The Computer Journal Advance Access published January 31, 2016

doi:10.1093/comjnl/bxv123

Churn Prediction System for Telecom using Filter–Wrapper and Ensemble Classification

Adnan Idris[1,2] and Asifullah Khan[1]*

[1]Department of Computer and Information Sciences, Pakistan Institute of Engineering and Applied Sciences, P.O Box Nilore, Islamabad 45650, Pakistan
[2]Department of Computer Sciences and Information Technology, University of Poonch Rawalakot, Rawalakot, AJK Pakistan
*Corresponding author: asif@pieas.edu.pk

Churn prediction in telecom is a challenging data mining task for retaining customers, especially, when we have imbalanced class distribution, high dimensionality and large number of samples in training set. To cope with this challenging task of churn prediction, we propose a new intelligent churn prediction system for telecom, named FW-ECP. The novelty of the FW-ECP lies in its ability to combine both filter- and wrapper-based feature selection as well as exploit the learning capability of an ensemble classifier built using diverse base classifiers. In the filter phase, Particle Swarm Optimization-based undersampling and mRMR feature selection are employed to reduce the effect of imbalanced class distribution and large dimensionality. In Wrapper phase, we employ Genetic Algorithm that further discards irrelevant and redundant features. Random Forest, Rotation Forest, RotBoost and SVMs are then employed to exploit the new feature space. Finally, the ensemble classifier is constructed using both majority voting and stacking. We have tested and compared the performance of proposed FW-ECP system on two publicly available standard telecom datasets: Orange and Cell2Cell. FW-ECP takes into account both the imbalanced nature and large dimensionality of the training sets and yields better prediction performances compared with existing state-of-the-art approaches. The feature spaces for the Orange and Cell2Cell datasets are reduced to 24D and 18D, from 260D and 76D, respectively. The AUCs obtained by FW-ECP are 0.85 and 0.82 for Orange and Cell2Cell datasets, respectively.

Keywords: telecom churn prediction; Filter–Wrapper; particle swarm optimization; minimum redundancy and maximum relevance; Genetic Algorithm; ensemble classification

Received 3 May 2015; revised 8 December 2015
Handling editor: Munevver Kokuer

1. INTRODUCTION

Telecom industry has rapidly progressed during the last decade and the number of mobile phone customers has increased to 6.8 billion, which is approaching the global population [1]. This means saturation is expected in the number of mobile phone customer and the telecom companies are in constant competition to retain customers. Given the expected saturation, it is more important to retain customers than to acquire new ones. Moreover, customer churn results in disrepute for the company and ultimately leads to truncated profits, whereas long-term customers appear less costly, generate higher profits and may introduce new referrals as well [2]. Marketing practices also claim that acquiring new customer costs five to six times more than retaining an existing customer [3]. Therefore, an effective churn prediction system can play an important role by predicting churners. Consequently, the expected churners could be lured with attractive tariff packages before they decide to switch to a different network.

Telecom companies focusing on customer retention thus use churn prediction system that pinpoints the customers who are most likely to churn [4, 5]. These systems use customers' demographics, contract information, call detail records, services log, complaints, billing and account information, etc. [6]. Recently, a number of computational intelligence-based data mining techniques have been used for predicting churners in various application domains. Random Forest, an ensemble of Decision Trees, has been used in predicting churners in banking and publishing sectors, and has been reported to attain good prediction performance [7]. However, Random Forest suffers while predicting telecom churners, largely because it works by generating trees on bootstrap samples, which may not work well due to imbalanced class distribution of telecom datasets. A variant of Random Forest known as Improved Balanced Random Forest (IBRF) is reported to show improved performance in predicting churners from an imbalanced dataset of the banking sector [8]. However, in [8], the results are evaluated using small-sized dataset, whereas in telecom sector datasets are very large in size.

Rotation Forest is an ensemble that rotates the input data by splitting feature space in random subsets. Using this approach, it tries to achieve both diversity and accuracy within the ensemble [9, 10]. Rotation Forest is extensively used for various classification problems but like Random Forest and IBRF, this ensemble may also lack effectiveness in predicting telecom churners with desired accuracy [11]. Similarly, RotBoost is another ensemble of Rotation Forest and AdaBoost. RotBoost inherits iterative approach of AdaBoost to attain higher performance [12]. RotBoost and its variants have also been used for churn prediction in various application areas including telecom but they are also unable to predict churners in telecom with good accuracy [13]. Other machine learning techniques such as logit regression, Decision Trees and artificial neural networks (ANNs) have also been used in ensemble mode to predict telecom churners [14]. The proposed churn prediction system in [14] is evaluated on datasets having fewer features. However, in real world telecom churn prediction problems, both number of samples and dimensionality are high. Therefore, prediction system trained on a relatively small dataset might not work for real churn prediction application due to curse of dimensionality effect and less generalization, $generalization = f(N, l)$. Therefore, such a churn prediction solution cannot be appropriate for telecom sector where datasets have large dimensionality.

In a separate interesting study, variable selection and sampling are utilized by separate ensembles of ANNs and logit regression for churn prediction. The method discards nominal features and employs random sampling to balance the training dataset, which exposes this method to overfitting and inconsistency [15].

Ensemble approaches have gained reasonable attention in the recent past largely due to their capability to improve prediction performance of industrial learning systems, but predicting telecom churners with appreciated accuracy is a challenging task for ensemble classifiers [11–13]. Some of the studies have also incorporated social network analysis of customers' calling data to attain better results of customer churn prediction [16–19], whereas in [20] focus is made about reactive winback campaigns and customer acquisition strategies.

In [6], a new set of features was reported and evaluated using six modeling techniques to predict churners from land-line services dataset. Experimental results show improved performance of the proposed approach; however, these features could be used only to identify land-line churners. Most of the contemporary studies, which target customer churn prediction in telecom, suffer from the enormous nature of telecom dataset. The issues of high dimensionality, imbalanced nature and large size of telecom datasets, recognize telecom churn prediction as a different application domain for predicting churners compared with other subscription-based businesses. Normally, few instances of minority class, which represent churners in an original dataset, make it difficult for a classifier to attain balanced learning. In one of the existing techniques, random oversampling of the minority class and undersampling of the majority class is separately performed to cope with the dominance of majority class in the training phase [14]. Oversampling practice sometimes leads to biased training of the learner toward oversampled instances. Similarly, undersampling results in loss of information affecting the learning capability of the classifier.

Telecom operators archive comprehensive information about customers, ranging from personal demographics to details regarding usage of services, which eventually results into a larger dimensionality of the telecom datasets. The large number of features in telecom datasets poses the problems associated with the curse of dimensionality. Therefore, predicting customer churn in telecom is a challenging problem due to the large dimensionality and imbalanced class distribution of the telecom datasets. Customer churn prediction is generally perceived as a difficult data mining problem considering the complex nature of telecom datasets. Separate studies have individually focused on partial problems, faced in predicting telecom churners. Burez discusses the issues associated with the imbalance present in the training set and exploits four different classification methods to predict churners [21]. The study exposes how oversampling of instances can lead to overfitting, while undersampling can degrade classifier's learning. Similarly, a χ^2-based method is presented that selects the significant features and reduces the large dimensionality of the telecom dataset [22]. Decision Trees and Neural Networks are also used for churn prediction in combination with feature selection [23]. In another study, ANN, K-nearest neighbor, support vector machine (SVM) and decision trees are used in a hybrid methodology with extracting influential features to predict telecom churners [24]. A new set of features is presented with three-input-window technique to attain improved churn prediction accuracy in land-line services [25]. In a study presented in [6], seven classification methods are used with a set of transformed features to predict churners of land-line services. Various studies have thus highlighted the partial problems faced in telecom churn prediction by separately focusing on either imbalanced distribution or high dimensionality of the telecom dataset. In this regard, we observed that the problem of churn prediction in telecom may be better handled by simultaneously dealing with imbalance distribution as well as large dimensionality of the training set.

In this paper, our focus is to effectively handle the large dimensionality and imbalance distribution of the telecom dataset for improving the prediction performance of a churn prediction system. This objective is achieved by constructing a new hybrid approach based on combining Filter–Wrapper and ensemble classification. In the filter phase, PSO-based method undersamples the instances of majority class to reduce the effect of imbalanced nature of the dataset on learning capability of classifiers. Then, minimum redundancy and maximum relevance (mRMR) technique is used for reducing the dimensionality by extracting features having maximum discriminating information. In the second phase, a wrapper method comprising Genetic Algorithm (GA) coupled with a classifier is applied to remove any further redundancy from the feature space. Finally, the ensemble is constructed by separately combining the predictions of Random Forest, Rotation Forest, RotBoost and SVMs through both majority voting and stacking. The proposed system termed as FW-ECP is tested on two publicly available standard telecom datasets of different socioeconomic backgrounds.

The rest of the manuscript is organized as follows. Section 2 presents methods and material including the details of proposed churn prediction system. This section contains the subsections on Filter–Wrapper and ensemble classifier construction. Performance measures are described in Section 3. Section 4 explains the telecom datasets, presents experimental results and provides relevant discussions. This section also includes the performance analysis of proposed churn prediction system compared with other existing systems. Finally, conclusions are drawn in Section 5.

2. METHODS AND MATERIAL

In our proposed approach, our main focus is to target both the challenges of large dimensionality and imbalanced distribution, emerging from the avalanche of the ever-increasing information in the telecom datasets. The main idea behind the proposed hybrid approach (FW-ECP) is removing unnecessary and less informative features and intelligently settling the imbalance class distribution of the telecom dataset. The obtained refined training set provides improved learning, which is ultimately exploited by an ensemble constructed through the majority voting of diverse classifiers. A stacking-based ensemble is also constructed and its prediction performance is compared with majority voting-based ensemble.

2.1. Classification methods

In this section, we have discussed the classification methods used in proposed FW-ECP. Random Forest is one of the used classifiers, which is constructed by developing an ensemble of Decision Trees, involving random feature selection [26]. Decision Trees are generated over bootstrap samples of the training set. Final predictions are drawn by aggregating the predictions made by all individual trees, as represented in Fig. 1.

PSO-based sampling in our proposed FW-ECP system balances training set, which provides fair data distribution to Random Forest and eventually improved prediction performance is attained. Rotation Forest is a recently introduced ensemble classifier that simultaneously improves diversity and accuracy [9]. Higher diversity is achieved through rotating the features that are extracted by applying principal component analysis (PCA) on the input data. A rotation matrix is developed which extends K-axis rotation and develops a new set of attributes. Equation (1) shows the prediction made by Rotation Forest ensemble C^* for an instance X is a rotation matrix, computed for each of the feature subsets.

$$C^*(X) = \left(\underset{y\in\phi}{\arg\max}\right) \sum_{t=1}^{T} I(C_t(XR_t^a) = y) \qquad (1)$$

where $C_t(1,\ldots,T)$ represents the base classifier and y denotes the output in the form of 0 or 1, showing that an instance X is either a churner or non-churner.Rotation Forest has reported good performance on numerous problems [11, 13], mainly for its capabilities to simultaneously improve diversity and accuracy. Thus, strong classification capabilities of Rotation Forest are favorable for its use in telecom churn prediction. RotBoost is another classifier, developed by combining AdaBoost and Rotation Forest [12]. RotBoost inherits the weight updating process from AdaBoost, while rotation matrix is computed in a similar way as in Rotation Forest. RotBoost shows good performance on various problems compared with bagging, CART and C4.5 [27]. Therefore, RotBoost is also included as an ensemble member in our proposed FW-ECP system. Generally, the good performance of an ensemble is also attributed to the base learning algorithm [27]. Since Decision Trees are sensitive to variations in the learning data, they are considered suitable to be used as base learners in Random Forest, Rotation Forest and RotBoost.

Besides trees-based classifiers, SVM is also considered in the development of proposed churn predictor. SVM is a classification methodology based on statistical learning theory [28]. SVM develops a linear hyper plane, for maximizing the margin of separation between churners and non-churners. The objective is to search a hyper plane that expresses minimum classification error. Therefore, SVM expresses the binary classification as quadratic optimization problem. LIBSVM implementation in WEKA is used for SVM training and prediction [29]. RBF is used as kernel function for SVM training. RBF is presented in equation (2):

$$K(x_i, x_j) = \exp\left\{-\gamma \left\| x_i - x_j \right\|^2\right\} \qquad (2)$$

where 'γ' represents the width of the Gaussian function.

