ISSN: 2277-9655
(I2OR), Publication Impact Factor: 3.785

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

HYBRID APPROACH OF BOOSTED TREE FOR CHURN PREDICTION IN MATLAB

Navneet Kaur *, Naseeb Singh
* M.Tech Student, Department of IT, AIET, Faridkot, Punjab, India
Asst. Professor, Department of CSE, AIET, Faridkot, Punjab, India

DOI: 10.5281/zenodo.46979

ABSTRACT

Organization puts much attempt to hold the churn clients in the company by identifying them as clients are beneficial persons to the growth of a company. Hybrid approach of Boosted tree is one of advance algorithm in data mining for the prediction of churn clients. The Hybrid approach of Boosted tree involved J48 and LogitBoost with some parameters. This Hybrid approach uses the feature of PSO that helps to predict the churn clients with more accuracy. This paper enhances the boosted tree with hybrid approach and produces more accurate results in the prediction of churn clients. The results obtained with Hybrid approach shows that the new approach is better in predicting the churn clients from existing boosted tree technique.

KEYWORDS: Churn clients, PSO, Data Mining, CRM, Boosted Tree, Hybrid Algorithm.

INTRODUCTION

Data mining is an approach for making patterns from large datasets. Data Mining is also interlinked with "KDD" to find and extract useful data [1]. Data mining techniques are used by many companies in order to make predictions of churn and loyal clients from the given dataset. Various DM techniques such as DT,NN, SVM, RF and LR etc. have been used in recent years in order to predict customers into loyal and churn categories [2]. CRISP-DM is a DM model that is made up of six different parts [3]. Data mining techniques have been used for the distinction of churn and loyal customers from the customers database so that proactive measures would be taken in future in order to retain them in the company as customers are god for them. Clients are the valuable persons for the profit of company. In now days, companies have moved their interest from product to clients and becomes client oriented to product oriented [4]. Churn clients becomes a major problem in the growth of an organization. It has been assumed that cost of attracting new clients is always greater than retaining existing customers. So organizations put more efforts to retain the churn clients after identifying them with the help of DM Techniques. There exist many variables for the prediction of churn customers. hurn clients can be grouped into voluntary and involuntary churners; where further voluntary churn can be divided into incident and deliberate churn [5].Churn is basically related to switching off the contract from one company. The managing of churn customers by the management is called Churn management. Churn clients can be handled by two approaches –target approach and untargeted approach [6]. Clients become churn clients and move to another company due to several reasons such as Dissatisfaction, quality, lack of features, No loyal brand etc [7]. CRM is a strategy to handle customers by providing retention methods. CRM always focus on loyal customers as company profit depends upon these loyal customers [8].

The rest of paper is organized as- Section 2 describes the related work in which different research papers have been read for my research. Section 3 is related to the Research problem from the base paper. Section 4 discusses the objectives of the research. Section 5 is about Methodology used for the research. Section 6 represents the results and at last section conclusion and future scope is represented.

RELATED WORK

There are many algorithms developed in recent years for the prediction of churn customers by using various Data Mining techniques. But these algorithms help in predicting churn customers with using fewer factors while making

decisions. Hence many of the algorithms have been enhanced with newer approaches. Some of related work has been discussed as-

In this paper [**K.Kaur and S.Vashist (2015)],** author enhanced the CRISP-DM methodology based on RFM and Boosted Tree by using Hybrid approach in order to predict churn customers on the dataset of retail store. CRISP-DM consists of six phases and prediction model is developed by using five different stages. The results are compared by AdaBoostM1 with Decision Stump and AODE and are proved as a better result than existing boosting technique. Hybrid approach is used in order to build classifiers and is an enhancement of boosted tree algorithm. The hybrid algorithm is implemented in cloud environment.

