Churn prediction in Telecom using Random Forest and PSO based data balancing in combination with various feature selection strategies

Adnan Idris a,b, Muhammad Rizwan a, Asifullah Khan a,*

aDepartment of Computer and Information Sciences, Pakistan Institute of Engineering and Applied Sciences, Nilore Islamabad, 45650 Pakistan

bDepartment of Computer Sciences and Information Technology, University of Poonch Rawalakot 12350 Kashmir, AJK, Pakistan

Abstract

The telecommunication industry faces fierce competition to retain customers, and therefore requires an efficient churn prediction model to monitor the customer’s churn. Enormous size, high dimensionality and imbalanced nature of telecommunication datasets are main hurdles in attaining the desired performance for churn prediction. In this study, we investigate the significance of a Particle Swarm Optimization (PSO) based undersampling method to handle the imbalance data distribution in collaboration with different feature reduction techniques such as Principle Component Analysis (PCA), Fisher’s ratio, F-score and Minimum Redundancy and Maximum Relevance (mRMR). Whereas Random Forest (RF) and K Nearest Neighbour (KNN) classifiers are employed to evaluate the performance on optimally sampled and reduced features dataset. Performance is evaluated using sensitivity, specificity and Area under the curve (AUC) based measures. Finally, it is observed through simulations that our proposed approach based on PSO, mRMR, and RF termed as Chr-PmRF, performs quite well for predicting churners and therefore can be beneficial for highly competitive telecommunication industry.

Keywords: Churn Prediction, Particle Swarm Optimization, PCA, Fisher’s Ratio, F-Score, mRMR, Random Forest, KNN and AUC.
1. Introduction

Telecommunication is one of the industries, where customer base plays a significant role in maintaining stable revenues and thus a serious attention is devoted to retain customers. The customers’ appetency to switch over to any other viable network varies for different reasons such as, call quality, more complimentary competitors’ pricing plan, customers’ billing problems, etc. The telecommunication industry always faces threat of financial loss from potential churners therefore, an efficient churn prediction model not only secures the revenues but also provides hints to management for targeting potential churners by reducing the market-relevant shortcomings. Hence, customer relationship management in a telecommunication company desires an efficient churn prediction model for predicting the potential churners.

The efficiency of churn prediction model, based on classification system relies on learning acquired through the available dataset. The appropriately preprocessed dataset helps the classifier to attain the required training level, which ultimately turns into a desirable performance. Telecommunication companies archive data by acquiring a lot of information about customers. Unfortunately, such a data has high dimensionality and imbalanced class distribution. Generally, information regarding demographics, contract nature, billing and payments, call details, services log etc. are maintained that eventually leads to the high dimensionality. Similarly, the number of churners in telecommunication industry is usually far less compared to non-churners and consequently, it results in an imbalanced dataset. This imbalance distribution in the dataset might cause weak learning by a classifier. Therefore, the preprocessing phase essentially requires a proper sampling and feature reduction strategy for accomplishing good learning by the classifier.

Principle Component Analysis (PCA) and Independent Component Analysis (ICA) [1] are mostly used feature selection strategies, which linearly operate to select the useful and discriminating features present in a dataset. PCA is based on data covariance while ICA uses higher order statistics for achieving data independence, along with reducing the dimensionality of the data. Similarly, some well known sampling techniques are Random Oversampling (ROS) and Random Undersampling (RUS)
where instances of the minority class are duplicated and majority class are discarded, respectively. Due to the random selection, involved in duplicating and discarding the data values, these approaches lack consistency and show varying performances. In addition, the RUS can discard some useful instances and ROS can lead to overfitting owing to replication. Similarly, One Sided Selection (OSS) removes the noisy and boundary line majority class instances, but it is slow when used on large datasets for using Tomek Links [3], which are proven costly. Cluster based oversampling identifies rare cases from the dataset and resamples the instances, but considered to be effective [4,5] for small sized training dataset. Synthetic Minority Oversampling Technique is an intelligent oversampling method, where new minority class samples are added synthetically, but it involves high computational cost [6] and thus is not suitable for large sized dataset. Data Boost -IM [7] is another approach used for sampling, where the predictive occurrences of both minority and majority classes are increased using synthetic data generation, this approach also involves high computation cost and therefore is not appropriate for large sized dataset. Most of the sampling techniques either use random selection for undersampling, which consequently introduces bias, or synthetic generation of minority class samples, which are proven costly. Therefore, an optimized sampling technique can be employed for sampling dataset, which can effectively mitigate the imbalance in data distribution.