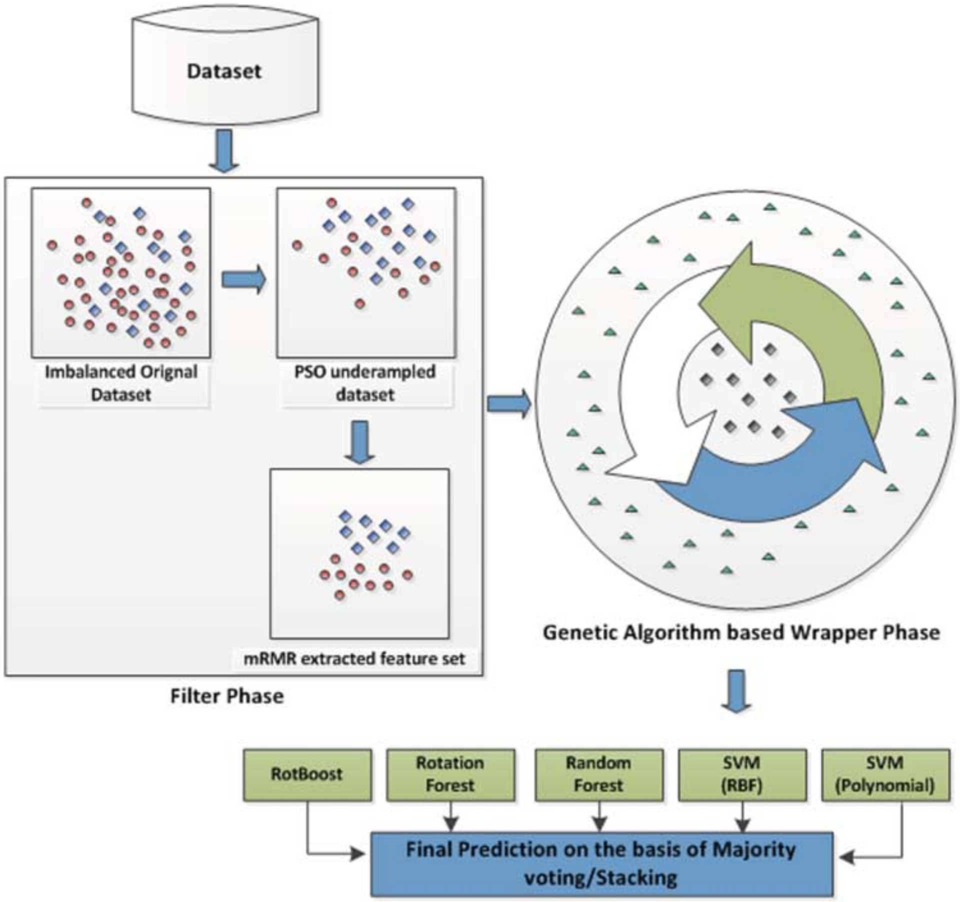

Dataset
Imbalanced Orignal Dataset
PSO underampled dataset
mRMR extracted feature set
Filter Phase
Genetic Algorithm based Wrapper Phase
RotBoost
Rotation Forest
Random Forest
SVM (RBF)
SVM (Polynomial)
Final Prediction on the basis of Majority voting/Stacking

FIGURE 1. Pictorial representation of Random Forest.

SVM with polynomial kernel of degree 2 is also employed to transform the features so that they could be made linearly separable in a bid to be used for churn prediction problem. Polynomial kernel function is expressed in equation (3):

$$K(x_i, x_j) = ((\phi(x_i).\phi(x_j)) + 1)^d \tag{3}$$

where d is the degree of the polynomial kernel that manipulates the decision boundary. In the current work, feature sets of the telecom datasets are significantly reduced through application of Filter–Wrapper method, whereas the number of instance is quite high; therefore, polynomial SVM can be considered as good choice among available kernels to be used for telecom churn prediction.

2.2. Proposed Filter–Wrapper-based Churn Prediction (FW-ECP) system

The proposed FW-ECP system is graphically represented in Fig. 2. The approach spans over two phases. The filter phase plays a significant role in handling the imbalanced distribution and high dimensionality of the telecom dataset, by employing PSO-based undersampling and mRMR feature extraction method, respectively. In the filter phase, the training set is initially preprocessed to handle the missing values and nominal values present in the telecom dataset. Then PSO-based undersampling methodology provides a balanced training set, subsequently used for extracting meaningful features. The method intelligently undersamples the majority class and selects those instances, which are useful in inducing the learner. Onward, mRMR method reduces the feature space to a set of discriminating feature set which lessens the computations and also extends improved learning to the classifiers. Usually, filter methods are meant to be used for the evaluation of feature subsets on the basis of certain criterion function such as Fisher's Ratio or other statistical measures [30]. On the contrary, in our proposed approach, a more natural and intuitive meaning of 'filter' is considered, which deals with filtering of irrelevant features from the training set and settles the imbalanced distribution as well.

In the second phase, GA with a classifier is used as wrapper to search the feature space and evaluate the goodness of selected feature subsets. The appropriateness of the selected feature subset is measured through performance obtained by the

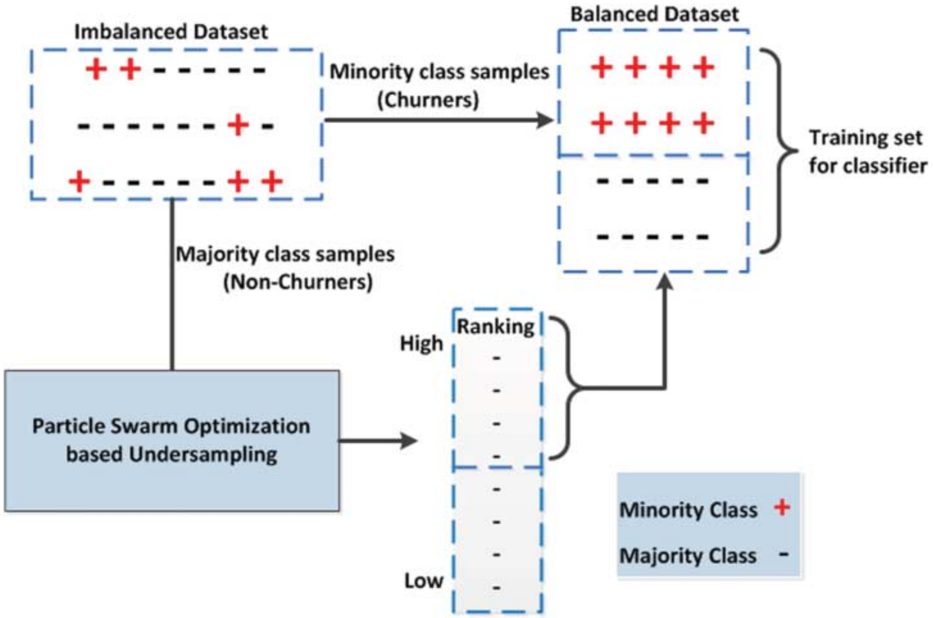


FIGURE 2. Graphical representation of the proposed FW-ECP system.

wrapped classifier. The wrapper method exploits GA's searching competitiveness and inductive learning process to evolve a subset of useful features, which helps the classifier in better learning the underlying patterns of churners. GA-based wrapper method further selects relevant features and thus ultimately enables majority voting-based ensemble to show improved prediction performance.

2.3. PSO-based undersampling

Undersampling is introduced in the filter phase of the proposed approach to handle the imbalance of training set. Imbalanced class distribution in the training set mostly causes the classification approaches to exhibit low prediction performance [8, 31–33]. Classification methods are biasedly trained when the training set is dominated by instances of a specific class. Telecom datasets generally comprise the majority of non-churner instances that result in classification approaches to show greater error in truly predicting the churners.

PSO-based undersampling [34] involves classification method to intelligently perform undersampling of the majority class. It recognizes a subsample of the dataset, which has maximum competency to extend meaningful information to the classifier. In addition, PSO-based undersampling method is an effective method compared with random undersampling, oversampling or synthetic oversampling method such as SMOTE, as reported in [33].

Figure 3 shows PSO-based undersampling process. The method develops a balanced dataset by finding and ranking the most informative instances of majority class and then combines these instances with minority class. Various subsets of the instances from the majority class are chosen and combined with the instances of minority class. The goodness of each subset is evaluated on the basis of area under the curve (AUC) attained by the classifier. Decision Trees and SVM classification methods are separately utilized to evaluate the candidate subsamples in PSO-based undersampling. This process ultimately builds a near-optimal balanced training set that considerably contributes in building improved churn prediction model for telecom by mitigating the imbalance of the training set.

PSO considers the subsets of instances as particles and steers these particles in the search space. The particles are moved and guided by their own best positions and swarm's best positions. The improved positions attained by the particles also guide the complete swarm to move.

Let us consider a population of n particles, where i is the index of a particle in the swarm $(i = 1, 2, 3, \ldots, n), j$ is the index of dimension $(j = 1, 2, 3, \ldots, m)$ and t shows the counter of iterations. The iterative process continues until a particle achieving good AUC is selected. The velocity $v_{i,j}(t)$ of ith particle and its position $x_{i,j}(t)$ is updated using the following equations:

$$v_{(i,j)}(t+1) = w.v_{i,j}(t) + c_1R_1 \cdot (pbest_{i,j} - x_{i,j}(t)) + c_2R_2(gbest_{i,j} - x_{i,j}(t)) \quad (4)$$

$$x_{i,j}(t+1) = \begin{Bmatrix} 0 & if\ random() \geq S(v_{i,j}(t+1)) \\ 1 & if\ random() < S(v_{i,j}(t+1)) \end{Bmatrix} \quad (5)$$

$$S(v_{i,j}(t+1)) = \frac{1}{1 + e^{-v_{i,j}(t+1)}} \quad (6)$$

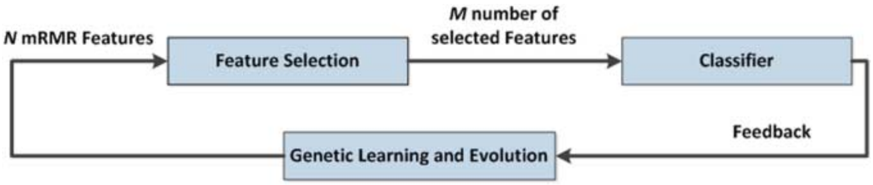


FIGURE 3. PSO-based undersampling model.

where $pbest_{i,j}$ and $gbest_{i,j}$ are the previous best position and the best position established by informers, respectively. c_1 and c_2 are cognitive and social accelerators, respectively, R_1 and R_2 are random numbers between 0 and 1, and w is the inertia weight.

Finally, majority class instances are ranked based on their selection frequency in model building as shown in Fig. 3. The highest rank is assigned to the instance that is selected more frequently in evaluating its prediction capability. Instances with higher ranks in the frequency list are combined with the instances of minority class to develop a balanced training set. PSO balanced training set enables the predictors (Random Forest, Rotation Forest, RotBoost and SVMs) to attain improved learning. Therefore, PSO balanced training set is further investigated for extracting features having maximum discriminating information and minimum redundancy.

PSO-based undersampling is implemented by considering particle space as m-dimensional space, equal to the size of majority class instances in the training set. Parameter setting of the PSO-based undersampling is accomplished empirically after conducting multiple runs and the best parameter values are fixed as presented in Tables 1–3.

2.4. mRMR-based feature selection

mRMR-based feature selection method extracts the subset of features that have a higher dependency on the class labels [34]. In order to reduce the redundancy, those instances are selected which are far away from each other. In this way, mRMR adopts broad criteria for feature selection based on mRMR. The maximum relevance is implemented with the help of the expressions given in equations (7) and (5):

$$\max D(S,c),\ D = \frac{1}{|S|}\sum_{x_i \in S} I(x_i;c) \tag{7}$$

$$\max D(S,c),\ D = I(\{x_i,\ i = 1,\ldots,m\};c) \tag{8}$$

where D is a 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 label c. The maximum relevance is sorted out by searching the feature set that satisfies the criteria in equation (7) and approximates the $D(S,c)$ in equation (8) with the mean value of all mutual information values between individual feature x_i and class c. A feature set S is chosen where the features have higher dependency on the respective class labels. Once the maximum relevant features are selected, there can be redundancy between them; therefore, one of the two such redundant features is removed, which would not change the discriminating power of the selected feature set. The

TABLE 1. Parameter setting of PSO-based undersampling method, using Decision Trees as classifier.

