the loss of company as they are moving from one company to another, where they found some extra profit.

2.1 Factors of customer churn behavior

Customer churn behavior can be predicted from the four factors: client behavior, client perceptions, and client demographics and macro-environments.

2.1.1 Client Behavior

refers to the treatment by the clients. These also relates to the main components that are utilized by the company and how the customers are dealing with them

2.1.2 Client Perception

it relates to the attachment of the customers to the company. It is basically relates to the activation and deactivation of the relation with the company.

2.1.3 Client demographic

includes personal information of customers, also used for churn calculation.

2.1.4 Macro environment

variable belongs to the changes available in the world and the different views of the customers about the product qualities.

3. SVM ALGORITHM

Support vector machines (SVMs, also support vector networks) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary 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 as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.

4. ARTIFICIAL BEE COLONY ALGORITHM

ABC algorithm is on the premise of the smart technique for the honey bees getting alongside each other. Most loved honey bees being social creepy crawlies split their perform among themselves: Used honey bees, Onlooker honey bees and Look Bees. Their measures are ordered specifically into four main times: Initialization point, used honey bee point, Onlooker honey bee point and Look honey bee stage. In introduction point, each utilized honey bee is furnished with different nourishment assets. In utilized honey bee point, each utilized honey bee figures the nectar measure of the foodstuff source associated with it and the separation of it from the hive. In pursuit honey bees point, the utilized honey bees whose nourishment source gets to be overlooked gets to be hunt honey bee. The primary component work of hunt honey bees is to discover new nourishment assets. Mostly of pc study and reason study, ABC is primarily utilized for answer of advancement issue. At the point when identified with improvement issue, the foodstuff potential outcomes would be the pair of shifted conceivable arrangements accessible.

4.1 Advantages of ABC

- Ease, freedom and robustness.
- Usage of fewer control variables compared a lot of different search techniques.
- Easy hybridization with different optimization algorithms.
- Volume to handle the point price with stochastic nature.
- Easy implementation with simple mathematical and affordable operations.

5. RELATED WORK

K.Kaur and S.Vashist (2015) enhanced the CRISP-DM methodology based on RFM and SVM by using Hybrid approach to observe churn customers on the dataset of retail store. CRISP-DM is made up of six parts and prediction model is developed by using five different stages. **A.churi and R.Mahe (2015)** discussed DM as powerful tool for churn prediction and explained statistical tool to predict customer churn by developing early-warning model with the use of DT, C4.5, Bayesian classifier and Naïve Bayes as a probabilistic classifier on the telecom dataset. Classification and Association rules are used as DM in CRM..**M. R. Israil and M. Makhtar (2015)** implemented MLP (Multilayer Perceptron) neural network on the telecommunication dataset to predict churn customers. MLP is compared with Regression analysis and Logistic Regression analysis and found as good to provide better results. MLP algorithm also provides statistical predictive approaches for churn prediction. For regression analysis, multiple and regression analysis was analyzed by them. The research design consists of different phases named Data Preprocessing and normalization, The drawback that arises with MLP is its complex structure as it consisted of fourteen inputs, one hidden and one output node units. **Y.Liu and Y.Zhang (2015)** described the concept of customer segmentation and misclassification cost for churn prediction. Customer segmentation is a way to distinct the clients into different groups on their behavior, preferences and demands. Decision tree is used for building model for high performance. Comparisons are made with C5.0, LR and ANN in order to make model for churn prediction. It helps to recognize the customer churn as well as maintain the customer strategy. Only prediction of churn clients is not appropriate for business profit, there must be efficiency improvement in customer management too.**A.O. Oyeniyi and AB Adeyemo (2015)** used K-mean clustering with rule-based algorithm in making predictions for churn customers. JRip was used for rule based algorithm. The raw data was cleaned and mined with Data mining tool named weka. The algorithm is applied on dataset of Nigeria bank and helps to identify churn customers so that appropriate solutions are to be made to retain the churn customers in company. K-means clustering technique and repeated incremental pruning to develop error reduction, which is also called JRip algorithm.