Besides the appropriate feature selection and sampling techniques required to handle the imbalanced telecommunication dataset, the classification models are the real tools, which perform the customer churn prediction. Researchers have used Decision Trees [8,9,10], Logistic Regression [10,11], Genetic Programming [12,26], Neural Network [13,14,15,16], Random Forest [17], Adaboost [19] Naive based algorithms[11] for various classification problems including churn prediction. Some of the techniques have also used nonlinear kernel methods in Support Vector Machines for churn prediction but they suffer from the high dimensionality of a dataset [8]. Other classification models such as SVM [20,27], KNN [11] etc. also show deteriorated performances in case of telecommunication churn prediction, because of the imbalanced nature of dataset [11]. Although some approaches, based on ensemble of KNN and logistic regression [18], additive grooves with multiple counts features evaluation [19] and hybrid two phased feature selection [20], have been suggested but the classification models could not achieve the needed performance. These ensemble approaches,
primarily curtail the data dimensionality by selecting features and introduce data balancing in the due course, but the classification performance suffers due to the loss of information resulting from application of improper sampling and feature reduction methods.

Realizing the challenges, being faced in customer churn prediction due to large size, high dimensionality and imbalanced nature of the telecommunication dataset, we initially analyzed RUS and PSO based [23] undersampling methods separately. The PSO based undersampling method initially subsamples the dataset and then evaluates each subsample against KNN and Random forest on the basis of AUC. Once an optimal subsample is selected then PCA, F-Score, Fisher’s Ratio and mRMR are applied separately and analyzed with RF and KNN classifiers. It is finally observed that our proposed approach based on PSO, mRMR and RF termed as Chr-PmRF provides best results among the other combinations of sampling, feature reduction and classification techniques.

The rest of the manuscript first presents the proposed churn prediction approach in section 2. Next, section 3 analyzes the simulated results and gives corresponding discussions. Finally, the conclusions are drawn in section 4.

2. Material and Methods

The telecommunication datasets generally face the problems of skewed data distribution and high dimensionality. This causes the classification algorithms to perform poorly for customers churn prediction. Therefore, in Chr-PmRF approach, we concentrate in handling these problems. The basic block diagram shown in Figure 1 highlights various steps involved in Chr-PmRF.

**Figure 1.** The Basic Block diagram representing the steps involved in proposed, Chr-PmRF approach.

We initially preprocess the dataset in order to handle the problems of missing values and nominal values present in the dataset. RUS and PSO based undersampling methods are employed to evaluate their effectiveness in improving the prediction performance. Various feature selection strategies such as PCA, F-Score, Fisher’s Ratio and mRMR are employed separately and their respective impact on classification is evaluated. In this work, RF and KNN are main classification schemes employed to evaluate the combinations of sampling and feature selection methods using AUC, sensitivity and
specificity. Chr-PmRF based on PSO, mRMR and RF shows best results among the other combinations of undersampling, feature selection and classification methods. The methods involved at various stages of experimentation are explained in later sections.

2.1. Dataset

French Telecom Company named Orange has provided processed version of the dataset for studying the problem of customers churn prediction [21]. The dataset used in this study has 50,000 instances with 260 features. The dataset comprises of 190 numerical and 70 nominal features. The dataset hides names of features to keep the customer’s information private. It has 3,672 instances of minority and 46,328 instances of majority class that amounts 7.3% scarce share of minority class in the whole dataset. Eighteen of the features have no value at all and five of the features have merely one value, thus these useless features have been discarded.

2.2. Dataset Preprocessing

The focus of our approach is to appropriately preprocess the dataset so that an optimal subsample of the dataset with useful features is extracted where classification algorithms are expected to produce the desired performance. The preprocessing phase starts with handling useless features and shrinks the dataset to 208 features. Then the next step handles 70 nominal features present in the dataset. The nominal values are transformed to numerical format by grouping the modalities [19] in three categories, which are small, medium and large depending upon the number of occurrences of instances in each category. Once the data is in numerical format, we further process the dataset for undersampling using RUS and PSO separately. Thereafter, PCA, Fisher’s ratio, F-Score, and mRMR are employed for selecting the features before employing classification models.

2.3. Random Undersampling

Sampling methodologies handle the imbalance of the data distribution. The majority class in the imbalanced dataset tends to overwhelm the classifier and thus minority class is suppressed. The RUS method is employed to undersample the large telecommunication dataset. RUS works by randomly selecting the samples from majority class, with equal in number of minority class, to establish a
balance data distribution [22]. Such an act benefits the classifier to attain improved training. This way, the dominating presence of majority class instances in the dataset is handled.