Used classification approach	Decision Tree
Size of particle population	150
Fitness evaluation criteria	AUC
Iterations	200
Update rule	Sigmoid function
Cognitive constant	1.23
Social acceleration constant	1.23
Inertia weight	0.69

TABLE 2. Parameter setting of PSO-based undersampling method, using SVM(RBF) as classifier.

Used classification approaches	SVM(RBF)
Size of particle population	150
Fitness evaluation criteria	AUC
Iterations	200
Update rule	Sigmoid function
Cognitive constant	1.49
Social acceleration constant	1.49
Inertia weight	0.68

TABLE 3. Parameter setting of PSO-based undersampling method, using SVM(Poly) as classifier.

Used classification approaches	SVM(Poly)
Size of particle population	150
Fitness evaluation criteria	AUC
Iterations	200
Update rule	Sigmoid function
Cognitive constant	1.43
Social acceleration constant	1.43
Inertia weight	0.68

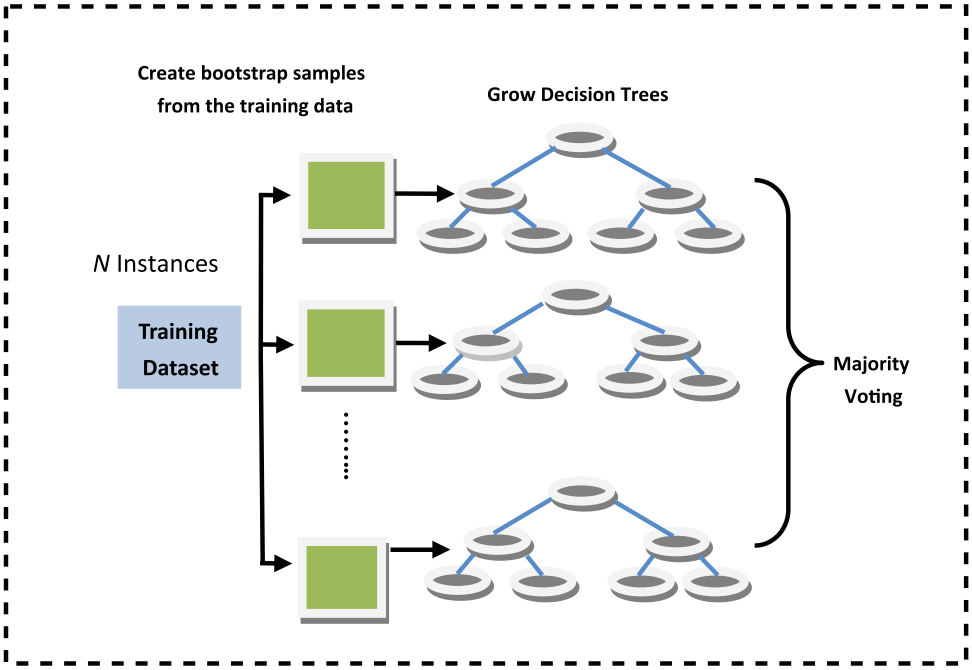


FIGURE 4. Feature extraction using classifier and feedback learning mechanism.

expression given in equation (9) helps in minimizing the redundancy among features:

$$\min R(S), \quad R = \frac{1}{[S]^2} \sum_{x_i x_j \in S} I(x_i, x_j). \tag{9}$$

The criteria of minimizing redundancy and maximizing relevance are then combined in one simple form, as given in equation (7), which is used to optimize both D and R, simultaneously.

$$\max \phi(D,R), \quad \phi = D - R. \tag{10}$$

The feature set obtained using mRMR is expected to be near-optimal, due to strong relevance of features with class targets and minimum redundancy between them. The features are descendingly ordered on the basis of mRMR criteria. Finally, a feature set is obtained having maximum mutual information that is selected for further investigations. The mRMR method reduces the feature space and serves the notion of providing meaningful features to the classifier for attaining improved learning.

2.5. GA-based Wrapper method for further feature selection

In order to optimize the prediction performance aimed in churn prediction, the feature selection can be vital. Relevant features having maximum useful information extend better learning, which ultimately improves the classifier's performance. A GA-based wrapper method is developed, by coupling feature selection with classification technique. A prediction feedback mechanism is incorporated that adjusts the selected features and ultimately a small feature subset is evolved that extends the maximum prediction accuracy. The wrapper method selects the features based on genetic learning and evolution.

The objective of the wrapper method is to further select a reduced set of features from the mRMR feature space, such that only useful features are included and features that contribute less in the learning process are excluded. Thus, mRMR feature space is further transformed into a new reduced feature space. This offers better separation of underlying pattern which in turn improves the prediction performance of classifier. Decision Trees and SVMs have been separately used to evaluate the fitness of feature subsets in GA-based wrapper method. AUC criterion is followed in evaluating the fitness of feature subsets. GA-based wrapper method uses performance of Decision Trees as a fitness measure in selecting the feature subset. The selected feature subset is subsequently used by Decision Trees-based predictors such as Random Forest, Rotation Forest and RotBoost. Similarly, SVM's performance is considered as a fitness measure to select the feature subset that is subsequently used by SVM-based predictors. The wrapper method finds a near-optimal transformation of mRMR feature space that results in both the lowest prediction error and smallest feature subset.

Figure 4 shows the working of GA-based wrapper method. Decision Trees are considered as a sensitive classifier with respect to the variation in the dataset [13]. Therefore, Decision Trees are employed to evaluate the prediction performance obtained on various feature subsets in GA iterations. Moreover, Decision Trees are also considered in wrapper method because they are used a base classifier in Random Forest, Rotation Forest and RotBoost, which are subsequently employed as predictors in our work. Similarly, SVMs are also used to evaluate the various feature subsets, so that selected feature subset may contribute significantly in inducing improved learning to SVM-based predictors. Each candidate feature subset is considered as chromosome, and its fitness is evaluated on the basis of AUC. The prediction performance of each candidate feature subset is evaluated using 5-fold cross-validation. In each generation, chromosomes are subject to the operations such as crossover, mutation and inversion. The best chromosomes are allowed to move to the next generation and within a specified number of generations, a near-optimal feature subset is evolved that attains highest AUC. The best feature set comprises the features that significantly contribute in attaining higher performance. GA-based wrapper method is developed using the implementation available in the WEKA data mining tool [29].

2.6. Developing churn predictor

Several classification techniques have been recently used for predicting churn in various areas. In this work, we have proposed churn prediction system FW-ECP, based on exploiting the capabilities of PSO-based undersampling and GA-based wrapper method to handle the issues regarding imbalanced class distribution and high dimensionality of the telecom datasets. Finally, diverse tree-based classifiers and SVMs are combined using majority voting and stacking methods, separately, to construct a predictor that can yield good prediction performance.

2.6.1. Majority voting-based ensemble

Majority voting is employed to combine the predictions of Random Forest, Rotation Forest, RotBoost and SVMs to construct an ensemble. The final decision is made in favor of a class that wins majority votes from the ensemble members. Mathematically, the majority voting can be expressed as given in equation (11):

$$class(x) = \underset{c_i \in dom(y)}{\arg\max} \left(\sum_k g(y_k(x), c_i) \right) \tag{11}$$

where $y_k(x)$ is the classification of the k'th classifier and $g(y, c)$ is an indicator function defined in equation (12).

$$g(y, c) = \begin{Bmatrix} 1 & y = c \\ 0 & y \neq c \end{Bmatrix} \tag{12}$$

Majority voting is one of the effective methods used to combine the output of various classifiers to construct an ensemble. Therefore, majority voting is applied in this work that produces an ensemble attaining improved prediction performance on telecom datasets.

2.6.2. Stacking-based ensemble

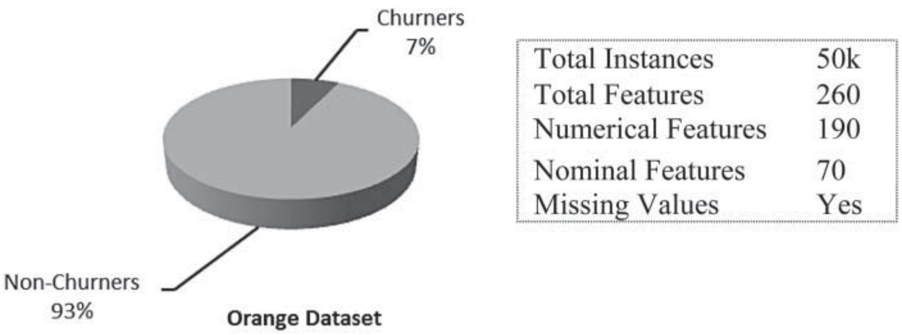
Besides majority voting, stacking is also used as an ensemble generation technique in this work. Stacking combines multiple base classifiers, which are trained on a single dataset using different learning algorithms such as $l_1, l_2, \ldots, l_N$ [35]. A set of base classifiers $c_1, c_2, \ldots, c_N$ is developed in the first phase and then a meta classifier is formed, which combines the outputs of the base classifiers in the second phase. A training set is developed for the meta classifier, combining the predictions obtained by base classifiers using 10-fold cross-validation. Thus, the training set involved in meta learning consists of examples in the form of $((y_i^1 \cdots y_i^n), y^i)$ where predictions obtained by base classifiers are features.

3. PERFORMANCE EVALUATION

Once an ensemble classifier is constructed, it is consequently used to make predictions regarding future conduct of the customers. In order to ensure that the proposed system exhibits good performance, predictions made by FW-ECP have to be assessed. AUC, sensitivity, specificity and F-measure are followed to conduct performance evaluation of the proposed system. These measures are commonly followed in contemporary literature to evaluate the performance of various classification methodologies used for churn prediction [2, 6, 36, 37]. In addition, lift curves are also drawn to evaluate the performance of proposed FW-ECP approach. Therefore, these performance measures are considered in this study so that we can conduct the performance comparison of our proposed FW-ECP system with other existing approaches.

3.1. Statistical measures

Area under a Receiver Operating Characteristics curve, known as AUC, is an effective scalar measure considered to evaluate various classification-based techniques for churn prediction [2, 21]. A perfect classifier has AUC 1, whereas a random classifier has AUC 0.5. Thus, a classifier attaining AUC close to 1 is considered as good performer. In customer churn prediction, TP shows the number of correctly predicted churners, FN shows the number of incorrectly predicted non-churners, TN shows the number of correctly predicted non-churners and FP shows the number of incorrectly predicted churners. TP + FN = P and TN + FP = N. The predictor assigns TP + FP examples to the churner class and TN + FN examples to the non-churner class. Using this information, we can define mostly used performance measures as follows: specificity = NT/N → 1-specificity = FP/N = FPrate, sensitivity = TP/P = TPrate = recall, Yrate = (TP + FP)/(P + N), precision = TP/(TP + FP), F-measure = 2 × ((precision × recall)/(precision + recall)) and accuracy = (TP + TN)/(P + N). Similarly, lift is defined as, lift = precision/(P/(P + N)) = sensitivity/Y rate, which makes comparison



Total Instances	50k
Total Features	260
Numerical Features	190
Nominal Features	70
Missing Values	Yes

FIGURE 5. Characteristics of the Orange dataset.

between precision and overall churn rate in the test set. AUC is computed with the help of the formula given in equation (13) [21]:

$$\mathrm{AUC} = \int_0^1 \frac{TP}{p} d\frac{FP}{N} = \frac{1}{p.N}\int_0^N TPdFP \qquad (13)$$

4. RESULTS AND DISCUSSION

This section presents the details of the used datasets in the current study and results obtained by the proposed FW-ECP, along with comparisons made with other relevant studies. A detailed comparative analysis is also presented to highlight the performance contribution made by each method in FW-ECP to attain an improved prediction performance. The results are evaluated on the basis of 10-fold cross-validation, which effectively measures the generalization capabilities of the predictor. The dataset $D$is decomposed into $D_1, D_2, D_3, \ldots, D_{10}$ parts. $D - D_i$-folds are used to train the majority voting-based ensemble and D_i fold to test the ensemble, in each iteration ($i = 1$ to 10). Finally, the results generated in each iteration are accumulated. Ten-fold cross-validation involves increasing computations considering the large size of telecom datasets but results obtained through this method are more reliable and generalized. Two of the publicly available telecom datasets are used in this work to evaluate the performance of proposed churn prediction system.