6. GAPS IN LITERATURE

It is observed that Churn Prediction has been major research problem with the growth of market development as customers asset more valuable persons for growth of company. The occurrence of churn customers is one of the crucial problems for the growth of a company, as it acquires higher costs. The task of churn prediction is to identify the customers who are pretending to shift from one company to another. As in the competitive environment, it becomes necessary to focus on retaining churn customers as well as attracting new customers and also the use of unsupervised filtering is ignored to reduce the effect of noise in the data, the use of evolutionary optimization is also neglected by the majority of existing researchers, the benefits of artificial bee colony like its speed and global best properties still need to be explored to predict churn users. We will propose the hybrid approach is an integration of two techniques named random forest and Support Vector Machine (SVM) that have feature of artificial bee colony (ABC), provides better and accurate results in the prediction of churn customers.

7. EXPERIMENTATION AND RESULTS

7.1 Methodology

1. Load 'churn Arff' data file into the memory.
2. Apply unsupervised filtering on target attribute and convert it from numeric to nominal.
3. Now split options are come in action to filter the data into 30/70 format i.e. 30 data as training and 70 data for testing.
4. Now set required constants.
5. Initialize random bees and global best and local best values.
6. Repeat the steps 6 to 10 while(loop<=max_it).
7. Evaluate velocity
8. Check best bees.
9. Now call meta algorithm i.e. SVM
10. Now update SVM values
11. Now define attribute types
12. Now evaluate nominal attribute values
13. Now define instances
14. classify test instances
15. Now call random forest Algorithm
16. Classify test instances
17. Now evaluate final outcomes.

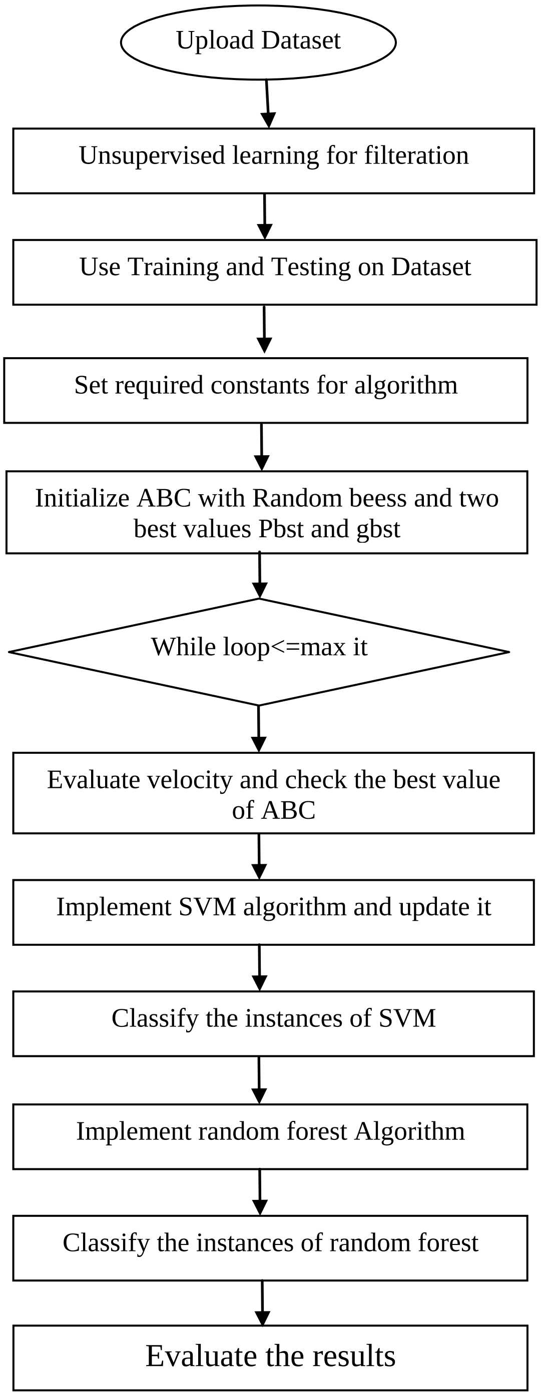
Upload Dataset
Unsupervised learning for filteration
Use Training and Testing on Dataset
Set required constants for algorithm
Initialize ABC with Random beess and two best values Pbst and gbst
While loop<=max it
Evaluate velocity and check the best value of ABC
Implement SVM algorithm and update it
Classify the instances of SVM
Implement random forest Algorithm
Classify the instances of random forest
Evaluate the results