In the paper **[M. Lapczynski (2014**)] author developed a hybrid model of C& RT- logit model by integration of decision tree and logistic model. Hybrid model produce improved results than basic logistic model as it used decision tree with it. It provider better results when compared to single DT. The hybrid approach also helps to obtain different probabilities of each test case. The limitation that arises during research is that Hybrid C& RT -logit can be applicable only to Single Decision tree.

In the paper **[R .Obiedat and O. Harfoushi (2013)] author** implemented Hybrid approach of K-mean clustering and Genetic Programming to predict churn customers. K-mean clustering is used to filter the dataset and Genetic Programming helps to classify the customers into churners and non-churners. Four clusters are to be used out of two clusters are discarded. Selected classifiers are loaded into model and results are compared with C4.5, ANN and GP with accuracy and churn rate. The accuracy rate does not classify exact churn and loyal customers, which is main limitation of this hybrid approach.

In the paper **[P. Dhandayudam et al. (2012)**] author, used RFM (Recency, Frequency and Monetary) values to enhance the clustering algorithm. The enhanced algorithm was compared with K-means, single link and complete link, which were traditional techniques. Cluster technique was used for the segmentation of customers.RFM were used in three techniques of clusters. Real dataset of customer transaction details were used for making clustering. Use of parallel merging is compared to traditional clustering. Clustering technique used for making segmentation of customers' into different groups. Clustering performance was measured in MSE; inter cluster distance, intra cluster distance and intra cluster distance divided by inter cluster distance.

In the paper **[Xie and Xiu(2009)],** author developed IBRF (improved Balanced Random Forest) on the dataset of real bank customer with ANN, DT and CWC-SVM and produce better results over RF algorithm such as BRF and WRF as it combines sampling technique. This novel approach is proposed by combining RF which used cost sensitive and Weighted Random forest which used sampling technique. This model is developed to put heavier effect on the misclassification of the minority class, which reduces the error case and achieve more and effective results.

PROBLEM FORMULATION

Boosted tree is used for making prediction of churn and loyal clients, but during the implementation more weights are to be assigned to each object until the classifier is updated into strong one. Building strong classifier with use of number of classifiers and iterations takes long time in making decisions, which is main limitation of boosted tree.

METHODOLOGY

The Purposed Hybrid algorithm is integrated with logitBoost and J48 tree. It contains the feature of PSO. PSO is particle swarm optimization that is used for searching best solution with two best values –pbst and gbst. Pbst is generally personal best and gbst is global best. PSO begins with providing initial values to all particles and then search for best solution. With two best values pbst (best solution) and gbst (best value), particle update its velocity and position. Maximum iterations are applied in order to update the best solution for pbst and gbst. At each iteration, each particle observe that best value and solution for themselves and then neighbors. The iterations are applied until we find best fitness value. The hybrid approach algorithm of Boosted tree for the prediction of churn clients is as follows:-

Step 1: Load the Dataset and apply training and testing on it.

Step 2: Initialize the components of PSO i.e. Velocity, Position, pbst, gbst, c1, c2, and random numbers.
Step 3: Repeat for the maximum iterations.
Step 4: Evaluate Velocity and position of PSO by using equations (1) and (2) to get the best fitness value.

v_nt=V(i,j)+R1*C1*(pbst(i,j) X_pso(i,j))+R2*C2*(gbst(1,j)-X_pso(i,j)); (1)

Xinter=X_pso(i,j)+V(i,j); (2)

Step 5: Call the LogitBoost and update it with pbst and gbst.
Step 6: Now define the attribute type, relationship name, size and instances of given dataset.
Update the attribute type and instances with PSO.
Step 7: Call the J48 and classify the test instances.
Step 8: Update the results.

RESULTS AND DISCUSSION

Hybrid approach of Boosted tree (J48 +LogitBoost) improves the results in term of accuracy in the prThe results are implemented in MATLAB on Intel core i3 with 4GB RAM on 32 bit operating system. The results are obtained by comparing with 8 different algorithms with different parameters.