4.1. Datasets

Public telecom datasets that can be used for churn prediction are scarcely available due to privacy of the customers. Contemporary research works on telecom churn prediction only explain the characteristics of the used telecom datasets and then present the analytical view of the performance obtained by predictors [2, 6, 8, 13]. Figure 5 shows the details of the Orange dataset that is quite large in size [38]. Eighteen of the features in the Orange dataset have no value at all and five of the features have only one value, and such useless features are discarded. Cell2Cell is the other dataset that is provided in balanced form by Duke University [2]. Figure 6 represents the details of the Cell2Cell dataset. The features having nominal values in both the datasets are converted to numerical format. This is accomplished by grouping the modalities in small, medium and large categories, depending on the number of instances in each category [39].

4.2. Performance analysis without Filter–Wrapper phase

Initially, we applied Random Forest, Rotation Forest, RotBoost and SVMs on the original form of the telecom datasets to evaluate the prediction performance. Tables 4 and 5 report the performance obtained by the used classifiers. It is clearly observed that all the used classifiers suffer from attaining good prediction performance on both datasets. Nonetheless, majority voting of the applied classifiers shows improvement in predicting churners compared with the performance shown by an individual classifier. However, the performance improvement is not satisfactory enough that may encourage using only majority voting of the base classifiers as a solution to telecom churn prediction. Such a deteriorated performance hints at employing effective preprocessing on the training dataset that could establish a favorable feature space and class distribution for the classifiers.

In the case of Orange dataset, the low sensitivity scores obtained by the predictors, as shown in Table 4, indicate the dominance of non-churners (92%) in the dataset, which leads to the biased training of the used classifiers. Moreover, SVMs also show deteriorated performances as the drawn hyper planes are unable to classify churners and non-churners with desired accuracy from the imbalanced dataset. However, in the case of Cell2Cell dataset where training set is provided in balanced form, an improved prediction performance is shown by the used classifiers as given in Table 5. Although the used classifiers are unable to attain higher prediction accuracy on Cell2Cell dataset as well due to the presence of less informative and redundant features in the training set, the balanced class distribution of Cell2Cell dataset considerably impacts the discriminating learning during the training phase. The figures

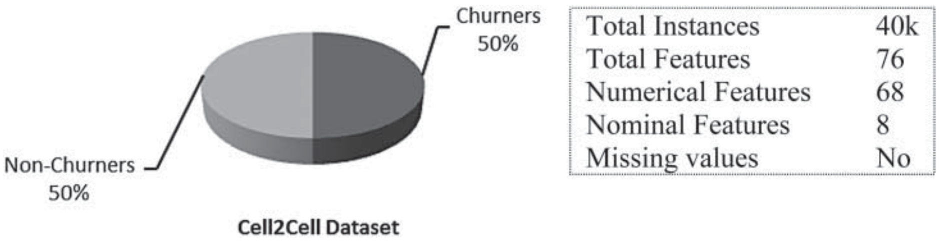


FIGURE 6. Characteristics of the Cell2Cell dataset.

TABLE 4. Performance evaluation on Orange datasets.

	Orange datasets		
	AUC	Sensitivity	Specificity
SVM(RBF)	0.301	0.0199	0.0214
SVM(Polynomial)	0.310	0.0214	0.3612
Random Forest	0.571	0.0049	0.9991
Rotation Forest	0.583	0.0026	0.9998
RotBoost	0.601	0.0291	0.7212
Majority voting-based ensemble	0.614	0.0301	0.8011

TABLE 5. Performance evaluation on Cell2Cell dataset.

	Cell2Cell dataset		
	AUC	Sensitivity	Specificity
Random Forest	0.592	0.690	0.601
Rotation Forest	0.610	0.666	0.646
SVM(RBF)	0.622	0.671	0.644
SVM(Polynomial)	0.645	0.681	0.644
RotBoost	0.699	0.684	0.646
Majority voting-based ensemble	0.704	0.699	0.681

shown in this work use ‘RandFor’ to represent Random Forest, ‘RotFor’ to represent Rotation Forest and ‘MVbasedEn’ to represent majority voting-based ensemble. Figure 7 shows that used classifiers attain improved performance on Cell2Cell dataset, unlike Orange dataset for the reasons of balanced class distribution of data and limited dimensionality. This supports the notion that balanced distribution of training set and selected discriminative feature set induce the classifiers with improved learning. Moreover, it is a better approach to select fewer discriminative features having maximum useful information compared with using large number of features with irrelevant information and less discriminating power.

Therefore, a Filter–Wrapper approach is essentially introduced in proposed FW-ECP system which supports the used classifiers to attain improved prediction performance through effectively handling the enormous characteristics of the telecom datasets.

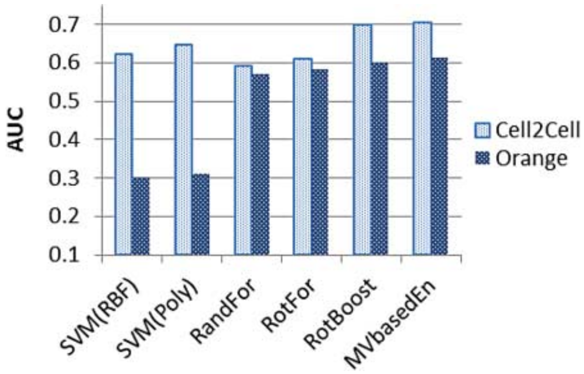


FIGURE 7. Prediction performance of predictors on Orange and Cell2Cell datasets.

4.3. Performance improvement using filter phase

The filter phase is introduced in the proposed churn prediction system to specifically address the issues associated with avalanche of the increasing information in the telecom datasets that leads to the imbalanced class distribution and large dimensionality.

4.3.1. Balanced learning using PSO-based undersampling
In the first step as shown in Fig. 2, filter phase employs a PSO-based undersampling method that settles the imbalance of the training set, through intelligently selecting useful instances of the majority class that are equal in number to minority class. Orange dataset has fewer instances of the minority class. In the complete dataset of 50 000 instances, 3276 churner instances are present. Figure 8 shows PSO-based method treats the imbalance of the Orange dataset.

Figure 9 shows that predictors used in this work attain improved performance once trained on the PSO balanced Orange dataset. It indicates that balanced data distribution between the classes in the training set plays a significant role in inducing the classifier with unbiased learning.

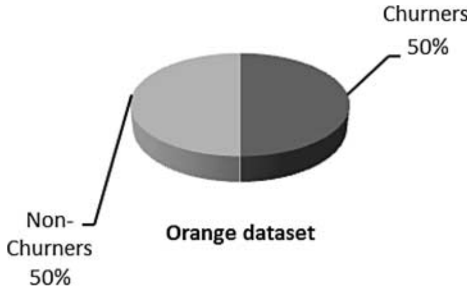


FIGURE 8. Balanced Orange dataset, using PSO-based undersampling method.

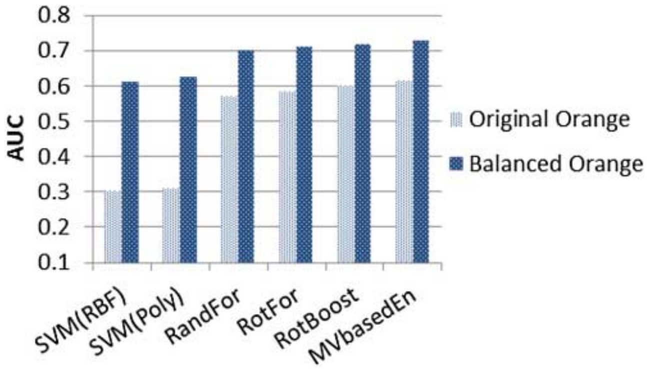


FIGURE 9. Performance improvement attained after applying PSO-based undersampling on Orange dataset.

Henceforth, PSO balanced dataset is used for feature extraction and further investigations. The Cell2Cell dataset is available in a balanced form. Therefore, it is not processed with PSO-based undersampling. PSO-based undersampling is preferred for its capabilities to exploit the original information and eliminate the instances of majority class, which are found less useful on the basis of their evaluated fitness. In contrast, in oversampling, instances of the minority class are duplicated, which sometimes leads to overfitness [21].

4.3.2. Average performance of PSO-based undersampling

Figure 10 shows the graph that represents the average performance of PSO-based undersampling when AUC of Decision Trees is used as the fitness evaluation criterion.

Each PSO-based simulation executes 200 iterations with the empirically fixed parameters given in Table 1. The best AUC value evolved in each simulation run is plotted in the graph, as shown in Fig. 10. The AUC values attained by Decision Trees, used as fitness evaluation criteria in the internal optimization of PSO-based undersampling, range from 0.62 to 0.70. Figure 10 shows the deviation in AUC over 30 PSO simulations. Error bars show the deviation of AUC from the mean value. Error bars are drawn using standard deviation 1 on both sides of the mean value. Sampling process involves internal 3-fold-stratified cross-validation for each of the classification methods used in creating balanced training set.

4.3.3. Performance improvement using mRMR

A number of experiments show that reduced feature spaces obtain better classification accuracy and faster computation compared with prediction results obtained using original features [40]. mRMR method is employed in filter phase as shown in Fig. 1 to select the most useful features that induce the classifiers with improved learning. mRMR has proved effective in extracting useful features that retain maximum discriminative information and extend better learning to classifiers [21]. Primarily, mRMR method improves interclass and intraclass proximities of the data instances through finding mutual dissimilarity between instances and their dependency on class labels. mRMR method extracts the minimum redundant features having maximum relevance with class labels. Numbers of mRMR features showing maximum performance with the specific classifier are selected using exhaustive search. Tables 6 and 7 report the prediction results obtained on PSO balanced Orange and Cell2Cell datasets, respectively. It is observed from the simulation results that RotBoost and majority voting-based ensemble obtain improved prediction performance using features selected through mRMR method. Thus, mRMR feature set not only helps in improved learning of the classifiers but also lessens the computations by extracting only meaningful features.

4.4. Performance improvement using wrapper phase

In the wrapper phase, we employ GA in conjunction with classification method. The appropriateness of the various feature subsets evolved in GA's iterations is evaluated on the basis of AUC value attained by classifier. mRMR transformed feature space offers good discriminating information that is further explored by GA-enabled wrapper method to remove any fine level of redundancy in features, which may be less useful in inducing meaningful learning. Hence, GA-based wrapper selects the most succinct features from mRMR transformed feature space. GA's searching competitiveness is exploited to further squeeze the feature space and select relevant and minimum number of features having sufficient discriminating power for attaining good prediction accuracy.

mRMR method extracts a set of 36 features from PSO balanced Orange dataset and a set of 31 features from Cell2Cell dataset that allows majority voting-based ensemble to attain highest prediction among other used individual predictors, as given in Tables 6 and 7. These feature sets (36 features from PSO balanced Orange dataset and 31 features from Cell2Cell) are further investigated by GA-based wrapper method to discard any fine level of irrelevance or redundancy in the feature space. Table 8 reports the performance of GA-based wrapper method, on PSO balanced Orange and Cell2Cell datasets. It is observed from simulation results that a set of only 24 of

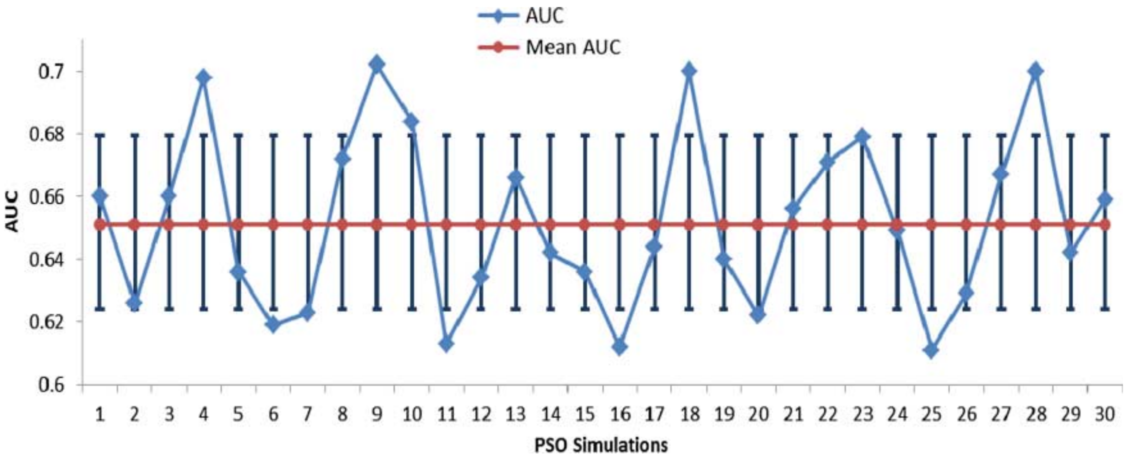


FIGURE 10. Average performance of PSO-based undersampling.