Fig3: Flowchart of Purposed Hybrid approach

7.2 Performance analysis

This paper has designed and implemented the proposed technique in MATLAB tool u2013a. The evaluation of proposed technique is done on the basis of following metrics i.e. Accuracy, F-measure, true positive rate, kappa statistics, error rate.A comparison is drawn between all the parameters with existing and proposed algorithm and figures shows all the results.

7.2.1 KAPPA Statistic, MAE, RMSE

Kappa coefficient is often a statistic which often measures inter-rater understanding for qualitative (categorical) objects. Root Mean Square Error (RMSE) can be determined as measurement of the difference between two distinct values that are being observed or calculated. Mean Absolute Error (MAE) is used to appraise that how close the anticipations are to the actual values.Table 1, it has been clearly indicated that proposed Hybrid approach provides the better Kappa statistic against all others, when comparisons are made in between them. Kappa statistics must be high and errors using RMSE and MAE are less. By using these values, Histogram is produced , which shows the Highest value of Kappa statistics and low range of errors that proved the proposed Hybrid approach gives improved results.

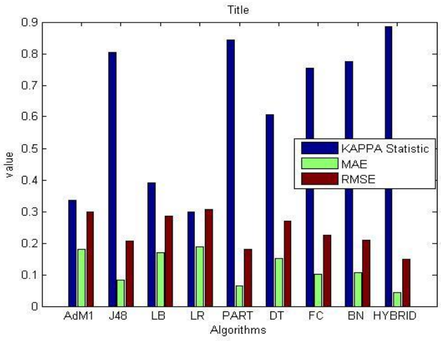

Fig4: Performance analysis using Kappa statistics, MAE and RMSE between existing and purposed algorithms

Table no. 1

Algorithm Name	**Kappa Statistics(**	**Mean Absolute Error(M**	**Root Mea Square Error(RM**
AdaBoostM1	0.3357	0.1815	0.2999
random forest	0.804	0.0821	0.2059
SVM	0.391	0.1701	0.2855
Logistic Regression	0.2976	0.1879	0.3068
PART	0.8426	0.0646	0.1798
Decision Table	0.6076	0.1505	0.2702
Filtered Classifier	0.7531	0.1004	0.2241
Bayes Network Classifier	0.7738	0.1065	0.2104
Purposed Hybrid approach	0.8845	0.0445	0.1492

7.2.2 TP Rate, FP Rate, Precision

TPR refers to True Positive Rate. It is also called Sensitivity or Recall in some fields. TPR is defined as measurement of positive cases that are correctly identified.

FPR is called False Positive Rate. It is defined as ration of those instances or objects that are incorrectly identified as positive. It is also known as fall-out.

Precision is defined as measurement of all positive cases that are identified when making calculations. Precision is also known as positive predictive value.

Table 2, it has been clearly indicated that proposed Hybrid approach provides the better, results. Recall, F-Measure and ROC provides the highest values in the result when we are making comparisons between existing and purposed algorithm.

