

Intelligent churn prediction in telecom: employing mRMR feature selection and RotBoost based ensemble classification

Adnan Idris · Asifullah Khan · Yeon Soo Lee

Published online: 21 April 2013
© Springer Science+Business Media New York 2013

Abstract Churn prediction in telecom has recently gained substantial interest of stakeholders because of associated revenue losses.

Predicting telecom churners, is a challenging problem due to the enormous nature of the telecom datasets. In this regard, we propose an intelligent churn prediction system for telecom by employing efficient feature extraction technique and ensemble method. We have used Random Forest, Rotation Forest, RotBoost and DECORATE ensembles in combination with minimum redundancy and maximum relevance (mRMR), Fisher's ratio and F-score methods to model the telecom churn prediction problem. We have observed that mRMR method returns most explanatory features compared to Fisher's ratio and F-score, which significantly reduces the computations and help ensembles in attaining improved performance. In comparison to Random Forest, Rotation Forest and DECORATE, RotBoost in combination with mRMR features attains better prediction performance

on the standard telecom datasets. The better performance of RotBoost ensemble is largely attributed to the rotation of feature space, which enables the base classifier to learn different aspects of the churners and non-churners. Moreover, the Adaboosting process in RotBoost also contributes in achieving higher prediction accuracy by handling hard instances. The performance evaluation is conducted on standard telecom datasets using AUC, sensitivity and specificity based measures. Simulation results reveal that the proposed approach based on RotBoost in combination with mRMR features (CP-MRB) is effective in handling high dimensionality of the telecom datasets. CP-MRB offers higher accuracy in predicting churners and thus is quite prospective in modeling the challenging problems of customer churn prediction in telecom.

Keywords Churn prediction · Telecom · DECORATE · RotBoost · mRMR

A. Idris · A. Khan
Pattern Recognition Lab, Department of Computer and Information Sciences, Pakistan Institute of Engineering and Applied Sciences, Nilore, Pakistan

A. Idris
e-mail: adnanidris@pieas.edu.pk

A. Khan
e-mail: asif@pieas.edu.pk

A. Idris
Department of Computer Sciences and Information Technology, University of Poonch Rawalakot, Rawalakot, Azad Jammu & Kashmir, Pakistan

Y.S. Lee (✉)
Department of Biomedical Engineering, College of Medical Science, Catholic University of Daegu, Daegu, South Korea
e-mail: yeonsoolee@cu.ac.kr

1 Introduction

Customer churn prediction in telecom is attaining serious attention of stakeholders in order to retain the customer's loyalty and improve the standard of customer relation management. The telecom operators realize the importance of retaining the customers instead of striving for adding new customers every time. The cost incurred to add a new customer is far more than retaining a customer whose appetite is not being properly served [1]. The telecom operators not only maintain the stable customer base but avoid customer churning by appropriately targeting the customers, that are predicted unsatisfied by a churn prediction model.

Customer churn prediction is a binary classification problem but the large dimensionality and less instances of minority class in the telecom datasets emerge as major hurdles for

conventional classifiers to show desired performance. KNN, a widely used classifier, shows good performance on various classification problems [3–6] and its hybridized form with Logistic Regression [7] also claims competitive performance for churn prediction. However, this performance is constrained to application domains where datasets do not possess high dimensionality and imbalance distribution. Few other ensemble classification algorithms have also been applied to model churn prediction in telecom. One such algorithm uses AdaBoost and ANN Boosting to predict churners in telecom [8]. Likewise, Logit and ANN ensembles [9], and Bagging with Stochastic Gradient Boosting [10] are also used for churn prediction in telecom. Another C5.0 Boosting ensemble is also reported in literature for predicting churners in telecom [11]. Though, ensemble classifiers are considered better performers compared to single classification algorithm [12–14] but they suffer in achieving desired accuracy for predicting churners in telecom.

Researchers have also used tree based ensemble classification methods, such as Random Forest [15], Balanced Random Forest [16], Rotation Forest [17], RotBoost [18] and its variants [19] for dealing with the problem of churn prediction, but these approaches also lack required effectiveness for predicting churn in telecom.

Telecom companies archive the detail call records of various services availed by the customers. The database also includes bill and payment information, customers' demographics, complaint record etc. Such a large set of information makes cellular dataset huge in size compared to the datasets generated in other subscription based domains. Moreover, cellular datasets also have large dimensionality, for considering multiple attributes ranging from personal demographics to detailed call records. Cellular dataset has also imbalanced nature with respect to the share of minority classes. This enormous size, large dimensionality and imbalance nature of cellular datasets recognize telecom churn prediction a different problem compared to other application areas.

In recent studies, a new set of features have been proposed for modeling churn prediction in telecom [20]. The new feature set has shown improvement with several modeling techniques for predicting churners, but the focus of this study is land-line customer churn prediction, thus the proposed feature set cannot be useful for churn prediction of cellular users.

The dataset of cellular users is different to the one of non-cellular users. Generally, landline service providers acquire limited information and mostly focus on billing and payments records only [21]. Therefore, both the datasets have different features, which eventually require different churn prediction models. Contemporary literature considers cellular and non-cellular users as different application domains of churn prediction [22–24]. In addition, there is a growing trend of increase in cellular users around the world,

which results in high competition between the cellular service providers. Thus, cellular service providers require an efficient churn prediction model to maintain low churn rate and meet the challenges of saturated markets. Therefore, in order to meet the requirements of concerned stakeholders, we have focused in this study to develop an efficient churn prediction model.

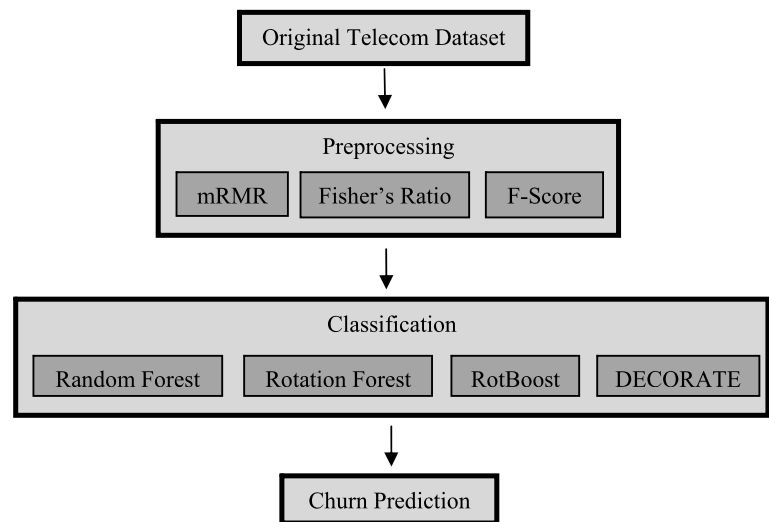
Moreover, contemporary literature presents rare studies specifically focusing to address the unique problems faced in churn prediction of cellular users. Therefore, it is hypothesized that a considerable margin of improvement exists to concentrate on dealing with enormous nature of the telecom dataset during preprocessing phase, which may extend better learning capabilities to the classifiers.

In this work, we have applied F-Score, Fisher's ratio and minimum redundancy and maximum relevance (mRMR) feature extraction schemes and it is found that mRMR scheme provides a reduced set of meaningful features that consequently extends better learning capabilities [25]. mRMR reduced feature set improves the overall performance of used classification methods, where RotBoost shows competitive performance over Random Forest, Rotation Forest and DECORATE for predicting churners. Results are evaluated on two standard telecom datasets. Thus RotBoost in combination with mRMR (CP-MRB) makes a unique contribution in the domain of churn prediction of cellular users. AUC, sensitivity and specificity based measures are applied to evaluate the performances of different classification ensembles using 10 fold cross validation.