Table 1: Kappa, MAE and RMSE

Algorithm	Kappa	MAE	RMSE
AdM1	0.3357	0.1815	0.2999
J48	0.804	0.0821	0.2059
LB	0.391	0.1701	0.2855
LR	0.2976	0.1879	0.3068
PART	0.8426	0.0646	0.1798
DT	0.6076	0.1505	0.2702
FC	0.7531	0.1004	0.2241
BN	0.7738	0.1065	0.2104
Hybrid approach	0.8845	0.0445	0.1492

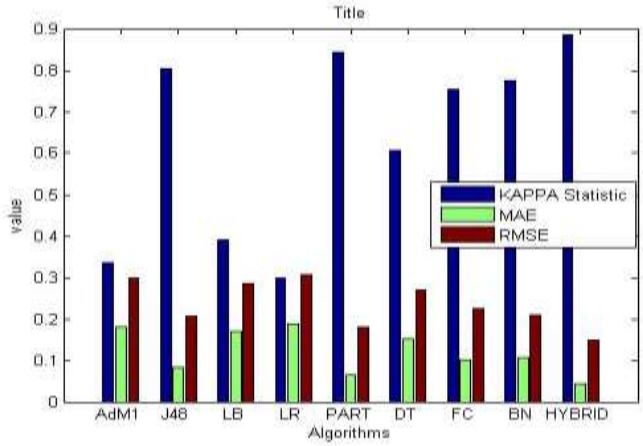

Fig1: Performance analysis with Kappa, MAE and RMSE

The table 1 and Fig. 1 depict higher Kappa statics and less error of MAE and RMSE. Higher Kappa statics shows the improvement taken by Hybrid approach of boosted tree.

Table 2: TP Rate, FP Rate and Precision

Algorithm	TP Rate	FP rate	Precision
AdM1	0.867	0.595	0.846
J48	0.956	0.221	0.0955
LB	0.884	0.582	0.87
LR	0.866	0.637	0.842
PART	0.962	0.153	0.962
DT	0.915	0.382	0.909
FC	0.942	0.233	0.94
BN	0.944	0.169	0.944
Hybrid approach	0.973	0.1390	0.973

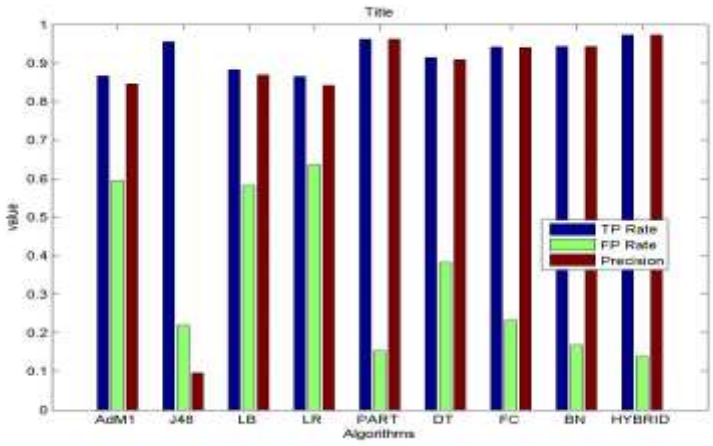

Fig 2: Performance analysis with TP rate, FP rate and Precision

Table 2 and Fig. 2 represent the higher TP rate and precision. TP rate and Precision are closely linked with each other and depict that hybrid approach shows the better enhancement among different algorithms.