### 2.4. PSO based Undersampling

PSO based [23] undersampling approach evolves a subsample of the dataset, optimized for RF and KNN in terms of AUC. Figure 2 represents the PSO based undersampling method.

**Figure 2.** PSO based undersampling model.

This approach finds and ranks the most informative instances of majority class and then combines these instances with minority class to develop an optimal balanced dataset. Different subsets of instances are chosen from minority and majority classes and the goodness of each subset is evaluated by KNN and RF using AUC. In this way PSO evolves an optimal dataset that helps in building a balanced churn prediction model.

PSO operates by steering the subsets of instances in search space which are referred as particles. These particles are moved, guided by their own best positions and swarm’s best positions. The improved positions of the particles also guide the whole swarm to move. The process iterates till a particle is chosen with good AUC. The movements of these particles are governed, using equations (1), (2) and (3).

\[
v_{i,j}(t + 1) = wv_{i,j}(t) + c_1R_1(pbest_{i,j} - x_{i,j}(t)) + c_2R_2(gbest_{i,j} - x_{i,j}(t))
\]

(1)

The velocity of \(i^{th}\) particle \(v_{i,j}(t)\) is computed using equation (1), where \(pbest_{i,j}\) represents the particle’s own best position and \(gbest_{i,j}\) represents the particle with best position in the complete swarm. \(c_1\) 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 used to update the particle’s velocity. A function given in equation (2) operates on calculated velocities of each particle and then the position of swarm particle is updated as given in equation (3).

\[
S(v_{i,j}(t + 1)) = \frac{1}{1 + e^{-v_{i,j}(t+1)}}
\]

(2)

\[
x_{i,j}(t + 1) = \begin{cases} 0 : \text{if random()} \geq s(v_{i,j}(t + 1)) \\ 1 : \text{if random()} < s(v_{i,j}(t + 1)) \end{cases}
\]

(3)
Finally, the majority class instances are ranked based on their selection frequency in model building. The more time an instance is selected for model building, the higher the rank is given to that instance. When the frequency list is obtained then instances with higher ranks from majority class are combined with the instances of minority class to create a balanced telecommunication dataset. The balanced dataset is used for RF and KNN training and produces improved churn prediction results on 10 folds cross validation. The improvement in churn prediction performance, attained using PSO undersampled dataset, suggests it to be used for further investigations.

2.5. PCA based feature selection

The PSO undersampled dataset is further treated with PCA in order to reduce the high dimensionality. PCA transforms the dataset into artificial components, which cover maximum variance present in dataset. In the first step, respective mean is subtracted from each of the data dimension as expressed in equation 4.

\[
\text{Cov}_{ij} = \frac{\sum (X_1 - M_1)(X_2 - M_2)}{n}
\]  

(4)

Where \(X_1\) and \(X_2\) represent instances of features under consideration, \(M_1\) and \(M_2\) are the respective means and \(n\) is the total number of instances. Covariance measurements are calculated in the form of a matrix as given in equation (5).

\[
c = \begin{bmatrix}
\nu(X_1) & c(X_1, X_2) & \cdots & c(X_1, X_p) \\
c(X_1, X_2) & \nu(X_2) & \cdots & c(X_2, X_p) \\
c(X_1, X_p) & c(X_2, X_p) & \cdots & \nu(X_p)
\end{bmatrix}
\]  

(5)

Eigenvectors are computed, which show the direction of maximum variance along a certain dimension. Eigenvectors are ordered using eigenvalues in descending order, which identify the insignificant components present in the dataset. The reduced feature vector is represented with selective significant components as given in equation (6);

\[
Feature\ Vector = (eig1, eig2, eig3, \ldots, eign)
\]  

(6)
Finally, this feature vector is transformed into new reduced feature vector by taking the transpose and multiplying it on the left of the original transposed dataset as given in equation (7).

\[
\text{New Dataset} = [\text{FeatureVector}]^T [\text{Data}]^T \tag{7}
\]

PCA reduces the feature space to a set of most coherent features covering 98% variance of the PSO undersampled dataset.

### 2.6. Fisher’s Ratio based Feature selection

Fisher’s Ratio is considered sensitive to the non-normality of the data and measures the discriminating power of features in a dataset. Fisher’s Ratio is computed as given in equation (8).

\[
\text{Fisher’s Ratio} = \frac{\mu_1 - \mu_2}{\sigma_1^2 - \sigma_2^2} \tag{8}
\]

Where, \(\mu_1\) and \(\mu_2\) being the means of binary classes involved and \(\sigma_1^2\) and \(\sigma_2^2\) the respective variances. PSO balanced dataset is also treated with Fisher’s Ratio based feature selection method and its impact on the performance of KNN and RF is evaluated.