The rest of the manuscript includes methodology in Sect. 2, which also elaborates the ensemble approaches and feature selection techniques used in this work. Section 3 presents results and detailed analysis with focus on how mRMR approach develops a feature space with maximum discriminative information, which consequently improves the ensembles' prediction performance. In the end discussion and conclusions are drawn in Sect. 4.

2 Proposed churn prediction model

Different data mining algorithms are available that are used to generalize the relationship between the decision of declaring a customer, churning or non-churning and predictor variables. The predictor variables exhibit the characteristics of the customers. The performance of such an algorithm to model churn prediction depends on the nature of input dataset. The cellular dataset carries comprehensive information that describes the relationship of a customer with the company. This comprehensive information turns cellular dataset large, in terms of dimensionality and size. Moreover, the rare presence of churners in cellular dataset also results in imbalance class distribution. Therefore, cellular dataset

Fig. 1 Proposed churn prediction model

requires an effective preprocessing prior to being used for training the classifier.

Generally, ensemble methods are considered a better tools for modeling churn prediction in various areas [7]. In the current work, a simulation based study is performed to analyze the capabilities of high performing ensembles in collaboration with various feature extraction methods to specifically model the problem of telecom churn prediction. The original dataset is initially processed to remove the useless and missing values with the help of filters available in the WEKA data mining tool. Nominal values present in the dataset are dealt by applying a sampling method that aggregates nominal values in three categories, i.e. small, medium and large, depending upon the number of instances in each category [26]. In case, the training dataset possesses imbalanced distribution between the classes, then under-sampling methodology is employed to handle the imbalance of the dataset. The under-sampling of the training dataset restrains ensembles from being biased towards the dominating class.

Afterwards, mRMR, Fisher's Ratio and F-Score feature reduction methods are employed in combination with four ensembles: Random Forest, Rotation Forest, RotBoost and DECORATE, as shown in Fig. 1. Feature selection methods are used to effectively handle the large dimensionality of the training set, so that it is convenient for the ensembles to perform effective learning [27]. The performance of each ensemble is evaluated in the context of applied feature selection method. Further searching the selected features that extend maximum learning to the ensemble chooses the best feature set.

mRMR in collaboration with RotBoost shows higher accuracy in predicting churners using two standard telecom datasets, compared to other combinations of feature extraction and ensemble methods used in this work. Thus, the

combination of effective preprocessing with high performing ensemble approach handles large dimensionality and produces improved prediction performance.

We have adopted 10-fold cross validation to assess the performance of various combinations of feature extraction methods and ensembles. 10-fold cross-validation is a helpful strategy in measuring the generalization capabilities of the classification approach [28]. The dataset D is divided in 10 sets, D_1, D_2, \dots, D_{10} . For each iteration ($i = 1$ to 10) the classifier is trained with $D - D_i$ folds and tested with D_i fold. In the end, the results are accumulated. Although, telecom datasets are large in size and the use of 10 fold cross validation increases the computations but accuracy results produces this way are more generalized.

2.1 Maximum Relevance and Minimum Redundancy (mRMR) based feature selection

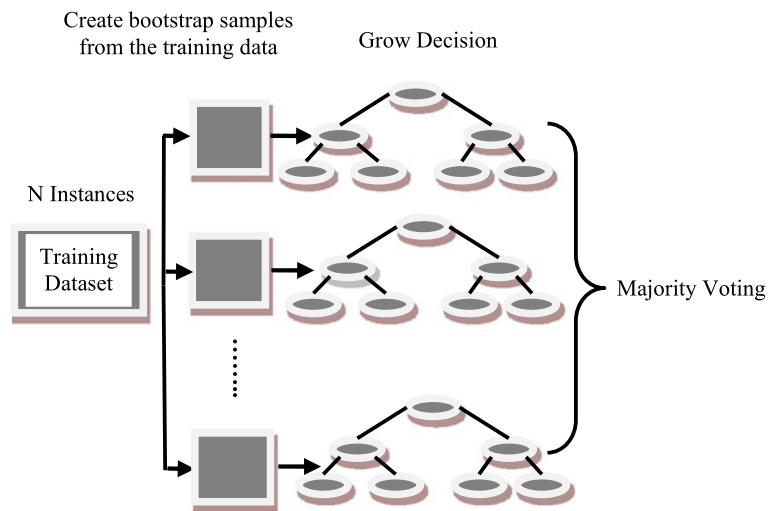
mRMR selects the features, which contain maximum discriminative information. This is accomplished by maximizing the interclass and minimizing the intraclass proximities. mRMR works by selecting the features, which are strongly correlated with class labels and not dependant on each other [25]. mRMR's criteria for feature selection is based on minimum redundancy and maximum relevance. The maximum relevance is implemented with the help of the expressions given in (1) and (2);

$$\max D(S, c), D = I(\{x_i, i = 1, \dots, m\}; c) \quad (1)$$

$$\max D(S, c), D = \frac{1}{|S|} \sum_{x_i \in S} I(x_i; c) \quad (2)$$

where D is dependency that is intended to be maximized in order to establish maximum relevance of the instances S with class labels c . $I(x_i; c)$ measures the mutual information between the instance x_i and the corresponding class

Fig. 2 Random Forest



label c . The maximum relevance is sort out by searching the feature set, which satisfy the criteria in (1) and approximates the $D(S, c)$ in (2) with the mean value of all mutual information values between individual feature x_i and class c . A features set S is chosen where the features have higher dependency on the respective class labels. The maximum relevant selected features may have redundancy among them. Therefore, in order to reduce this redundancy, any single redundant feature is selected which does not compromise the discriminative power of the feature set. The expression given in (3) minimizes the redundancy:

$$\min R(S), R = \frac{1}{|S|^2} \sum_{x_i, x_j \in S} I(x_i, x_j) \tag{3}$$

The criteria of minimizing redundancy and maximizing relevance are combined in a simple form, where Φ is defined as given in (4). This simplest form is used to optimize both D and R .

$$\max \Phi(D, R), \quad \Phi = D - R \tag{4}$$

The feature set obtained using mRMR is expected to be discriminative, for showing strong relevance with the class targets and at the same time having features with maximum unique values.

2.2 Fisher’s Ratio based feature selection

Fisher’s Ratio is considered to be sensitive to the normalization of data and measures the discriminating power of the features in the dataset. Fisher’s Ratio is computed as given in (5).

$$\text{Fisher’s Ratio} = \frac{(\mu_1 - \mu_2)}{\sigma_1^2 - \sigma_2^2} \tag{5}$$

where μ_1 and μ_2 are the means of binary classes involved, and σ_1^2 and σ_2^2 the respective variances.

2.3 F-score based feature selection

F-score is a simple technique, which measures the discrimination of two sets of real numbers. Given training vectors $x_k, k = 1, \dots, m$, if the number of instances of churner and non-churner classes are n_+ and n_- , respectively, then F-score of the i th feature is defined as:

$$F_i = \frac{(\bar{x}_i^{(+)} - \bar{x}_i)^2 + (\bar{x}_i^{(-)} - \bar{x}_i)^2}{\frac{1}{n_+ - 1} \sum_{k=1}^{n_+} (x_{k,i}^{(+)} - \bar{x}_i^{(+)})^2 + \frac{1}{n_- - 1} \sum_{k=1}^{n_-} (x_{k,i}^{(-)} - \bar{x}_i^{(-)})^2} \tag{6}$$

where \bar{x}_i is the mean value of i th feature, $\bar{x}_i^{(-)}$ is the mean value for negative instances, and $\bar{x}_i^{(+)}$ is the mean value of the i th feature for positive instances. F-score minimizes the intraclass distance, whereas maximizes the interclass distance of the instances as shown in (6). The larger the F-score is, more likely the feature is discriminative.