TABLE 6. Performance evaluation on selected mRMR features of balanced Orange dataset.

	Balanced Orange dataset	
	Number of features	AUC
SVM(RBF)	67	0.663
SVM(Polynomial)	61	0.678
Rotation Forest	39	0.701
Random Forest	36	0.751
RotBoost	35	0.761
Majority voting-based ensemble	36	0.781

TABLE 7. Performance evaluation on selected mRMR features of Cell2Cell dataset.

	Cell2Cell dataset	
	Number of features	AUC
SVM(RBF)	68	0.734
SVM(Polynomial)	63	0.741
Rotation Forest	35	0.762
RotBoost	31	0.816
Majority voting-based ensemble	31	0.836

the mRMR features show highest fitness for inducing the Decision Trees. Similarly, feature sets comprising 34 features and 36 features show highest fitness using SVM(RBF) and SVM(Poly), respectively. Thus, GA-based wrapper method reduces the feature space to a set of only fewer meaningful features which show maximum performance with respect to a wrapped classifier as given in Table 8. In the case of Cell2Cell dataset, GA-based wrapper method concludes only 18, 28 and 31 features using Decision Trees, SVM(RBF) and SVM(Poly), respectively.

4.4.1. *Average performance of GA-based wrapper method*

Figure 11 shows the graph that represents the average performance of GA-based wrapper method. The graph includes 30 runs of GA-based simulation, which are conducted using Decision Trees' AUC as fitness measure. The population size and maximum generations are set to 100 and 50, respectively. Probabilities for crossover and mutation are empirically set at 0.60 and 0.033, respectively. Each GA simulation searches a reduced feature set using AUC performance attained by the wrapped classifier. It can be observed from Fig. 12 that fitness function ranges between 0.61 and 0.76 AUC. Error bars using standard deviation 1 are drawn in Fig. 11 that represents the deviation in AUC from mean value in each GA simulation run. GA-based wrapper method selects a set of relevant features from both the training sets, which helps in attaining improved performance.

4.5. Performance improvement using ensemble classification

A majority voting-based ensemble is constructed by combining the predictions of Random Forest, Rotation Forest, RotBoost and SVMs. Figures 12–14 show the improvement attained by the employed classifiers after the application of filter and wrapper phases on training sets. The classifiers have shown gradual improvement in prediction performance, when imbalanced class distribution and high dimensionality of Orange dataset are treated by hybridizing filter and wrapper phases. Experimental results given in Figs. 12–14 show that PSO-based undersampling and mRMR-based feature extraction have considerably contributed in ordering the training set to extend improved learning. Moreover, GA's efficient searching capabilities in wrapper phase search the most relevant features

TABLE 8. Number of features extracted from GA-based wrapper method.

	Balanced Orange dataset	Cell2Cell dataset
Number of mRMR feature	36	31
Number of features extracted through GA-based wrapper method using Decision Trees	24	18
Number of features extracted through GA-based wrapper method using SVM(RBF)	34	28
Number of features extracted through GA-based wrapper method using SVM(Poly)	36	31

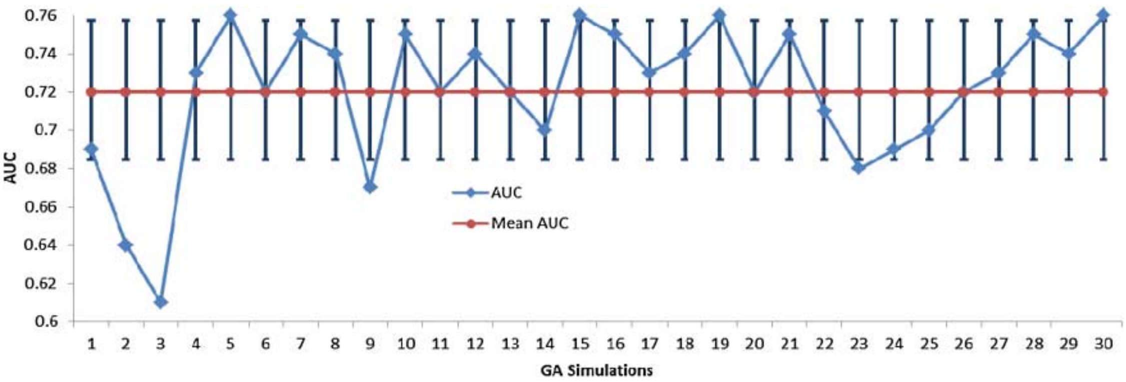


FIGURE 11. Average performance of GA-based wrapper method.

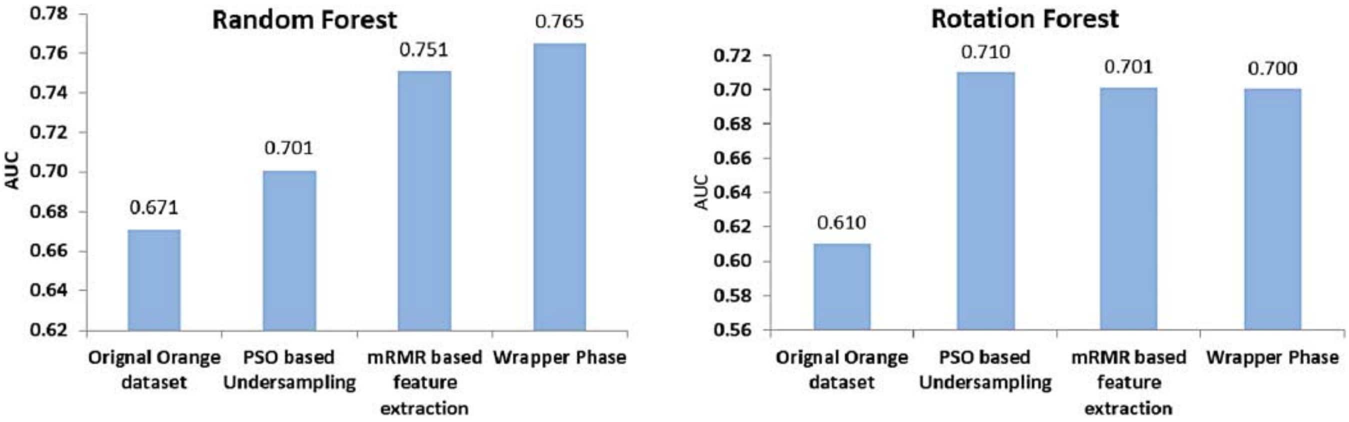


FIGURE 12. Performance improvement attained by Random Forest and Rotation Forest.

having maximum information that ultimately enable FW-ECP to exhibit better performance in predicting churners.

4.5.1. *Performance analysis on Orange dataset*

Orange dataset has complex nature for its highly imbalanced class distribution and large dimensionality. *Random Forest* initially achieves 0.671 AUC score for predicting churners from the original Orange dataset as shown in Fig. 12. The original Orange dataset has only fewer instances of churner's class compared with the number of non-churners. Random Forest may be susceptible to biased learning when training set is highly imbalanced like in the case of Orange dataset. The experimental results given in Fig. 12 show that Random Forest attains improved performance once the imbalance and high dimensionality of the training set are appropriately handled. An accuracy of 0.765 AUC is attained by Random Forest on Orange dataset when presentation of training set is improved through filter and wrapper phases.

Rotation Forest also improves prediction performance with the availability of balanced training set as shown in Fig. 12. Rotation Forest attains an accuracy of 0.710 AUC when provided with PSO balanced training set. However, it suffers

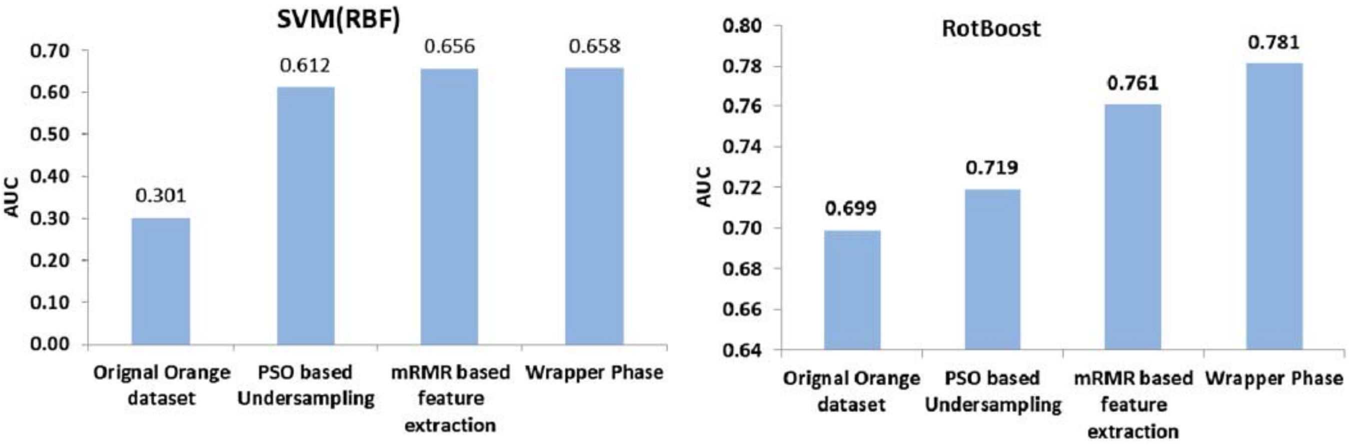


FIGURE 13. Performance improvement attained by RotBoost, SVM(RBF).

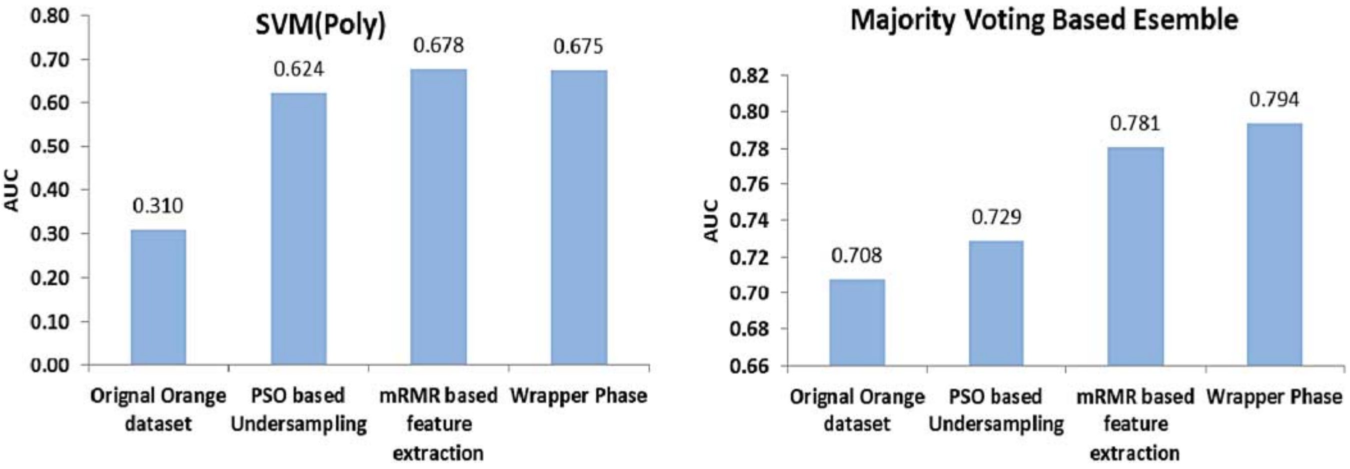


FIGURE 14. Performance improvement attained by SVM(Poly) and majority voting-based ensemble.

some level of deterioration in performance when provided with mRMR extracted feature set.