### 2.7. F- Score based Feature Selection

PSO undersampled dataset is also treated with F-Score based feature selection. F-score is a simple technique, which measures the discrimination power of two sets of real numbers. Given training vectors \(x_k, \quad k = 1, \ldots, m\) if the number of instances of churner and non-churner classes are \(n^+\) and \(n^-\), respectively, then the F-score of \(i^{th}\) feature is defined as given in equation (9):

\[
F_i = \frac{1}{n^+-I} \sum_{k=1}^{n^+} (\bar{x}_i^{(+)} - \bar{x}_i)^2 + \frac{1}{n^-I} \sum_{k=1}^{n^-} (\bar{x}_i^{(-)} - \bar{x}_i)^2 + \frac{1}{n^-I} \sum_{k=1}^{n^-} (\bar{x}_i^{(+)} - \bar{x}_i)^2 \tag{9}
\]

Where, \(\bar{x}_i\) is the mean value of \(i^{th}\) feature, \(\bar{x}_i^{(-)}\) for negative instances, and \(\bar{x}_i^{(+)}\) is the mean value of \(i^{th}\) feature for positive instances. The F-score minimizes the intraclass distance whereas maximizes the interclass distance of the instances as shown in equation (9). The larger the F-score is, the more likely the feature to be more discriminative.
2.8. mRMR based Feature Reduction

The mRMR works by selecting those features which show strong correlation with class labels while not being dependent on each other [24]. mRMR extracts the subset of features from the PSO undersampled dataset, which have strong correlations with class labels thus showing maximum relevance. While maintaining high correlation with class labels, the instances can be selected to be mutually far away from each other, resulting into minimum redundancy. The maximum relevance criterion is implemented with the help of expressions given in equation (10) and (11):

$$\max D(S, c), D = I \left( \{ x_i, i = 1, ..., m \}; c \right)$$  \hspace{1cm} (10)

$$\max D(S, c), D = \frac{1}{|S|} \sum_{x_i \in S} I( x_i ; c )$$  \hspace{1cm} (11)

Where, $D$ is the dependency that is intended to be maximized in order to establish maximum relevance of the instances $S$ with class labels $c$. The $I(x_i ; c)$ measures mutual information between the instance $x_i$ and the corresponding class label $c$. The maximum relevance is sort out by searching the feature sets, which satisfy the criteria in equation (11) and approximates the $D(S, c)$ in equation (10) with the mean value of all mutual information values between individual feature $x_i$ and class $c$. A feature set $S$ is chosen where features have higher dependency on the respective class labels. The feature set $S$ having maximum relevance with class labels can be redundant as well. Therefore, one of the two redundant features is removed that would not change the discriminating power of the feature set. The expression given in equation (12) minimizes the redundancy, $R$:

$$\min R(S), R = \frac{1}{|S|^2} \sum_{x_i, x_j \in S} I( x_i, x_j )$$  \hspace{1cm} (12)

Both the above criteria of minimizing redundancy and maximizing relevance are combined and $\Phi$ operator is defined as given in equation (13). This simplest form is used to optimize both $D$ and $R$.

$$\max \Phi(D, R), \Phi = D - R$$  \hspace{1cm} (13)

mRMR selects a feature subset from the PSO balanced dataset, which has strong relevance with class targets and at the same time having maximum unique values, which consequently improves the classification performance.
2.9. Random Forest and KNN based Classification

RF and KNN classifiers have been used twice in our Chr-PmRF approach. Initially, PSO optimizes subset-samples of the dataset for RF and KNN using AUC. During classification, these classifiers are used as predictors to evaluate various preprocessing variants involved in this study. RF is an ensemble of decision trees, which are grown from the bootstrap samples [8]. The final prediction is made on the basis of majority voting from the ensemble trees. These bootstrap samples are created from training data. RF is chosen as classifier for its capability to minimize over all classification error. Therefore, if a balanced dataset is provided to the random forest it shows good performance.

KNN is known as simplest of the classifiers and predicts the label of a test instance by comparing its proximity with the instances from training vector. KNN works following simple computations, by finding the distance between \( X \) and \( X_i \), where \( X \) is the test instance and \( X_i \), is the instance from the training vector;

\[
S(X, X_i) = 1 - \frac{X \cdot X_i}{\|X\| \|X_i\|} \quad (i = 1, 2, 3, ..., N) \tag{14}
\]

The \( X \cdot X_i \) represents the dot product of two vectors containing the feature values while \( \|X\| \) and \( \|X_i\| \) represent their respective moduli. The minimum distance between \( X \) and \( X_k \) \((k=1,2,3...N)\) is computed as;

\[
S(X, X_k) = \min\{S(X, X_1), S(X, X_2), ..., S(X, X_N)\} \tag{15}
\]

The best match is identified where the minimum distance value of \( X \) is found in comparison with training examples \( X_k \). Then target label is assigned to \( X \), showing if \( X \) is from positive or negative class. KNN is a proximity based classifier and its performance is important for evaluating the feature selection strategies. Because the feature selection techniques used in this study impact the proximity of the dataset.