2.4 Random Forest

Random Forest [29] is an ensemble of decision trees, which are grown over bootstrap samples of the training dataset, involving random feature selection in the process of tree construction. The final predictions are made by aggregating the predictions of all individual trees, as represented in Fig. 2. Random Forest, being the ensemble of the decision trees, certainly exhibit substantial performance improvement over single tree based classifiers. Although, Random Forest is considered to be good choice in handling the large sized data, but it suffers in the case of imbalanced training dataset. Random Forest minimizes the overall error rate, therefore in the case of imbalanced dataset; the higher total accuracy sometimes undermines the true prediction of the minority class. The telecom datasets normally suffer from higher degree of skewness, therefore Random Forest sometimes suffers to show appreciable performance.

2.5 Rotation Forest

Rotation Forest [17, 19] is a new ensemble classifier that operates by simultaneously improving diversity and accuracy. Rotation Forest achieves high diversity by employing rotation through linear feature extraction methods such as Principal Component Analysis (PCA), Independent Component Analysis (ICA) etc. on the input data. The original dataset is divided in K subsets (K is a Rotation Forest parameter) derived from L original feature space. The desired accuracy is attained by utilizing all the components of each subset during learning of the base classifier. This also preserves the variability information in the data. The original feature space L is split into K subsets. Then, PCA is applied on each subset, which results in K axis rotation and forms the new attributes for a base classifier. Rotation Forest encourages diversity by using PCA as a feature extraction method for each base classifier. Whereas, better accuracy is achieved by utilizing all the principal components for training each base classifier. Let, the class of an instance X is predicted with the Rotation Forest ensemble C^* as given in (7):

$$C^*(X) = \operatorname{argmax}_{y \in \Phi} \sum_{t=1}^T I(C_t(X R_t^a) = y) \tag{7}$$

C_t ($1 \dots T$) shows the base classifier and y corresponds to either 0 or 1 for the binary nature of churn prediction problem. Whereas, R_t^a is the rotation matrix, derived for each of the feature subsets ($1 \dots T$). I is an indicator function that assigns the instance X , 0 or 1.

Rotation Forest reports superior performance over bagging, AdaBoost and Random Forests on a number of problems. This improved performance is mainly accredited to simultaneous improvement in diversity and accuracy. The increased diversity within the ensemble is achieved by using feature extraction from the training data and using decision trees as base classifiers, which are considered sensitive to variations in training data. Similarly, the increased accuracy is obtained by utilizing all extracted features, which are principal components when PCA is used as feature extraction method. As Rotation Forest is proven successful in modeling various problems, we have also considered Rotation Forest in combination with various feature extraction methods to model churn prediction in telecom.

2.6 RotBoost

RotBoost is an ensemble classifier generation technique that is developed by combining AdaBoost and Rotation Forest [18, 19]. Adaboost operates in a sequential manner where each new classifier is constructed, considering the performance of previous classifier. In this method, a set of weights is maintained over the original training set, where initially

for $s = 1, 2, \dots, S$

- i. Compute the rotation matrix, R_s^a following the steps specified by Rotation Forest. $I^a = [X R_s^a y]$ is the training dataset extracted corresponding to s_i feature subset for classifier C_S .
- ii. Initialization of the weight distribution for I^a as $D_1(i) = \frac{1}{N}$ ($i = 1, 2, \dots, N$) governed by AdaBoost

for $t = 1, 2, \dots, T$

- iii. A new set I_t^a is developed performing N extractions from I^a with replacement.
- iv. A classifier C_t^a is trained on I_t^a set using W base learning algorithm. The error of C_t^a is computed and weights over distribution D^t are updated as per Adaboost algorithm.
- v. End for

$$C_s(x) = \operatorname{argmax}_{y \in \Phi} \sum_{t=1}^T \alpha_t I(C_t^a(x) = y)$$

Prediction

The ensemble C^* predicts an instance of dataset as churner or non-churner:

$C_s(x) = \operatorname{argmax}_{y \in \Phi} \sum_{s=1}^S I(S_s(x) = y)$ where, $I(\cdot)$ is an indicator function.

Fig. 3 The pseudocode of RotBoost algorithm

they all are kept equal. In subsequent iterations, the misclassified instances are given more weights and correctly classified instances are given less weights. In this way, subsequently trained classifiers better handle hard instances. In RotBoost, weight updation over the training data distribution is taken from AdaBoost, while rotation matrix is computed in similar fashion as in Rotation Forest. Generally, the good performance of an ensemble is also attributed to the base learning algorithm [20]. Since decision trees are sensitive to little permutations in the learning data, they are considered suitable to be used as base learner in RotBoost. The pseudo code of the RotBoost algorithm is provided in Fig. 3.

S and T parameters specify the number of iterations to be carried out by Rotation Forest and AdaBoost, respectively. K is the parameter that is provided to the Rotation Forest part of RotBoost that decides the number of feature subsets. W represents a base learning algorithm. x is a data instance that is to be classified. I shows the training dataset, where $I = \{(x_i, y_i)\}$ ($i = 1, 2, 3 \dots, N$) denoted by two random variables $[X Y]$. X is a matrix containing attributes and their corresponding values in an order of $N \times p$, while Y is N dimensional column vector keeping the class targets.

RotBoost is a newly introduced ensemble method and also shows good performance on various problems compared to bagging, CART and C4.5 [20]. Therefore, we also

include RotBoost in our work, to be explored in combination with various feature extraction methods for modeling telecom churn prediction.

2.7 Decorate

DECORATE (Diverse Ensemble Creation by Oppositional Relabeling of Artificial Training Examples) is an ensemble generation technique that develops a diverse hypothesis using artificially created training instances in addition to the original training set [30]. DECORATE generates an ensemble, following an interactive process, where a classifier is first learnt and then added to the current ensemble. The ensemble members, in successive iterations, are trained with combination of original and artificially created training set. Artificial training instances are created in each iteration from the data distribution, where the size of these artificially created instances is specified as a fraction of original training sets. The labels of these artificially created instances are chosen so as to be differing maximally from the predictions made by current ensemble. The artificial instances for numerical features are computed by first calculating mean and standard deviation from the training set. Computed statistics, are then used to generate values for numeric attributes by using them to define Gaussian distribution. Artificial values for nominal attributes are computed by calculating the probability of occurrence of each distinct value in its domain and generating values based on this distribution. Laplace smoothing is also applied, in order to ensure the non-zero probability of occurrence for nominal attribute not being represented in the training set. The artificially generated instances are assigned labels based on the predictions made by ensemble in each iteration.