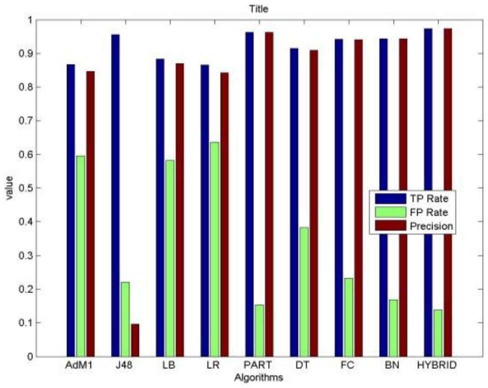

Fig5: Performance analysis using TP, FP and Precision between existing and purposed algorithms

Table 2

Algorithm Name	**TP Rate**	**FP rate**	**Precision**
AdaBoostM1	0.867	0.595	0.846
random forest	0.956	0.221	0.0955
SVM	0.884	0.582	0.87
Logistic Regression	0.866	0.637	0.842
PART	0.962	0.153	0.962
Decision Table	0.915	0.382	0.909
Filtered Classifier	0.942	0.233	0.94
Bayes Network Classifier	0.944	0.169	0.944
Purposed Hybrid approach	0.973	0.1390	0.973

7.2.3 Recall, F-measure, ROC

Recall is the division of the written documents that are applicable to the question which have been winningly recovered.

F-Measure is also called F1 score. It contains both precision and recall. It is generally use to check the accuracy and reliability.

ROC (Receiver Operating Characteristics) is a sort of graphical plot that demonstrates the execution of a binary classifier technique process as its discrimination limit is different.

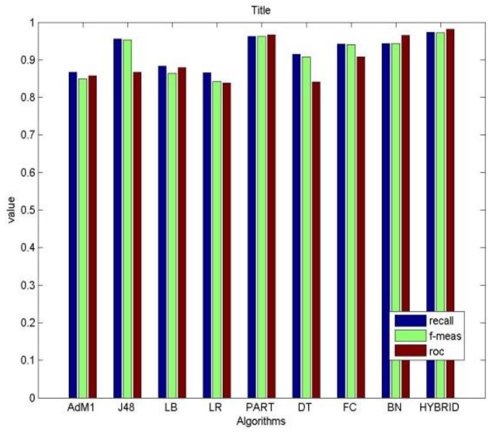

Fig6: .Performance measure using Recall, F-measure and ROC between existing and purposed algorithm

Algorithm Name	Recall	F-Measure	ROC
AdaBoostM1	0.867	0.849	0.857
random forest	0.956	0.953	0.867
SVM	0.884	0.864	0.879
Logistic Regression	0.866	0.843	0.838
PART	0.962	0.962	0.967
Decision Table	0.915	0.908	0.841
Filtered Classifier	0.942	0.941	0.908
Bayes Network Classifier	0.944	0.944	0.965
Purposed Hybrid approach	0.973	0.972	0.981

8. CONCLUSION AND FUTURE WORK

The occurrence of churn customers puts adverse impact on the profit of company various algorithms based on Data Mining and Hybrid approach of random forest and SVM provides better and accurate results when comparisons are made on various algorithms. The use of ABC with two best values local best and global best makes it more effective to obtain the more pleasing and comfortable results. Different kinds of parameters used in order to enhance the purposed algorithm. The purposed Hybrid approach is implemented in MATLAB with statistics toolbox. The experiment results reveal that purposed Hybrid approach outperforms better results when comparisons are made over eight different algorithms. Present Hybrid approach is implemented in MATLAB, but in future it can be incorporated to multiple software platforms to achieve better results as well as Variety of different software tools can be used for further analysis.Different algorithms can be used in prediction of churn customers to obtain more accurate results.

9. REFERENCES

[1] Adnan Idris,Asifullah Khan and Yeon Soo Lee(2012) ," Genetic Programming and Adaboosting based churn prediction for Telecom", Korean National Research Foundation, COEX, Seoul, Korea.

[2] Afaq Alam Khan, Sanjay Jamwal and M.M.Sepehri (2010), "Applying Data Mining to Customer Churn Prediction in an Internet Service Provider," Vol. 9, No.7, Pp.8-14.