Table 3: Recall, F-measure and ROC

Algorithm	Recall	F-Measure	ROC
AdM1	0.867	0.849	0.857
J48	0.956	0.953	0.867
LB	0.884	0.864	0.879
LR	0.866	0.843	0.838
PART	0.962	0.962	0.967
DT	0.915	0.908	0.841
FC	0.942	0.941	0.908
BN	0.944	0.944	0.965
Hybrid approach	0.973	0.972	0.981

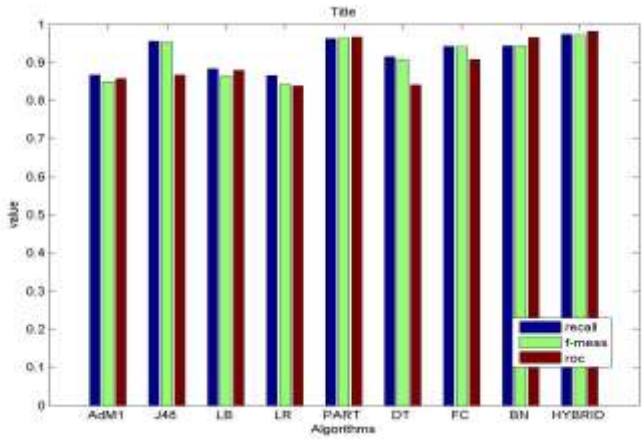

Fig 3: Performance analysis with Recall, F-measure and ROC

Table 3 and Fig. 3 represent the Recall, F-measure and ROC. Higher values of Recall, F-measure and ROC show the improvement taken by Hybrid approach of Boosted tree over 8 different data mining algorithms.

Table 4: Accuracy, Error rate and RAE

Algorithm	Accuracy	Error rate	RAE
AdM1	86.6787%	13.3213%	73.1989%
J48	95.5596%	4.4404%	33.0881%
LB	88.3588%	11.6412%	68.59%
LR	86.5887%	13.4113%	75.7602%
PART	96.2496%	3.7504%	26.0667%
DT	91.4791%	8.5209%	60.6888%
FC	94.2394%	5.7606%	40.4962%
BN	94.3894%	5.6106%	42.9417%
Hybrid approach	97.2997%	2.7003%	17.9552%

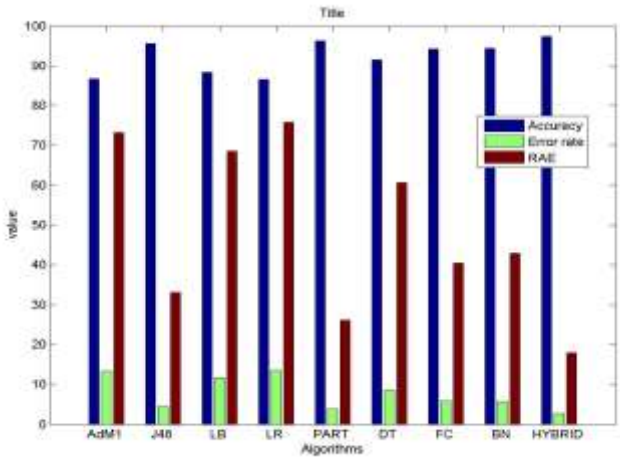

Fig4: Performance analysis with Accuracy, Error rate and RAE

Table 4 and Fig. 4 shows the higher accuracy produced by Hybrid approach of boosted trees which shows the enhanced results taken by it when comparisons are made over different algorithms.

Table 5: RRSE, Coverage of cases and Mean rel. region size

Algorithm	RRSE	Coverage	Mean rel.
AdM1	85.20%	99.4899%	91.2091%
J48	58.50%	95.5596%	50%
LB	81.09%	98.7999%	76.2976%
LR	87.15%	99.0099%	81.0681%
PART	51.07%	99.0099%	60.5911%
DT	76.75%	98.9449%	85.2835%
FC	63.65%	98.1998%	63.9664%
BN	59.77%	99.4899%	68.5119%
Hybrid approach	42.38%	98.9499%	54.4554%