The procedure works by first finding the class membership probabilities using ensemble. The zero probabilities are replaced with a small non-zero value and are normalized to make a distribution. Then, labels are selected in a way that probability of selection is inversely proportional to the predictions made by current ensemble. So, if $P_y(x)$ represents the class membership probabilities, predicted by current ensemble as given in (8), then a new label is designated based on the distribution:

$$P_y^{-1}(x) = \frac{1/P_y(x)}{\sum_y 1/P_y(x)} \quad (8)$$

This artificially generated training set is referred as diversity data. The new classifier is trained on the union of original training data and the diversity data, which enables it to be different from the current ensemble. Thus, adding such a classifier to ensemble increased its diversity. In addition, the accuracy is also maintained by not considering the new classifier, if it deteriorates ensemble's accuracy. This

BaseLearner—decision trees used as base learner

T—training dataset, with labels

C_{size}—desired ensemble size

I_{max}—maximum number of iterations

R_{size}—size of artificial instances, to be generated

- i. $i = 1$
- ii. $trials = 1$
- iii. $C_i = BaseLearner(T)$
- iv. *Initializing Ensemble* $C^* = \{C_i\}$
- v. *Ensemble Error*, $\epsilon = \frac{\sum_{x_j \in T} C^*(x_j) \neq y_j}{m}$
While $i < C_{size}$ and $trials < I_{max}$
- vi. $R = R_{size} \times T$ training instances are created, based on distribution of training data
- vii. Labels are assigned to instances in R , with probability of class labels inversely proportional to prediction of C^*
- viii. $T = T \cup R$
- ix. $C' = BaseLearner(T)$
- x. $C^* = C^* \cup \{C'\}$
- xi. $T = T - R$, removing the artificial data
- xii. Compute the error ϵ' for classifier C^* as shown in step v
- xiii. If $\epsilon' < \epsilon$
- xiv. $i = i + 1$
- xv. $\epsilon = \epsilon'$
- xvi. else
- xvii. $C^* = C^* - \{C'\}$
- xviii. $trials = trials + 1$

Fig. 4 The pseudocode of DECORATE algorithm

iterative process continues till the criteria are met, either committee size is reached or number of iterations is exceeded.

Let's an instance x be classified using DECORATE. The class membership probabilities of x , are computed by each of the base classifier, C_i in ensemble C^* . C_i computes probability $PC_{i,y}(x)$ for an instance x belonging to class y . Then, the class membership probabilities for the entire ensemble are computed as given in (9):

$$P_y(x) = \frac{\sum_{C_i \in C^*} PC_{i,y}(x)}{|C^*|} \quad (9)$$

where $P_y(x)$ represents probability of instance x belonging to the class y . In case of predicting churn, the problem involves binary classes, therefore most probable class is chosen as the label for an instance x , as given in (10):

$$C^*(x) = \operatorname{argmax}_{y \in Y} P_y(x) \quad (10)$$

The pseudocode of the DECORATE algorithm is produced in Fig. 4.

3 Results and discussion

In current work, we have performed the detailed experimentation involving feature selection and classification approaches to model churn prediction for telecom. The implementation of Random Forest, Rotation Forest, RotBoost, DECORATE and other feature extraction methods are accomplished using Matlab. The considered ensemble classifiers comprise of 100 constituent members per class, while decision trees are used as base classifiers in each of the ensembles. Generally, AUC is considered a better measure to evaluate the performance of a classifier [22]. Therefore, it is used along with sensitivity and specificity measures to study the impact of feature extraction methods on ensembles to model churn prediction in telecom.

3.1 Dataset

Public datasets are rarely available which are to be used for telecom churn prediction. Most of the churn prediction studies only explain the characteristics of the used datasets and then analyze the performance of the predictors [16, 19, 23, 24, 31]. The privacy of the customers' restrains researchers and companies from publicly publishing the dataset. However few processed telecom datasets are public, which are used in this study to model the churn prediction problem. Table 1 describes the characteristics of the used datasets. Orange dataset is large in size with 50k instances and 260 features [32]. The dataset is comprised of 190 numerical and 70 nominal features. Eighteen of the features in the Orange dataset have no value at all and five of the features have only one value, thus these useless features are discarded. Orange dataset is also imbalanced in nature as the

Table 1 The characteristics of the used telecom datasets

	Orange Telecom	Cell2Cell
Total Instances	50k	40k
Total Features	260	76
Numerical Features	190	68
Nominal Features	70	8
Data Distribution	Imbalanced (7.3 % minority class)	Balanced
Missing values	Yes	No

Table 2 Performance evaluation on Cell2Cell and Orange datasets

	Cell2Cell Dataset			Orange Dataset		
	AUC	Sensitivity	Specificity	AUC	Sensitivity	Specificity
Random Forest	0.592	0.690	0.601	0.571	0.0049	0.9991
Rotation Forest	0.610	0.666	0.646	0.583	0.0026	0.9998
RotBoost	0.699	0.664	0.632	0.601	0.0291	0.7212
DECORATE	0.561	0.491	0.641	0.490	0.0020	0.7011

minority class which represents churners, has only 7.3 % of scarce share in the whole dataset. The other data set is provided by Duke University [33] referred as Cell2Cell dataset. Cell2Cell dataset is already processed and provided in balanced shape with 40K instances. The dataset contains 8 nominal and 68 numerical features. The nominal values present in both the datasets, are transformed to numerical format by grouping the modalities [26] in three categories, i.e. small, medium and large depending upon the number of occurrences of instances in each category.

3.2 Performance evaluation without involving feature reduction

Table 2 reports the performances of Random Forest, Rotation Forest, RotBoost and DECORATE on the original form of two standard telecom datasets used in this study. All the four used ensembles suffer in accurately predicting churners on both the datasets, however, RotBoost comparably performs better compared to the other four ensembles but the performance is not satisfactory enough which may encourage using RotBoost in this form as a modeling technique for churn prediction. The deteriorated performance implies that unless appropriate features are not selected, the ensembles suffer in attaining good training which consequently results in low prediction accuracy. The results obtained using the original forms of datasets clearly hint for employing efficient feature selection technique, which may help ensembles to show better performance. Moreover, the low sensitivity scores obtained on Orange dataset by all the used ensembles as given in Table 2, clearly indicate the dominance of non-churners (92 %) in the dataset. Thus an under-sampling method is essentially required to establish a balance distribution between the training instances of both classes in order to avoid the biased learning. As a result, a sampling methodology is employed to establish a normal distribution between the instances of churning and non-churning classes, and hereafter the balanced Orange dataset is used in this work for further investigations.

3.3 Impact of feature reduction on Random Forest

Feature reduction is performed in the preprocessing phase in order to provide the meaningful and discriminative fea-

Table 3 Performance evaluation of Random Forest on Cell2Cell Dataset

Random Forest	Cell2Cell Dataset			
	D^*	AUC	Sensitivity	Specificity
mRMR	34	0.772	0.720	0.671
F-Score	36	0.746	0.689	0.646
Fisher's Ratio	37	0.723	0.656	0.647

Table 4 Performance evaluation of Random Forest on Orange Dataset

Random Forest	Orange Dataset			
	D^*	AUC	Sensitivity	Specificity
mRMR	36	0.7511	0.6883	0.6739
F-Score	38	0.7214	0.6508	0.6553
Fisher's Ratio	40	0.7185	0.6704	0.6489

tures to the classifiers. Mostly, classifiers suffer from irrelevant feature space [34]. mRMR, F-Score and Fisher's ratio methods are applied to analyze the impact on prediction performance of used ensembles. For each feature extraction method, exhaustive search is performed to select the maximum discriminative features which enable ensembles in attaining better prediction performance. Exhaustive search is separately conducted for mRMR, F-Score, and Fisher's Ratio in combination with Random Forest.

Table 3 reports the prediction performance of Random Forest in collaboration with the used feature extraction methods, on Cell2Cell dataset. Results show that 34 best mRMR features extend most discriminative and meaningful information to Random Forest. Random Forest attains highest AUC of 0.722 with mRMR features. Table 4, gives the performance of Random Forest on Orange dataset.

Random Forest predicts the churners with highest accuracy of 0.7511 AUC for Orange dataset using only 36 best mRMR features. Results in Table 3 and Table 4 clearly indicate that Random Forest attains improved accuracy in predicting churners for both the datasets using mRMR features. Moreover, results also show that Random Forest obtains better prediction performance with less number of mRMR features compared to F-Score and Fisher's Ratio based features.