Similarly, wrapper phase also does not contribute much in improving the prediction performance of Rotation Forest. Such deterioration in performance is primarily because of undermining inbuilt comprehensive feature extraction process of Rotation Forest.

Rotation Forest encourages diversity by using PCA for extracting a number of feature subsets that are used to construct rotation matrix, responsible for rotating input data space. Further, Rotation Forest is also based on utilizing all the extracted principal components to attain maximum accuracy. But, Rotation Forest's own feature extraction process is undermined when mRMR method is applied for feature extraction. This also compromises the higher diversity targeted by Rotation Forest. Although, most relevant features are selected by GA-based wrapper method for boosting the performance but it does not contribute in improving Rotation Forest's prediction performance.

Rotation Forest shows maximum performance when PSO balanced dataset is provided for training because effective feature extraction process is already deployed within it, and by constructing rotation matrix attains maximum diversity and accuracy. Therefore, mRMR feature extraction method and wrapper phase have not much contributed in improving the Rotation Forest's performance.

RotBoost combines Adaboosting and Rotation Forest to attain the maximum performance. AdaBoost part of the RotBoost ensures to tackle the hard instances in iterative approach and attains maximum performance, whereas Rotation Forest operates by constructing a rotation matrix ensuring higher diversity. The experimental results shown in Fig. 13 clearly indicate that RotBoost gradually improves the performance as the training set is appropriately treated in filter and wrapper

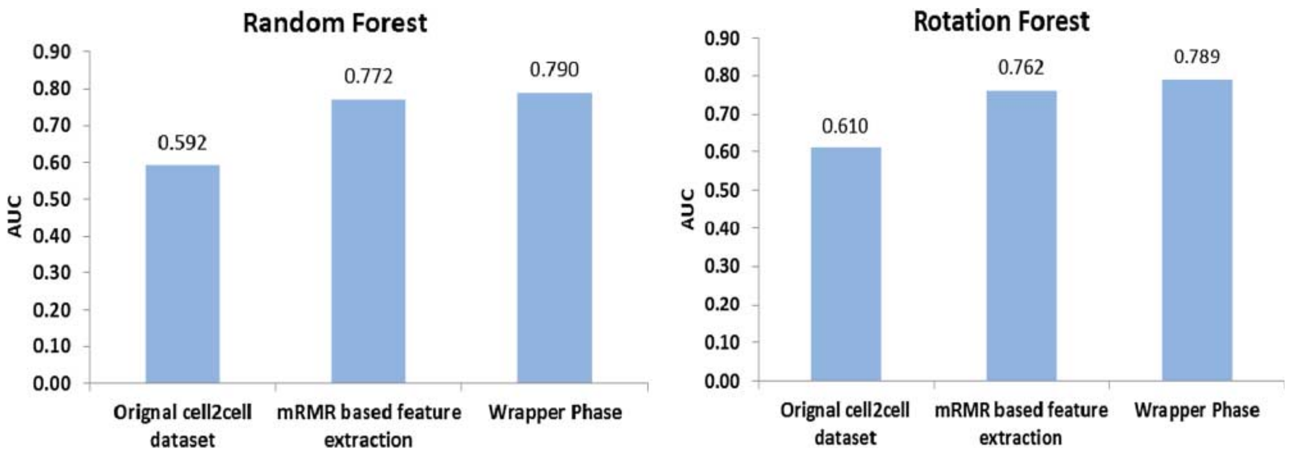


FIGURE 15. Performance improvement attained by Random Forest and Rotation Forest.

phases. RotBoost initially achieves 0.699 AUC on the original form of Orange dataset. Thereafter, the prediction performance gradually improves to 0.781 AUC, when wrapper phase extracts a set of 24 features from mRMR-transformed feature space. RotBoost achieves the improved performance for the Adaboosting involved in the algorithm. Adaboosting adopts an iterative process that handles the hard instances by assigning weights, which results in improved performance.

SVM(RBF) and SVM(Poly) also show gradual improvement in the performance of Orange dataset as shown in Figs. 13 and 14, respectively, but the overall performance is low compared with other used classifiers. SVM(RBF) uses Gaussian in the kernel function, where output of the kernel is dependent on the Euclidean distance of instances from support vector. SVM(Poly) is useful in situations where the dimensionality of the feature space is low and the number of instances is relatively higher. In our case, filter and wrapper phases reduce the dimensionality of the Orange dataset, whereas the number of instances is more compared with features but still SVM(Poly) suffers to attain improved performance. In SVM(Poly), '*d*' parameter is the degree of the polynomial kernel which controls the flexibility of the classifier. Thus, kernel parameters considerably affect the performance of SVM(Poly). SVMs attain below power performance primarily due to the absence of optimized values of kernel parameters.

Finally, a *Majority Voting-based Ensemble* is constructed by combining the predictions of all the base classifiers. Majority voting is a simple and effective method that strengthens the prediction process by assigning the target label to an instance that wins the maximum votes. Figure 14 shows that majority voting-based ensemble improves the prediction performance on Orange dataset and finally achieves 0.794 AUC. This performance, to our best of knowledge, is the highest prediction performance so far reported on Orange dataset. Majority voting of Random Forest, Rotation Forest, RotBoost and SVMs constructs a high performing ensemble that presents an effective solution for churn prediction in telecom. Moreover, the hybridization of filter and wrapper phases also enables majority voting-based ensemble to attain higher performance.

4.5.2. Performance analysis on Cell2Cell dataset

Cell2Cell dataset is provided with balanced class distribution, but it has also large dimensionality. The simulation results shown in Figs. 15–17 represent the gradual improvement in the prediction performance attained by the used classifiers, when training set is processed with Filter–Wrapper approach.

Random Forest shows improved performance of 0.772 AUC when trained with mRMR selected features. The prediction accuracy is further improved to 0.790 AUC using 18 features selected through GA-based wrapper method as shown in Fig. 15.

Rotation Forest also improves the prediction performance with the application of Filter–Wrapper processing on Cell2Cell training set, contrary to the performance delivered on Orange dataset after application of Filter–Wrapper processing. Rotation Forest ultimately attains 0.789 AUC compared with 0.610 AUC attained on the original Cell2Cell training set as shown in Fig. 15.

RotBoost has shown notable improvement in prediction performance after the application of filter and wrapper phases on Cell2Cell training set. Eighteen of the features selected using wrapper method enable RotBoost to attain as high as 0.820 AUC as shown in Fig. 16. Thus, filter and wrapper phases make significant contribution to provide informative and reduced feature space to RotBoost and thus help in attaining improved prediction performance.

SVM(RBF) and SVM(Poly) also attain improved performance on Cell2Cell dataset contrary to the performance shown on Orange dataset. SVM(RBF) and SVM(Poly) achieve 0.799 AUC and 0.811 AUC, respectively, as shown in Figs. 16 and 17. The improved performance of classifiers obtained on Cell2Cell

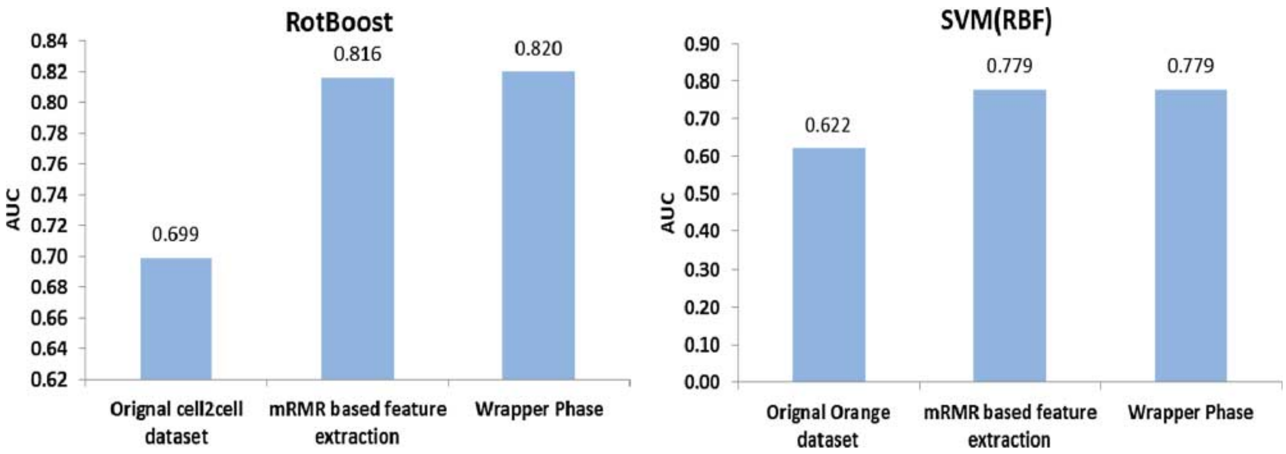


FIGURE 16. Performance improvement attained by RotBoost and majority voting-based ensemble Cell2Cell dataset.

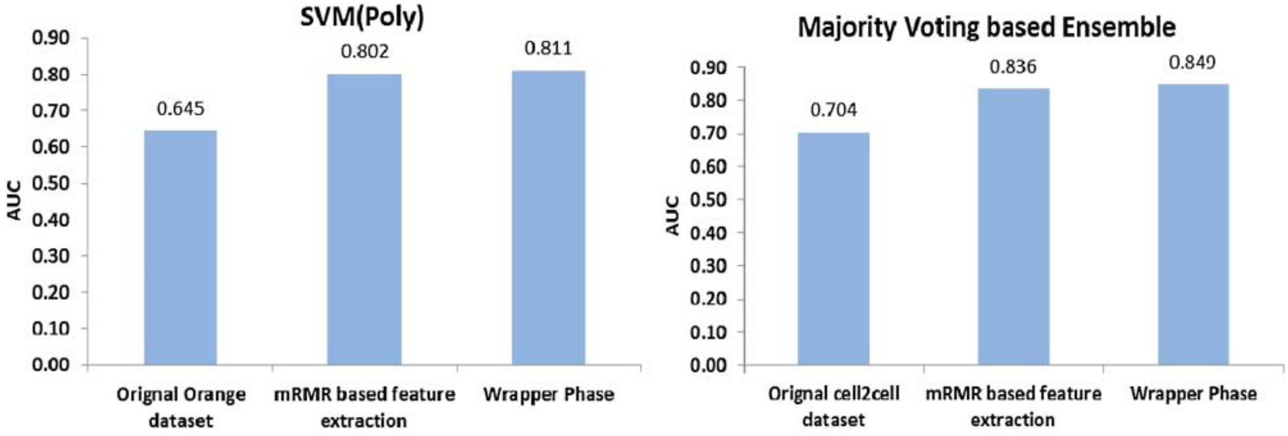


FIGURE 17. Performance improvement attained by SVM(Poly) and majority voting-based ensemble on Cell2Cell dataset.

dataset can be attributed to the underlying distinguishing pattern already existing in the Cell2Cell dataset that facilitates the classifiers to attain good prediction performance.

Majority voting-based ensemble combines the predictions of Random Forest, Rotation Forest, RotBoost and SVMs to attain the highest prediction performance of 0.849 AUC on Cell2Cell dataset as shown in Fig. 17. This result is so far best reported prediction performance on Cell2Cell dataset. Although Cell2Cell dataset is provided with balanced class distribution, filter phase effectively transforms the feature space using mRMR method that enhances the level of learning required by ensemble to attain higher prediction. The combination of Filter–Wrapper and majority voting is thus quite effective in handling enormous nature of the telecom datasets.

4.6. Performance comparison

Figure 18 represents the performance of various predictors in terms of AUC on Orange and Cell2Cell datasets. It can be clearly observed that majority voting-based ensemble achieves the highest prediction performance on both datasets, compared with individual base classifiers. Similarly, RotBoost in individual capacity achieves second best prediction performance on both datasets. Majority voting-based ensemble combines the predictions of Random Forest, Rotation Forest, RotBoost and SVMs and finally produces improved results on both datasets. Effective preprocessing of the training set accomplished through Filter–Wrapper improves the overall prediction performance of the classifiers but majority voting-based ensemble achieves the highest prediction performance on both datasets.