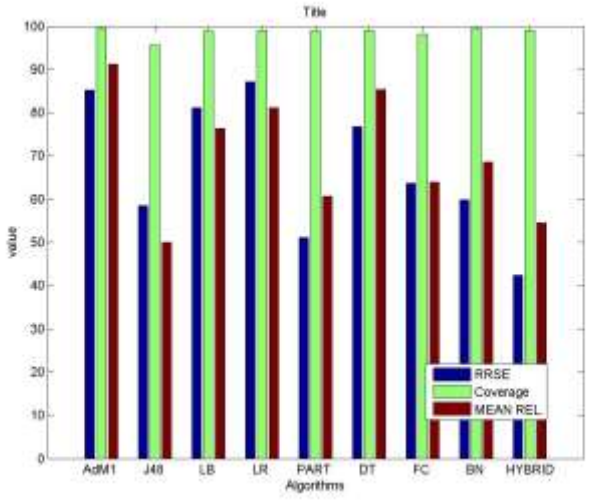

Fig 5: Performance analysis with RRSE, coverage of cases and Mean Rel.

It has been shown from the above Table 5 and Figure5 that Purposed hybrid approach provides the better results by producing less RRS errors. For the predictions and to make accurate results; coverage of cases increases in order to enhance the performance by making positive and correct predictions.

CONCLUSION AND FUTURE SCOPE

Every organization puts much effort in predicting the accurate churn clients by using different data mining, so that proactive measures could be attained in future for holding the churn clients. In this research paper, we compare 8 different algorithms of boosted tree with new hybrid approach of boosted tree and conclude the result that all algorithms contain number of imperfections. The new hybrid approach of boosted tree overcomes this problem. The proposed Hybrid approach uses the feature of PSO for making prediction of churn and loyal clients of given dataset that contains 3333 instances and 21 attributes. Different Kinds of Parameters are used for making accurate predictions. It has been shown from graphs (Fig.1 to Fig. 5) that hybrid approach provides best results over churn prediction.

The purposed hybrid approach provides the accuracy of 97.29:% when comparisons are made over 8 different algorithms. In Future, a new algorithm can be used in new simulator for making predictions of churn and loyal clients.

REFERENCES

[1] S. Janakiraman and K. Umamaheswari(2013),” *A Survey on Data Mining Techniques for Customer Relationship Management*,” International Journal of Engineering, Business and Enterprise Applications, Vol.7, No.1, pp.55-61

[2] Dr. M. Madan and K. Nijhawan(2015), “*A Review on: Data Mining for Telecom Customer Churn Management*,” Vol. 5, No. 9,pp. 813-817

[3] A. M. Almana 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

[4] Nabavi, S. and Jafari, S.(2013), “ *Providing a Customer Churn Prediction Model using Random Forest Technique*” , In proceedings of 5th IEEE-Conference on Information and Knowledge Technology (IKT), pp. 202-207

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

[6] 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

[7] 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

[8] A. Sharma and Dr.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

[9] Kiranjot Kaur and Sheveta Vashisht, " *A Novel Approach for Providing the Customer Churn Prediction Model using Enhanced Boosted Trees Technique in Cloud Computing*, " Vol. 114 , No. 7 Pp.1-7

[10] M.Lapczynski(2014) , "*Hybrid C&Rt-Logit Models In Churn Analysis*, " Folia Oeconomica Stetinensia, Pp.37-52

[11] R.Obiedat, M. Alkasassbeh, H. Faris and O. Harfoushi(2013) , " *Customer churn prediction using a hybrid genetic programming approach*”, Vol. 8, No. 27, Pp. 1289-1295

[12] P.Dhandayudam, Dr. Krishnamurthi (2012) , “*An improved Clustering Algorithm for customer segmentation*”, International Journal of Engineering Science and Technology, Vol. 4, No. 2, Pp. 99-102

[13] Yaya Xie , Xiu Li , E.W.T. Ngai and Weiyun Ying (2009), " *Customer churn prediction using improved balanced random forests*, "Expert Systems with Applications. Vol. 36 , Pp. 5445–5449