3.4 Impact of feature reduction on Rotation Forest

Rotation Forest encourages diversity by employing PCA and rotating the input data space using rotation matrix. Rotation Forest achieves higher accuracy by utilizing all principal components extracted from feature subsets. Table 5 and Table 6 present prediction performance of Rotation Forest on both datasets. Rotation Forest in combination with set of 35

Table 5 Performance evaluation of Rotation Forest on Cell2Cell Dataset

Rotation Forest	Cell2Cell Dataset			
	D^*	AUC	Sensitivity	Specificity
mRMR	35	0.762	0.721	0.583
F-Score	37	0.691	0.670	0.629
Fisher's Ratio	37	0.652	0.603	0.610

Table 6 Performance evaluation of Rotation Forest on Orange Dataset

Rotation Forest	Orange Dataset			
	D^*	AUC	Sensitivity	Specificity
mRMR	39	0.7011	0.5983	0.6739
F-Score	46	0.6014	0.5608	0.6353
Fisher's Ratio	50	0.5685	0.4704	0.6489

Table 7 Performance evaluation of DECORATE on Cell2Cell dataset DECORATE

	Cell2Cell Dataset			
	D^*	AUC	Sensitivity	Specificity
mRMR	35	0.789	0.728	0.669
F-Score	37	0.740	0.682	0.639
Fisher's Ratio	37	0.713	0.601	0.638

best mRMR features obtain highest prediction performance of 0.762 AUC on Cell2Cell dataset. Similarly, Rotation Forest in combination with 39 best mRMR features produces improved prediction performance on Orange dataset. Thus, mRMR features extend sufficient information, which enables Rotation Forest to yield good performance. Whereas, Rotation Forest suffers in predicting churners from both the datasets when provided with F-Score and Fisher's ratio based features. Fisher's Ratio and F-Score methods are ranking based feature extraction methods, which do not consider the relationship between features and respective class labels. While, mRMR's criteria of selecting features is more succinct, which considers correlation of features with class labels and also ensures minimal inter dependency of features. This is the main reason for Rotation Forest's attaining better prediction performance in combination with mRMR features.

3.5 Impact of feature reduction on DECORATE

Tables 7 and 8 show the performance of DECORATE in combination with mRMR, F-Score and Fishers' Ratio, in predicting churners for Cell2Cell and Orange datasets,

Table 8 Performance evaluation of DECORATE on Orange dataset

DECORATE	Orange Dataset			
	D^*	AUC	Sensitivity	Specificity
mRMR	39	0.749	0.688	0.673
F-Score	40	0.711	0.670	0.645
Fisher's Ratio	44	0.699	0.660	0.632

Table 9 Performance evaluation of RotBoost on Cell2Cell Dataset

RotBoost	Cell2Cell Dataset			
	D^*	AUC	Sensitivity	Specificity
mRMR	31	0.816	0.765	0.746
F-Score	36	0.726	0.679	0.627
Fisher's Ratio	37	0.724	0.675	0.629

Table 10 Performance evaluation of RotBoost on Orange Dataset

RotBoost	Orange Dataset			
	D^*	AUC	Sensitivity	Specificity
mRMR	36	0.761	0.729	0.673
F-Score	38	0.731	0.681	0.655
Fisher's Ratio	40	0.721	0.630	0.668

respectively. DECORATE creates new artificial instances which are combined with original dataset, and then ensemble members are trained on more diversified training dataset. DECORATE attains 0.789 AUC using 35 best mRMR features from Cell2Cell dataset. Similarly, DECORATE scores 0.749 AUC with the set of 39 best mRMR features for predicting churners from Orange dataset. DECORATE's performance in combination with F-Score and Fisher's ratio deteriorates. Sensitivity measure indicates that DECORATE, when used with Fisher's Ratio, weakens prediction performance on both the datasets, as reported in Table 7 and Table 8. DECORATE assigns labels to the artificially created instances in inverse proportion to the labels of original dataset for introducing diversity. Thus, if selected features are unable to provide meaningful information, then the artificially created instances of DECORATE also do not significantly contribute in enabling better learning to the classifier.

3.6 Impact of feature reduction on RotBoost

RotBoost ensemble efficiently exploits the advantages of Rotation Forest and Adaboost methods. The Adaboost's iterative approach to tackle hard instances assists RotBoost

in attaining improved accuracy. Table 9 and Table 10 list the results that RotBoost has achieved in predicting churners for Cell2Cell and Orange datasets. It is clearly observed that RotBoost in collaboration with only 31 best mRMR features produce highest prediction performance on Cell2Cell dataset. Similarly, RotBoost also achieves highest prediction performance on Orange dataset using 36 best mRMR features. RotBoost achieves 0.816 AUC for Cell2Cell dataset and 0.7612 AUC for Orange dataset respectively that is appreciable prediction performance compared to the other combinations of feature selection and ensembles methods.

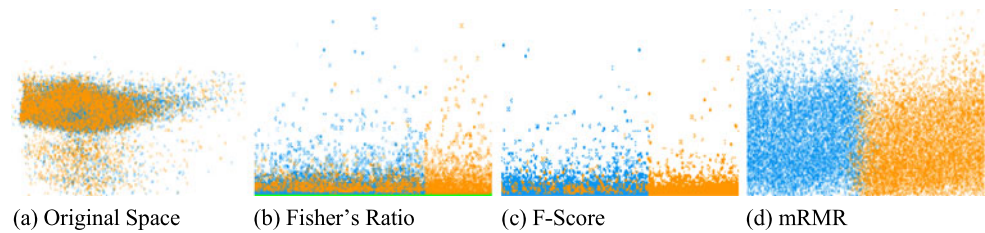
mRMR feature extraction method proves its effectiveness in combination with RotBoost as well, compared to F-Score and Fishers' ratio methods. mRMR method adopts criteria that characterizes the relationship of a feature with class label and also improves the proximity of the instances in data space, which ultimately helps ensembles to attain better learning. RotBoost shows highest accuracy in predicting churners with only 31 best mRMR features as shown in Table 9. A set of 36 and 37 best features are extracted using F-Score and Fisher's Ratio separately which yield 0.726 and 0.724 AUC, respectively, with RotBoost. However, RotBoost achieves the highest accuracy of 0.816 AUC with only 31 best mRMR extracted features.

3.7 Discrimination analysis of feature extraction methods

Fisher's ratio measures the discrimination offered by the features by considering mean and variance of the binary classes in the telecom dataset. Fisher's ratio is considered sensitive to the non-normality of the dataset. While F-Score adopts better criteria, which minimizes the intraclass, distance and maximizes the interclass distance of the instances. It is observed from the results shown in Tables 2–7, that F-Score's criteria of selecting features is effective compared to Fisher's ratio and it consequently impacts in attaining better learning.

F-Score's ability of enabling ensembles to produce better prediction results compared to Fisher's Ratio lies in minimizing the intraclass and maximizing interclass distance of data instances. mRMR method selects the features which show strong correlations with class labels, while being not dependent on each other. This way, the proximity of instances is improved that ultimately extends better learning to ensembles. mRMR method selects the features which are minimally redundant and maximally relevant. The results shown in Tables 2–7 clearly indicate that mRMR is an effective method compared to F-Score and Fisher's Ratio in selecting discriminative features which enable ensemble to attain improved prediction performance. Additionally, in order to study the impact of feature extraction method on data space, PCA is applied on dataset before and after applying a specific feature extraction method. PCA encompasses the maximum variance present in the dataset and yields artificial principal components. The initial principal components

Fig. 5 Feature spaces of Cell2Cell dataset



encompass maximum variance of the dataset. Principal component one (PC1) is plotted against principal component two (PC2) to see the separability introduced between instances after applying a feature extraction method. Figure 5, shows the feature spaces before and after applying a feature selection method.