4.6.1. *Comparing majority voting with stacking-based ensemble*

We have also compared the performance of majority voting with stacking-based ensemble generation technique in our work. Stacking refers to combining multiple classifiers, which are trained on a single dataset. The predictions made

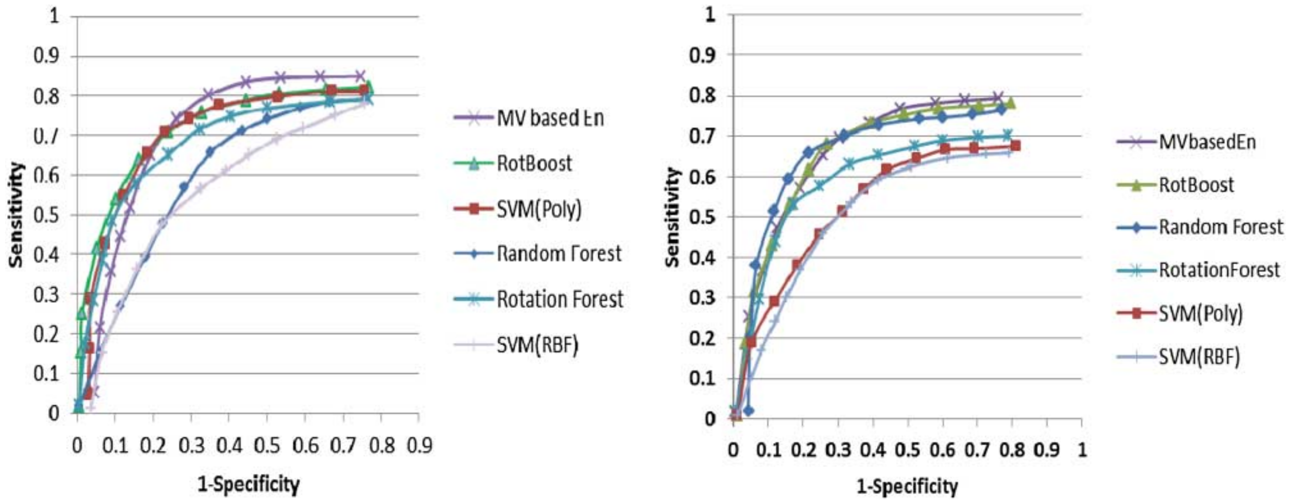


FIGURE 18. Performance comparison in terms of AUC on Orange (left) and Cell2Cell (right) datasets.

TABLE 9. Performance comparison of majority voting and stacking-based ensemble generation techniques.

Method	Filter–Wrapper processed Orange dataset				Wrapper processed Cell2Cell dataset			
	Sensitivity	Specificity	F-measure	AUC	Sensitivity	Specificity	F-measure	AUC
Stacking-based ensemble	0.612	0.596	0.586	0.712	0.800	0.781	0.853	0.862
Majority voting-based ensemble	0.741	0.736	0.727	0.794	0.792	0.776	0.846	0.849

by base classifiers are provided as a training set to the meta classifier. The training set is developed by using predictions obtained by base classifiers following 10-fold cross-validation which is onward used for meta learning. Decision trees are used as meta classifier. Table 9 shows the comparison between stacking and majority voting-based ensemble generation techniques. F-measure is also included here for comparing both the ensemble classification methods. F-measure is the harmonic mean of precision and recall. The higher value of F-measure ensures that both precision and recall are equitably higher and predictor is not biased to a particular class in the test set.

Majority voting-based ensemble scores higher AUC and F-measure on Filter–Wrapper processed Orange dataset, whereas stacking-based ensemble attains higher AUC and F-measure on Cell2Cell dataset as given in Table 9. However, majority voting method is overall more promising for showing higher prediction performance on both of the datasets. Stacking-based ensemble shows inconsistent performance because it attains highest prediction performance on Cell2Cell dataset while showing deteriorated performance on Orange dataset.

Moreover, lift curves are also employed for comparing the performance of both ensemble classification methods. Lift curves represent what ratio of customers needs to be targeted to approach a certain ratio of all churners [2, 19, 41]. Lift curves are different from AUC because lift curves are based on the churn rate, whereas AUC is independent of the churn rate. Figure 19 shows the lift curves for stacking and majority voting-based ensembles on Orange and Cell2Cell datasets, respectively. Lift curve provides the criterion about how many churn customers could be recognized by the prediction system when certain percentage of cumulative customers is considered. Majority voting-based ensemble achieves higher lift scores for (5, 10, 20%) top decile, on both of the used datasets and thus considered to be used as a classification tool in proposed FW-ECP.

In majority voting method, meta learning is not involved rather a simple premise of plural voting for combining the classifiers is adopted. A test instance is assigned a class that wins maximum votes of the base classifiers using majority voting. But, stacking method is based on meta learning, exploiting the prediction outputs obtained by base classifiers.

In our work, stacking-based ensemble develops a meta training set using predicted class labels by base classifiers. But, stacking attains improved results when meta training set is developed using probability distributions instead of only class values [35]. In this way, prediction outputs as well as confidence level of base classifiers are also incorporated in meta training set. Moreover, the dimensionality of the training set involved in meta learning is also important. The increase in dimensionality of meta training set through adding more base classifiers

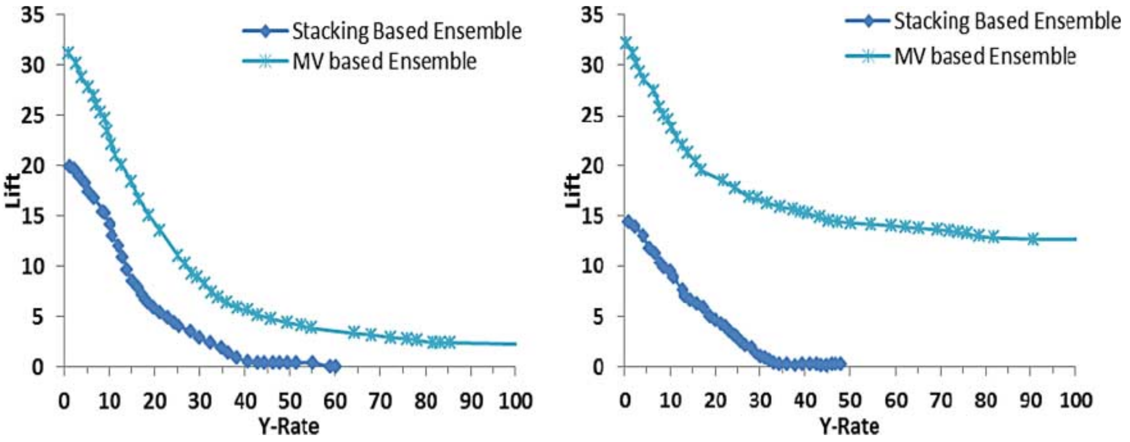


FIGURE 19. Lift curves for stacking and majority voting-based ensemble on Orange (left) and Cell2Cell (right) datasets.

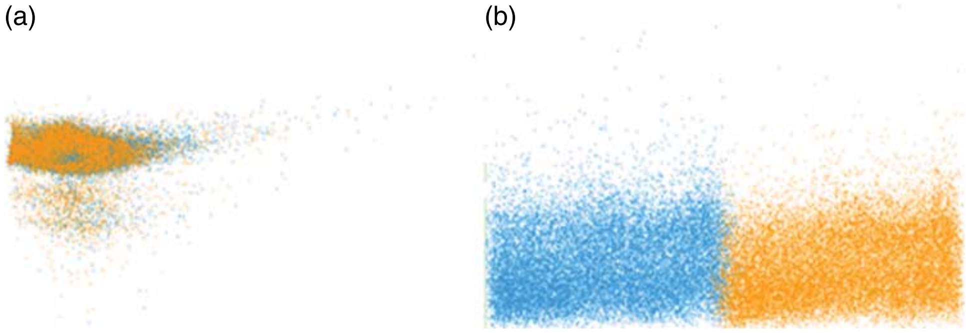


FIGURE 20. Sparse diagrams of Original Orange dataset and Filter–Wrapper processed dataset.

can be a possible way to improve stacking performance. Therefore, stacking-based ensemble can attain consistent prediction performance for telecom churn prediction subject to incorporating probability distributions in meta training set and adding more base classifiers.

4.7. Discrimination analysis of Filter–Wrapper

PCA has been applied on the training set before and after processing through Filter–Wrapper. Sparse diagrams given in Fig. 20 visually represent the separability introduced in the instance space of Orange dataset after processing through Filter–Wrapper. PCA covers the maximum variance present in the dataset and generates artificial principal components. The initial components carry the maximum variance present in the dataset. Principal component one (PC1) is plotted against principal component two (PC2) to see the separability introduced in the data space after processing the training set through Filter–Wrapper method. It can be observed in Fig. 20 that underlying pattern in data space is easier for classifiers to learn once the training set is processed using Filter–Wrapper. Thus, the application of Filter–Wrapper effectively handles the training set and reduces the effect of issues related to imbalanced class distribution and curse of dimensionality of telecom dataset.

4.8. Comparison of FW-ECP with other approaches

The FW-ECP system combines the filter and wrapper-based methods to effectively handle the issues associated with enormous nature of telecom datasets. This study attempts to specifically address the core issues faced in telecom churn prediction problem in more systematic and intuitive manner. Considering the fact that classification approaches show better performance when provided with balanced class distribution [21] and relevant features in training set [25, 42–44], FW-ECP also employs Filter–Wrapper to preprocess the telecom training set. Filter–Wrapper helps in eliminating redundant features and thus leads to improved prediction performance.

The wrapper phase is introduced in FW-ECP in a bid to use GA's searching capabilities for further reducing the feature space by further discarding less useful features. Wrapper

TABLE 10. Performance comparison of proposed FW-ECP with other existing approaches.

Method	AUC
Performance comparison on Orange dataset	
FW-ECP	0.794
CP-MRB [45]	0.760
Chr-PmRF [33]	0.751
Gradient boosting machine [36]	0.737
Stochastic gradient boosting [46]	0.728
Decision stump-based model [47]	0.725
Decision tree-based model [47]	0.715
Bayesian Net (BN)-based approach [2]	0.714

TABLE 11. Performance comparison of the FW-ECP approach with other existing approach.

Method	AUC
Performance comparison on Cell2Cell dataset	
FW-ECP	0.849
CP-MRB [45]	0.816
Naïve Bayes (NB)-based approach [2]	0.818

method further improves the performance as shown in Figs. 12–17, except in the cases of SVM(Poly) on Orange dataset and SVM(RBF) on Cell2Cell dataset. On the basis of a simulation result, our proposed approach has shown improved performance compared with other existing approaches by considerable margin as given in Tables 10 and 11. The performance comparison is conducted using AUC measure because other existing churn prediction approaches, which use same datasets, also follow AUC measure for performance evaluation. FW-ECP outperforms other approaches, which have used Orange dataset. The performance of FW-ECP is compared with that of Chr-PmRF churn prediction system. Chr-PmRF employs a preprocessing phase for handling imbalanced class distribution and high dimensionality of the Orange telecom dataset and then Random Forest attains a prediction performance of 0.751 AUC. Similarly, another prediction system (CP-MRB) based on mRMR feature extraction method and RotBoost attains 0.760 AUC on Orange dataset. The results obtained in this work are also compared with the approach based on gradient boosting machine [36] that achieves 0.737 AUC on Orange dataset. 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 the mean response for each quantile is calculated using half of the training data. The calculated mean is applied to the other half and then AUC is calculated which ranks the variables. This method lacks a systematic approach, and is based on imputation method for feature selection.

The results obtained using FW-ECP are also compared with an AdaBoost-based approach [47]. AdaBoost-based model is optimized with multi-armed bandits (MABs). In this approach, AdaBoost builds a classifier in a stepwise fashion by adding simple base classifiers to pool and using simple 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 the whole space. The results in Table 10 show that the 0.725 AUC and 0.715 AUC are attained [47], using tree and stump-based learners with AdaBoost, respectively.

A comparison with Stochastic Gradient Boosting algorithm is also made [46], which applies boosting with Decision Trees as the classification method for churn prediction. This method scores 0.7282 AUC for Orange dataset. In another study, a Bayesian network with oversampling attains 0.714 AUC to predict churners from Orange dataset [2]. But, our FW-ECP approach achieves 0.794 AUC on Orange dataset that is the best score on Orange dataset, to best of our knowledge, reported so far. Thus hybridizing filter and wrapper methods in collaboration with majority voting-based ensemble prove its effectiveness for being more competent in addressing the challenges faced in telecom churn prediction.