[3] Fröhlich, Ahn, P. Hana, and S. Lee (2006), "Customer churn analysis: Churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry, Telecommunications Policy", International Journal of Computer Applications, Vol. 30(10-11), pp. 552-568

[4] A.Churi, M. Divekar and Reena Mahe(2015) , " Prediction Of Customer Churn In Mobile Industry Using Probabilistic Classifiers, " International Journal of Advance Foundation And Research In Science & Engineering, Vol. 1, No. 10,Pp. 41-49

[5] Alejandro Correa Bahnsen, Djamila Aouada and Bjorn Ottersten (2015), "A novel cost-sensitive frameworks for customer churn predictive modeling, " Decision Analytics a SpringerOpen Journal, Pp.1-15

[6] Alok Kumar Rai and Medha Srivastava(2012) , " Customer Loyalty Attributes: A Perspective, " NMIMS Management Review, Vol. 22, Pp.49-76

[7] Amal M. Almana, Mehmet Sabih Aksoy and Rasheed Alzahrani(2014) , "A Survey On Data Mining Techniques In Customer Churn Analysis For Telecom Industry," Int. Journal of Engineering Research and Applications, Vol. 4, No.5, Pp.165-171

[8] A. Hudaib, R. Dannoun, O. Harfoushi, R. Obiedat and H. Faris(2015), " Hybrid Data Mining Models for Predicting Customer Churn " Int. J. Communications, Network and System Sciences, Vol. 8, Pp. 91-96

[9] Anuj Sharma and Dr. Prabin Kumar Panigrah(2011), "A Neural Network based Approach for Predicting Customer Churn in Cellular Network Services", International Journal of Computer Applications, Vol. 27, No.11,pp. 26-31

[10] A. O. Oyeniyi and A.B. Adeyemo(2015), " Customer Churn Analysis In Banking Sector Using Data Mining Techniques, " African Journal of Computing & ICT, Vol 8. No. 3, Pp.165-174

[11] Bart Baesens, Geert Verstraeten, Dirk Van den Poel, Michael Egmont (2004), " Bayesian network classifiers for identifying the slope of the customer lifecycle of long life customers", European Journal of Operational Research, Vol. 156, Pp. 508-523

[12] Burez J. and Van D. (2008), "Separating Financial from Commercial Customer Churn: A Modeling Step towards Resolving the Conflict between the Sales and Credit

Department," Expert Systems with Applications, Vol. 35, Issue 1, pp. 497-514

[13] Chih-Fong Tsai and Mao-Yuan Chen (2009), "Variable selection by association rules for customer churn prediction of multimedia on demand," Expert Systems with Applications, Vol.30, Pp. 1-10

[14] Chris Rygielski , Jyun-Cheng Wang and David C. Yen(2002) , " Data mining techniques for customer relationship management, " Technology in Society, Vol. 24, Pp. 483–502

[15] Dr. M.Balasubramanian, M.Selvarani(2014), "Churn prediction in mobile telecom system", International Journal of Scientific and Research Publications", Vol. 4, No. 4, Pp.1-5

[16] Dr. U. Devi Prasad and S. Madhavi (2012)," Prediction Of Churn Behavior Of Bank Customers Using Data Mining Tools, " Business Intelligence Journal, Vol.5, No.1, Pp. 96-101

[17] H. Abbasimehr, M. Setak, and M. Tarokh (2014), "A Comparative Assessment of the Performance of Ensemble Learning in Customer Churn Prediction," The International Arab Journal of Information Technology, Vol. 11, No. 6, Pp. 599-606

[18] Emtiyaz, Mohammad Reza Keyvanpour(2011), "Customers Behavior Modeling by Semi-Supervised Learning in Customer Relationship management ", Advances in information sciences and Service Science, Vol.3, No. 9, Pp. 229-236

[19] E. Shaaban, Y. Helmy, A. Khedr and M. Nasr(2012) , " A Proposed Churn Prediction Model, " International Journal of Engineering Research and Applications, Vol. 2, No. 4, Pp.693-697

[20] Georges D. Olle Olle and Shuqin Cai(2014), "A Hybrid Churn Prediction Model in Mobile Telecommunication Industry, " International Journal of e-Education, e-Business, e-Management and e-Learning, Vol. 4, No. 1,Pp.55-62