3.7.1 mRMR vs. Fisher's Ratio vs. F-Score

Data space of the original dataset is shown in Fig. 5(a). It can be clearly observed that instances share same coordinates in the data space. Therefore, it will be difficult for a classifier to learn a distinctive pattern from such a complex data space. mRMR method transforms the feature space which ultimately establishes favorable classification pattern for an ensemble compared to Fisher's Ratio and F-Score as shown in Fig. 5(d). mRMR extracted features establish strong relevance with class targets and also ensure maximum unique instances present in the dataset which ultimately establishes distinguishable separability between churners and non-churners. The plots shown in Fig. 5 clearly indicate the difference in the feature space, before and after applying mRMR method and other feature extraction methods. mRMR method considers the correlation of feature with class labels, which improves the feature extraction criteria compared to ranking criteria of F-Score and Fisher's ratio. mRMR features introduce separability between the instances of churners and non-churners which consequently improve the prediction performance of used ensembles. Therefore, mRMR has emerged as the promising technique compared to F-Score and Fisher's Ratio for extracting most discriminative features, which improves the prediction performance.

3.8 Performance comparison of Random Forest, DECORATE, Rotation Forest, and RotBoost

Random Forest and rotation based ensembles have already been used to model churn prediction problems in telecom [18, 19], but they do not achieve the desired accuracy in predicting churners. The complex nature of the telecom dataset is the reason behind such a deteriorated performance. Therefore, in this study, a churn prediction model is proposed that handles the enormous nature of the telecom dataset and

addresses the unique problems faced in predicting telecom churners.

The results shown in Tables 2–7 clearly highlight that the extraction of useful features from the dataset enables classifier in attaining improved performance. Random Forest, Rotation Forest, DECORATE and RotBoost have shown improved prediction performance when mRMR extracted features are provided for training. Simulation results show that appropriate feature selection method is a vital factor in attaining higher prediction performance. mRMR method appears most effective compared to Fisher's Ratio and F-Score. The ensemble methods improve the prediction performance on both the datasets when provided with mRMR features. Moreover, the best mRMR features which produce good results are less in number compared to the best feature sets selected through F-Score and Fisher's Ratio as shown in Tables 2–7. The reduced number of mRMR features provides the most meaningful and discriminative information, that ensemble requires to learn pattern from complex telecom datasets.

Random Forest and DECORATE in combination with mRMR features, produce better prediction performance for both datasets, compared to Rotation Forest as shown in Figs. 7 and 8. Whereas, RotBoost in combination with mRMR (CP-MRB), outperforms all other combination of ensemble and feature extraction methods. Although, Rotation Forest does not show good prediction results on both the datasets, but when it is combined with Adaboost in RotBoost ensemble, it achieves maximum accuracy in predicting churners for both the datasets as shown in Fig. 8 and Fig. 9. RotBoost exploits the advantages of both, Rotation Forest and Adaboost methods. RotBoost inherits the better tradeoff of variance-bias term from Rotation Forest. Additionally, Adaboosting in RotBoost contributes the iterative process of updating weights over training instances to maximize the prediction performance. Rotation Forest (part of RotBoost) develops a rotation matrix, using principal components extracted from feature subsets. Principal components are rearranged in the rotation matrix to rotate the training data in order to maximize the diversity. The Adaboost part works in a sequential manner, which considers the performance of previously generated classifiers and updates the weights of training instances accordingly. This iterative approach of Adaboost to focus on hard instances makes RotBoost achieve higher accuracy in predicting scarce churners,

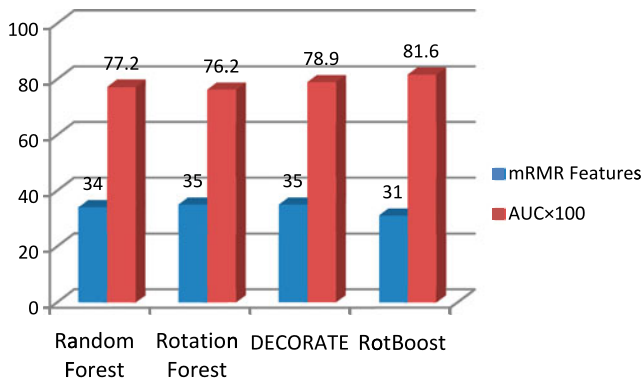


Fig. 6 Performance comparison of Random Forest, Rotation Forest, DECORATE and RotBoost on Cell2Cell dataset

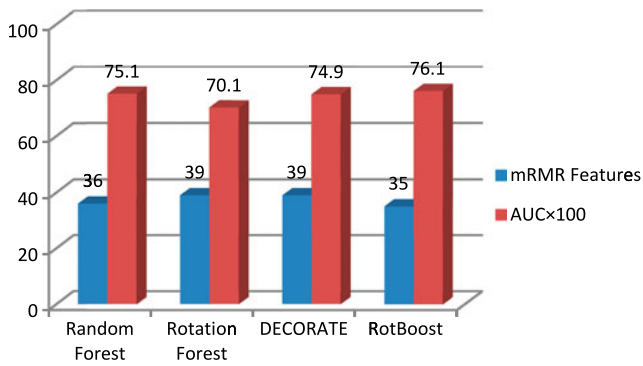


Fig. 7 Performance comparison of Random Forest, Rotation Forest, DECORATE and RotBoost on Orange dataset

compared to Random Forest, Rotation Forest and DECORATE.

In addition to the best mRMR features provided to RotBoost, the implicit feature selection is also performed in developing rotation matrix, using PCA which further improves the learning of pattern by rotating the training set. The best mRMR features are also less in number which results in fewer computations involved in Adaboosting process of RotBoost. The combination of mRMR and RotBoost in our proposed CP-MRB approach appears as most promising in predicting churners for telecom datasets. CP-MRB achieves an accuracy of 0.816 AUC with only 31 best mRMR features, on Cell2Cell dataset, as shown in Fig. 6. While CP-MRB obtains 0.761AUC on Orange dataset with only 35 best mRMR features, as given in Fig. 7.

3.9 CP-MRB churn predictor using mRMR and RotBoost

CP-MRB approach is based on mRMR and RotBoost methods to model the challenging problem of churn prediction in telecom. mRMR method, adopts a unique criteria of minimal-redundancy and maximal-relevance for selecting the features. Features, which have strong relevance with

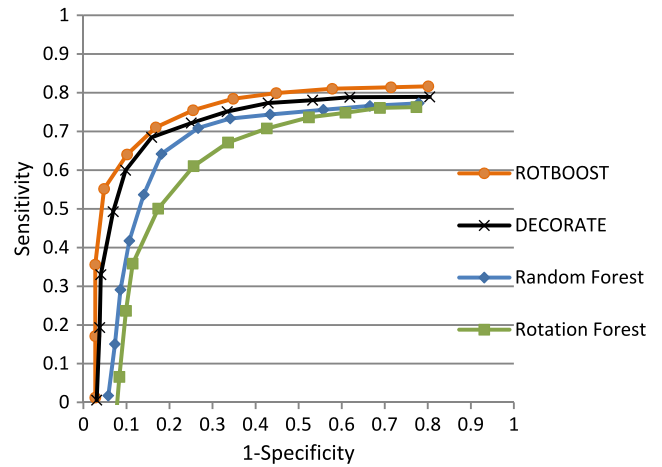


Fig. 8 Performance comparison of Random Forest, Rotation Forest and RotBoost in terms of AUC, on Cell2Cell dataset