Table 11 shows the comparison of our FW-ECP with other existing approach using Cell2Cell dataset. FW-ECP attains improved prediction performance on Cell2Cell dataset as well, compared with Naïve Bayes's performance. In Verkerke's work [2], Naive-based approach shows good predictions performance only on Cell2Cell dataset and lacks in attaining good prediction performance on other datasets used in the study, resulting an inconsistent performance. Moreover, in our FW-ECP system, 10-fold cross-validation is deployed to evaluate the prediction performance, whereas a single random split of the training set is performed to evaluate performance in Verkerke's work. The performance of our FW-ECP is also compared with that of CP-MRB, which also reports prediction results from Cell2Cell dataset. Our CP-MRB attains an improved prediction performance of 0.849 AUC compared with CP-MRB's prediction performance of 0.816 AUC.

In addition, the proposed FW-ECP may also be beneficial to offer a benefit for business context. FW-ECP approach focuses to remove the features and instances which are less effective in extending useful learning to classifier. Therefore, final features extracted through Filter–Wrapper process can intuitively be investigated to identify the underlying reasons of customers' churning.

5. CONCLUSION

In this manuscript, FW-ECP system is presented for telecom churn prediction based on hybridizing Filter–Wrapper and ensemble classification in a new way. To cope with enormous nature of telecom datasets, the proposed churn prediction system adopts a systematic approach to settle the imbalanced distribution and high dimensionality of the training set through employing Filter–Wrapper method. Filter phase applies PSO-based undersampling to handle the imbalance distribution in the training set and then mRMR method is used to remove redundant and less informative features. As a result, in filter phase, 36 and 31 features are selected from Orange and Cell2Cell datasets, respectively. Then, wrapper phase exploits GA's global searching capabilities using a classifier's performance as the feedback to remove any further redundancy in feature space. It selects a set of 24 best features from Orange and 18 from Cell2Cell datasets. The Filter–Wrapper processed training set helps the classifiers to effectively learn and predict the churners. Finally, majority voting-based ensemble of Random Forest, Rotation Forest, RotBoost and SVMs is observed as an effective method in predicting churners. Our proposed FW-ECP system uniquely adopts a broad framework based on hybridizing Filter–Wrapper and ensemble classification, which effectively addresses the core issues faced in telecom churn prediction. Thus, FW-ECP is expected to be beneficial for achieving enhanced prediction of churners in telecom industry.

FUNDING

This work is supported by the Higher Education Commission of Pakistan (HEC) as per award no. 17-5-6(Ps6-002)/HEC/Sch/2010.

REFERENCES

[1] ITU-ICT-2014 (2014) *The World in 2014 ICT Facts and Figures*. Union, I. T., Geneva, Switzerland.

[2] Verbeke, W., Dejaeger, K., Martens, D., Hur, J. and Baesens, B. (2012) New insights into churn prediction in the telecommunication sector: a profit driven data mining approach. *Eur. J. Oper. Res.*, **218**, 211–229.

[3] Athanassopoulos, A.D. (2000) Customer satisfaction cues to support market segmentation and explain switching behavior. *J. Bus. Res.*, **47**, 191–207.

[4] Masand, B., Datta, P., Mani, D.R. and Li, B. (1999) CHAMP: a prototype for automated cellular churn prediction. *Data Min. Knowl. Discov.*, **3**, 219–225.

[5] Rosset, S., Neumann, E., Eick, U. and Vatnik, N. (2003) Customer lifetime value models for decision support. *Data Min. Knowl. Discov.*, **7**, 321–339.

[6] Huang, B., Kechadi, M.T. and Buckley, B. (2012) Customer churn prediction in telecommunications. *Expert Syst. Appl.*, **39**, 1414–1425.

[7] Verikas, A., Gelzinis, A. and Bacauskiene, M. (2011) Mining data with random forests: a survey and results of new tests. *Pattern Recogn.*, **44**, 330–349.

[8] Xie, Y., Li, X., Ngai, E.W.T. and Ying, W. (2009) Customer churn prediction using improved balanced random forests. *Expert Syst. Appl.*, **36**, 5445–5449.

[9] Rodriguez, J.J., Kuncheva, L.I. and Alonso, C.J. (2006) Rotation Forest: a new classifier ensemble method. *IEEE Trans. Pattern Anal.*, **28**, 1619–1630.

[10] Zhang, C.-X. and Zhang, J.-S. (2010) A variant of Rotation Forest for constructing ensemble classifiers. *Pattern Anal. Appl.*, **13**, 59–77.

[11] Kuncheva, L.I. and Rodriguez, J.J. (2007) *An Experimental Study on Rotation Forest Ensembles*. In Michal, H., Josef, K. and Fabio, R. (eds), Multiple Classifier Systems. Berlin, Heidelberg, Springer.

[12] Zhang, C.-X. and Zhang, J.-S. (2008) RotBoost: a technique for combining Rotation Forest and AdaBoost. *Pattern Recogn. Lett.*, **29**, 1524–1536.

[13] Bock, K.W.D. and Poel, D.V.d. (2011) An empirical evaluation of rotation-based ensemble classifiers for customer churn prediction. *Expert Syst. Appl.*, **38**, 12293–12301.

[14] Mozer, M.C., Wolniewicz, R., Grimes, D.B., Johnson, E. and Kaushansky, H. (2000) Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry. *IEEE Trans. Neural Netw.*, **11**, 690–696.

[15] Kim, Y. (2006) Toward a successful CRM: variable selection, sampling, and ensemble. *Decis. Support Syst.*, **41**, 542–553.

[16] Verbeke, W., Martens, D. and Baesens, B. (2013) Social network analysis for customer churn prediction. *Appl. Soft. Comput.*, **14**, 431–446.

[17] Nanavati, A.A., Singh, R., Chakraborty, D., Dasgupta, K., Mukherjea, S., Das, G., Gurumurthy, S. and Joshi, A. (2008) Analyzing the structure and evolution of massive telecom graphs. *IEEE Trans. Knowl. Data Eng.*, **20**, 703–718.

[18] Kyoungok, K., Chi-Hyuk, J. and Jaewook, L. (2014) Improved churn prediction in telecommunication industry by analyzing a large network. *Expert Syst. Appl.*, **41**, 6575–6584.

[19] Droftina, U., Stular, M. and Kosir, A. (2014) A diffusion model for churn prediction based on sociometric theory. Available at 10.1007/s11634-014-0188-0.

[20] Gerpott, T.J. and Ahmadi, N. (2015) Regaining drifting mobile communication customers: Predicting the odds of success of winback efforts with competing risks regression. *Expert Syst. Appl.*, **42**, 7917–7928.

[21] Burez, J. and Poel, D.V.d. (2009) Handling class imbalance in customer churn prediction. *Expert Syst. Appl.*, **36**, 4626–4636.

[22] Huang, Y., Huang, B.Q. and Kechadi, M.T. (2010) A New Filter Feature Selection Approach for Customer Churn Prediction in Telecommunications. 2010 IEEE Int. Conf. on Industrial Engineering and Engineering Management, Macao. IEEE.

[23] Shin, Y.H., David, C.Y. and Hsiu, Y.W. (2006) Applying data mining to telecom churn management. *Expert Syst. Appl.*, **37**, 3665–3675.

[24] Keramati, A., Jafari-Marandi, R., Aliannejadi, M., Ahmadian, I., Mozaffari, M. and Abbasi, U. (2014) Improved churn prediction in telecommunication industry using data mining techniques. *Appl. Soft. Comput.*, **24**, 994–1012.

[25] Huang, B.Q., Kechadi, T.M., Buckley, B., Kiernan, G., Keogh, E. and Rashid, T. (2010) A new feature set with new window techniques for customer churn prediction in land-line telecommunications. *Expert Syst. Appl.*, **37**, 3657–3665.

[26] Breiman, L. (2001) Random Forests. *Mach Learn.*, **45**, 5–32.

[27] Dietterich, T.G. (2000) An experimental comparison of three methods for constructing ensemble of decision trees: bagging, boosting and randomization. *Mach Learn.*, **40**, 139–157.

[28] Wang, C.-W. and You, W.-H. (2013) Boosting-SVM: effective learning with reduced data dimension. *Appl Intell.*, **39**, 465–474.

[29] Bob, D., Eibe, F., Lyn, H., Geoff, H., Mike, M., Bernhard, P., Tony, S. and Ian, W. (2013) The University of Waikato WEKA. http://www.cs.waikato.ac.nz/ml/weka/ (accessed January 1, 2011).

[30] Yijun, S., Sinisa, T. and Steve, G. (2010) Local-learning-based feature selection for high-dimensional data analysis. *IEEE Trans. Pattern Anal.*, **32**, 1610–1625.

[31] Mikel, G., Alberto, F., Edurne, B., Humberto, B. and Francisco, H. (2012) A review on ensembles for the class imbalance problem: bagging–boosting- and hybrid-based approaches. *IEEE Trans. Syst. Man Cybernatics-Part C: Appl. Rev.*, **42**, 463–484.

[32] Kotsiantis, S.B. (2011) Cascade generalization with reweighting data for handling imbalanced problems. *Comput J.*, **54**, 12.

[33] Adnan, I., Muhammad, R. and Asifullah, K. (2012) Churn prediction in telecom using Random Forest and PSO based data balancing in combination with various feature selection strategies. *Comput. Electr. Eng.*, **38**, 1808–1819.

[34] Peng, H., Long, F. and Ding, C. (2005) Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal.*, **27**, 1226–1238.

[35] Dzeroski, S. and Zenko, B. (2004) Is combining classifiers with stacking better than selecting the best one. *Mach. Learn.*, **54**, 255–273.

[36] Miller, H., Clarke, S., Lane, S., Lonie, A., Lazaridiz, D., Petrovski, S. and Jones, O. (2009) *Predicting customer behaviour: The University of Melbourne's KDD Cup Report*. JMLR Workshop and Conference Proceedings, Paris, France, June 28, pp. 45–55. MIT Press, Cambridge.

[37] Owczarczuk, M. (2010) Churn models for prepaid customers in the cellular telecommunication industry using large data marts. *Expert Syst. Appl.*, **37**, 4710–4712.

[38] SIGKDD (2009) KDDCup 2009 Challenge. http://www.sigkdd.org/kdd-cup-2009-customer-relationship-prediction (accessed March 18, 2012).

[39] Sorokina, D. (2009) *Application of Additive Groves Ensemble with multiple counts Feature Evaluation to KDD Cup'09 Small Data Set*. JMLR Workshop and Conference Proceedings, Paris, France, June 28, pp. 101–109. MIT Press, Cambridge.

[40] Pacharawongsakda, E. and Theeramunkong, T. (2013) Multi-label classification using dependent and independent dual space reduction. *Comput. J.*, **56**, 1113–1135.

[41] Ning, L., Hua, L., Jie, L. and Guangquan, Z. (2012) A customer churn prediction model in telecom industry using boosting. *IEEE Trans. Ind. Informat.*, **10**, 1659–1665.

[42] Huang, Y., Huang, B.Q. and Kechadi, M.T. (2010) *A New Filter Feature Selection Approach for Customer Churn Prediction in Telecommunications*. 2010 IEEE Int. Conf. on Industrial Engineering and Engineering Management, Macau, China, December 7–10, pp. 338–342. IEEE, NY.

[43] Huang, B.Q., Kechadi, T.M., Buckley, B., Kiernan, G., Keogh, E. and Rashid, T. (2010) A new feature set with new window techniques for customer churn prediction in land-line telecommunications. *Expert Syst. Appl.*, **37**, 3657–3665.

[44] Vinh, L., Lee, S., Park, Y.-T. and Auriol, D.B. (2012) A novel feature selection method based on normalised mutual information. *Appl Intell.*, **37**, 100–120.

[45] Adnan, I., Asifullah, K., Yeon Soo, L. (2013) Intelligent churn prediction in telecom: employing mRMR feature selection and RotBoost based ensemble classification. *Appl Intell.*, **39**, 659–672.

[46] Komoto, K., Sugawara, T., Tetu, T. I. and Xuejuan, X. (2009) Stochastic Gradient Boosting. http://www.kdd.org/kdd-cup/view/kdd-cup-2009/Results (accessed January 16, 2016).

[47] Busa-Fekete, R. and Kegl, B. (2009) *Accelerating AdaBoost using UCB*. JMLR Workshop and Conference Proceedings, Paris, France, June 28, pp. 111–122. MIT Press, Cambridge.