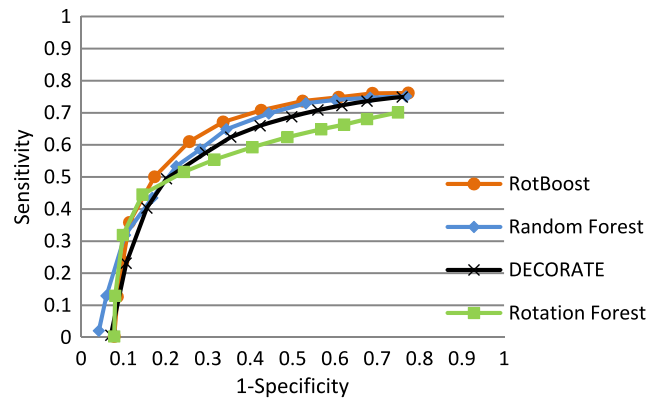


Fig. 9 Performance comparison of Random Forest, Rotation Forest and RotBoost in terms of AUC, on Orange dataset

class targets, are selected through calculating mutual information between each feature and class label. The maximum relevance criteria may lead to the feature set which are redundant i.e. dependency among such features could be large. Therefore one of the two dependent features is removed, which does not alter the respective class’s discriminating power. Thus, mRMR method selects the features considering their relevance with class labels, unlike Fisher’s ratio and F-score’s ranking criteria. This way, a reduced feature set is selected which is less redundant and contains maximum discriminative information, which consequently extend better learning to ensembles. All the used ensembles (Random Forest, Rotation Forest, DECORATE and RotBoost) obtain highest prediction performance using mRMR extracted features. Whereas, RotBoost obtains best prediction performance on the datasets by efficiently exploiting the reduced and meaningful mRMR features as shown in Fig. 8 and Fig. 9. Thus, the best results produced by CP-MRB are attributed to:

Table 11 Performance comparison of the proposed CP-MRB approach with existing approaches

Performance Comparison on Orange Dataset	
Method	AUC
CP-MRB	0.761
Gradient Boosting Machine [35]	0.737
Stochastic Gradient Boosting [37]	0.728
Decision stump based model [36]	0.725
Decision tree based model [36]	0.715
Bayesian Net (BN) based approach [31]	0.714

- Use of mRMR method in selecting most relevant features, in addition to the inherent feature selection performed by Rotation Forest in RotBoost.
- Maximum diversity obtained within ensemble (RotBoost) by using feature extraction on training data, utilizing all principle components and using decision trees as base classifier, considered to be sensitive to variations in training data [19].
- Iterative Adaboosting process in RotBoost, to adjust the weights of hard instances, which plays vital role in truly predicting the scare churners in the telecom datasets and contributes significantly for achieving improved accuracy.

3.10 CP-MRB comparison with other existing approaches

CP-MRB approach emerges as most effective in accurately predicting churners for Cell2Cell and Orange datasets as shown in Fig. 8 and Fig. 9. Table 11 gives the comparison of CP-MRB with other existing approaches, on the basis of AUC scores attained on Orange dataset.

Results produced by CP-MRB are compared with a method that use a gradient boosting machine to predict churners and achieve highest AUC of 0.737, reported so far on the small version of Orange dataset [35]. The gradient boosting based approach uses decision trees as base classifier with boosting, and adopts ranking based feature selection criteria. The instances are split into 1 % quantiles and mean response for each quantile is calculated using half of the training data. The calculated mean is applied to the other half and AUC is calculated which rank the variables. This method lacks a systematic approach, and is based on imputational method for feature selection.

We have also compared our results with an Adaboost based approach [36]. In this work, Adaboost based model is optimized with multi armed bandit (MABs). Adaboost builds a classifier in a step wise fashion by adding simple base classifiers to pool and use voting for the final prediction. The approach constructs the data subsets optimized through MABs and then ultimately Adaboost only searches these subsets instead of optimizing the base classifier over

Table 12 Performance comparison of the proposed CP-MRB approach with existing approach

Performance Comparison on Cell2Cell Dataset	
Method	AUC
CP-MRB	0.816
Naïve Bayes (NB) based approach [31]	0.818

the whole space. The results in Table 12 show that the 0.7258 AUC and 0.7158 AUC are attained [36], using tree and stump based learners with Adaboost respectively for Orange dataset.

A comparison with Stochastic Gradient Boosting algorithm is also made [37], that applies boosting with decision trees as classification method to model churn prediction. This method scores 0.7282 AUC for Orange dataset. In another study, Bayesian networks with oversampling attains 0.714 AUC, amongst other classification methods used to prediction churners for Orange dataset [31].

In our CP-MRB approach a systematic methodology is followed by first, selecting features which have maximum explanatory power, using mRMR and then high performing RotBoost is employed to predict churners. CP-MRB attains highest AUC of 0.761 on Orange dataset that is competitive to AUC scores produced by other existing methods as shown in Table 11. The highest AUC score, reported so far on Orange dataset is 0.737 [35] therefore, our CP-MRB approach outperforms other existing approaches and achieves highest accuracy of 0.761 AUC.

CP-MRB approach attains competitive prediction performance on Cell2Cell dataset as well, compared to Naïve Bayes's performance [31]. The difference in performance is marginal, therefore CP-MRB approach is regarded as more generic compared to the Naïve Bayes based approach. In Verkerke W. et al. [31] work different classification techniques produce high prediction accuracy on different datasets, while our CP-MRB approach has attained high prediction performance on both the datasets. Moreover, we have adopted 10 fold cross validation to evaluate performance of our CP-MRB approach, whereas a single random split of the data set is performed in Verkerke W. et al. [31] work to evaluate performance. Therefore CP-MRB approach is considered as a more generic solution to model telecom churn prediction that efficiently handles the telecom datasets and produces high accuracy in predicting churners.

4 Conclusion

mRMR and RotBoost appear as promising feature extraction and classification tools, respectively to model the challenging problem of churn prediction in telecom. mRMR efficiently reduces the feature space which in turn provides

better learning capabilities to RotBoost. Hence, mRMR appears as an efficient feature reduction technique to mitigate the high dimensionality of the telecom dataset compared to Fishers' Ratio and F-Score. mRMR not only reduces the feature space but also provides the most discriminative features which eventually help RotBoost to attain the highest prediction accuracy. RotBoost is a rotation based ensemble that achieves high accuracy and diversity. It operates in an iterative manner, thus small sized feature space also lessens the computations involve in training and testing phases. Thus our CP-MRB proposed approach, uniquely employs mRMR and RotBoost that effectively handles the main hurdle of high dimensionality of telecom dataset, and attains good prediction performance. Consequently, CP-MRB is believed to be beneficial for enhanced prediction of the churners in telecom industry.

Acknowledgement This work is supported by the Higher Education Commission of Pakistan (HEC) as per award No. 17-5-6(Ps6-002)/HEC/Sch/2010 and Korean National Research Foundation as per grant No. (NRF-2011-0006806).

References

- Reinartz WJ, Kumar V (2003) The impact of customer relationship characteristics on profitable lifetime duration. *J Mark* 67(1):77
- Lee T-S, Chiu C-C, Chou Y-C, Lu C-J (2004) Mining the customer credit using classification and regression tree and multivariate adaptive regression splines. *Comput Stat Data Anal* 50(4):1113–1130
- Ruta D, Nauck D, Azvine B (2006) K nearest sequence method and its application to churn prediction. In: *Intelligent data engineering and automated learning—IDEAL 2006. Lecture notes in computer sciences*, vol 4224, pp 207–215
- Khan A, Khan MF, Choi T-S (2008) Proximity base GPCRs prediction in transform domain. *Biochem Biophys Res Commun* 371(3):411–415
- Tan S (2006) An effective refinement strategy for KNN text classifiers. *Expert Syst Appl* 30(2):290–298
- Zhao L, Wang L, Xu Q (2012) Data stream classification with artificial endocrine system. *Appl Intell* 37(3):390–404
- Zhang Y, Qi J, Shu H, Cao J (2007) A hybrid KNN-LR classifier and its application in customer churn prediction. In: *IEEE international conference on systems, man and cybernetics*, pp 3265–3269
- Mozer MC, Wolniewicz R, Grimes DB, Johnson E, Kaushansky H (2000) Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry. *IEEE Trans Neural Netw* 11(3):690–696
- Kim Y (2006) Toward a successful CRM: variable selection, sampling, and ensemble. *Decis Support Syst* 41(2):542–553
- Lemmens A, Croux C (2006) Bagging and boosting classification trees to predict churn. *J Mark Res* 43(2):276–286
- Bose I, Chen X (2009) Hybrid models using unsupervised clustering for prediction of customer churn. *J Organ Comput Electron Commer* 19(2):133–151
- Dietterich TG (2000) Ensemble methods in machine learning. In: *MCS'00 proceedings of the first international workshop on multiple classifier systems*. Springer, London, pp 1–15
- Bauer E, Kohavi R (1999) An empirical comparison of voting classification algorithms: bagging, boosting and variants. *Mach Learn* 36(2):105–139
- Wang C-W, You W-H (2013) Boosting-SVM: effective learning with reduced data dimension. *Appl Intell*. doi:10.1007/s10489-013-0425-9
- Verikas A, Gelzinis A, Bacauskiene M (2011) Mining data with random forests: a survey and results of new tests. *Pattern Recognit* 44(2):330–349
- Xie Y, Li X, Ngai EWT, Ying W (2009) Customer churn prediction using improved balanced random forests. *Expert Syst Appl* 36(3):5445–5449
- Rodriguez JJ, Kuncheva LI, Alonso CJ (2006) Rotation forest: a new classifier ensemble method. *IEEE Trans Pattern Anal Mach Intell* 28(10):1619–1630
- Zhang C-X, Zhang J-S (2008) RotBoost: a technique for combining rotation forest and AdaBoost. *Pattern Recognit Lett* 29(10):1524–1536
- Bock KWD, Van den Poel D (2011) An empirical evaluation of rotation-based ensemble classifiers for customer churn prediction. *Expert Syst Appl* 38(10):12293–12301. doi:10.1016/j.eswa.2011.04.007
- Dietterich TG (2000) An experimental comparison of three methods for constructing ensemble of decision trees: bagging, boosting and randomization. *Mach Learn* 40(2):139–157
- Huang BQ, Kechadi TM, Buckley B, Kiernan G, Keogh E, Rashid T (2010) A new feature set with new window techniques for customer churn prediction in land-line telecommunications. *Expert Syst Appl* 37(5):3657–3665
- Huang B, Kechadi MT, Buckley B (2012) Customer churn prediction in telecommunications. *Expert Syst Appl* 39(1):1414–1425. doi:10.1016/j.eswa.2011.08.024
- Burez J, Van den Poel D (2009) Handling class imbalance in customer churn prediction. *Expert Syst Appl* 36(3):4626–4636. doi:10.1016/j.eswa.2008.05.027
- Owczarczuk M (2010) Churn models for prepaid customers in the cellular telecommunication industry using large data marts. *Expert Syst Appl* 37(6):4710–4712
- Peng H, Long F, Ding C (2005) Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans Pattern Anal Mach Intell* 27(8):1226–1238
- Sorokina D (2009) Application of additive groves ensemble with multiple counts feature evaluation to KDD cup '09 small data set. In: *JMLR workshop and conference proceedings*, Paris, France, June 28, 2009, vol 7, pp 101–109
- Vinh L, Lee S, Park Y-T, Auriol BD (2012) A novel feature selection method based on normalized mutual information. *Appl Intell* 37(1):100–120
- Li H, Wu X, Li Z, Wu G (2013) A relation extraction method of Chinese named entities based on location and semantic features. *Appl Intell* 38(1):1–15
- Breiman L (2001) Random forests. *Mach Learn* 45(1):5–32
- Zhang C-X, Wang G-W, Zhang J-S (2012) An empirical bias-variance analysis of DECORATE ensemble method at different training sample sizes. *J Appl Stat* 39(4):829–850
- Verbeke W, Dejaeger K, Martens D, Hur J, Baesens B (2012) New insights into churn prediction in the telecommunication sector: a profit driven data mining approach. *Eur J Oper Res* 218(1):211–229
- KDDCup 2009 challenge (2009) <http://kddcup-orange.com>
- The Center for Customer Relationship Management, Duke University. <http://www.fuqua.duke.edu/centers/ccrm/>
- Marquez-Vera C, Cano A, Romero C, Ventura S (2013) Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data. *Appl Intell* 38(3):315–330

35. Miller H, Clarke S, Lane S, Lonie A, Lazaridiz D, Petrovski S, Jones O (2009) Predicting customer behaviour: the University of Melbourne's KDD Cup report. In: JMLR workshop and conference proceedings, Paris, France, June 28, 2009, vol 28, pp 45–55
36. Busa-Fekete R, Keg1 B (2009) Accelerating AdaBoost using UCB. In: JMLR workshop and conference proceedings, Paris, France, June 28, 2009, vol 7, pp 111–122
37. Komoto K, Sugawara T, Tetu TI, Xuejuan X (2009) Stochastic gradient boosting. <http://www.kddcup-orange.com/factsheet.php?id=23>



Adnan Idris received his M.S. degree in Computer System Engineering from GIK Institute of Engineering Sciences and Technology Topi, Pakistan in 2006. Prior to that, he has earned his master degree in software engineering from COMSTATS Institute of IT, Islamabad in 2002. Further, he has more than 10 years of research & teaching experience at university level. Currently he is doing Ph.D. from Pakistan Institute of Eng. & Applied Sciences, Islamabad. His research areas include Customer Churn Prediction,

Machine Learning and Evolutionary algorithms.



Asifullah Khan received his M.Sc. degree in Physics from University of Peshawar, Pakistan in 1996 and his M.S. degree in Nuclear Engineering from Pakistan Institute of Engineering and Applied Sciences (PIEAS), Islamabad, Pakistan, in 1998. He received his M.S. and Ph.D. degrees in Computer Systems Engineering from Ghulam Ishaq Khan Institute of Engineering Sciences and Technology (GIK Institute), Topi, Pakistan, in 2003 and 2006, respectively. He has carried out two-years Post-Doc Research

at Signal and Image Processing Lab, Department of Mechatronics, Gwangju Institute of Science and Technology, South Korea. He has more than 15 years of research experience and is working as Associate Professor in Department of Computer and Information Sciences at PIEAS. His research areas include Digital Watermarking, Pattern Recognition, Image Processing, Evolutionary Algorithms, Bioinformatics, Machine Learning, and Computational Materials Science.



Yeon Soo Lee studied Metallurgy and Mechanical engineering in his B.S., M.S., and Ph.D. courses at Chonnam National University, Gwangju, Korea. During 2002–2004, He worked as a post-doc researcher for medical application of composite materials at Prof. Ben Goich's Lab, College of Manufacturing Technology, Nihon University, Chiba, Japan, and Clinical Orthopaedic Biomechanics at Prof. Thay Q. Lee's OBL Lab, Department of Orthopaedic surgery, UC Irvine, CA, USA. In 2004–2006, he

worked for clinical biomechanics and computer aided robot/navigation surgery at Chonnam National University Hwasun Hospital, Korea. Since 2007, he had worked at Gwangju Institute of Science and Technology as a research assistant professor. March 2010, he has joined Department of BioMedical Engineering, College of Medical Science, Catholic University of Daegu, Rep. Korea.