2.10. Performance Measures

We have used Area under the curve (AUC), sensitivity and specificity as performance measures to analyze the performance of classifiers on processed dataset. AUC is a useful technique for
visualizing and evaluating the performances [28]. The AUC is a single number measure of a graphical plot of the sensitivity, or true positive rate, versus false positive rate for a binary classifier system as its discrimination threshold is varied. AUC is used as a fitness measure, for achieving an optimization for PSO that guides the updation of the sample's best and swarm’s best positions to evolve an optimal subsample of the dataset. Later AUC is also used to evaluate the prediction performance of KNN and RF on processed dataset.

3. Proposed Chr-PmRF approach

Besides various combinations of sampling, feature selection and classification methodologies employed, we have observed that PSO based undersampling in combination with mRMR based feature selection and RF classifier yields best churn prediction results. Therefore, in what follows, we will focus on this particular combination denoted as Chr-PmRF. Our proposed Chr-PmRF efficiently utilizes a PSO based undersampling method, which not only undersamples the dataset but also optimizes chosen instances for KNN and RF. This surely improves the KNN and RF performances when tested on 10 fold cross validation. The PSO based sampling approach works by partitioning the original dataset into an external test set and two external training sets using 3-fold stratified cross validation as shown in Figure 3.

**Figure 3.** Block diagram of Original Dataset partitioning, using 3-fold stratified cross validation for PSO

Each of the external training set is further divided using 3-fold stratified cross validation, which consequently produces internal training sets and internal test sets. The sampling is performed using internal training sets while internal test sets are used for guiding the optimization process. The external test is kept separated and only used for the evaluation of the balanced dataset. PSO generates different subsets of instances from majority class and combines them with instances from minority class for internal test folds classification. The subsets, which show good classification performance for KNN and RF, are favored and optimized in each iteration. Those instances, which have high frequency of selection from majority class, are combined with minority class samples to create a
balanced dataset. Ultimately an optimal dataset is evolved that produces improved classification score for KNN and RF in terms of AUC. We used K=3 for KNN classifier and 200 tree size for RF in our experimentation.

The PSO balanced dataset, reduces the size of dataset to great extent and shows good performances for KNN and RF when tested with 10 fold cross validation. Further, we use mRMR as feature selection technique, which is found efficient compare to other feature selection methods. mRMR selects the features, which have strong correlation with class labels and are less redundant as well. Chr-PmRF is expected to reduce the complexities involved in customers churn prediction by handling enormous nature of telecommunication datasets.

4. Results and Discussion

The proposed Chr-PmRF approach is validated with the comprehensive experimentation conducted employing various combinations of sampling, feature selection and classification methodologies. The 10 folds cross validation testing is adopted for analyzing the performance attained during the experimentation using AUC, sensitivity and specificity based performance measures.

4.1. Performance based on Basic PreProcessing

Initially, KNN and RF are applied on original dataset without involving any sampling or feature reduction strategy. The table 1 shows, both the classifiers exhibit unsatisfactory performance in terms of AUC, sensitivity and specificity.

Table1. The performance of RF and KNN on original dataset.

The original dataset has 50000 instances with 205 features. Although the dataset is not processed for undersampling and feature reduction but useless and features with missing and nominal values are treated. The used dataset has high dimensionality and the minority class has fewer instances as compare to majority class, therefore the current state of the dataset extends weak learning that ultimately leads to suffered performances by the used classifiers. KNN is a proximity based classifier and its performance is associated with matching proximity of test samples with train samples, thus if data suffers curse of dimensionality, then KNN produces unsatisfactory results. On the other hand, RF
is considered as suitable choice for large dataset due to its characteristic of implicitly performing the feature selection but it does not produce the satisfactory results as well. The imbalanced data distribution does not provide the training level RF requires for producing desirable performance. Therefore the dataset requires a balancing act between the churners and non-churners. The dataset also has high dimensionality, which needs to be dealt with feature reduction strategies.

### 4.2. Performance based on Random Undersampling

The RUS is employed to establish a normal distribution of the instances between churners and non-churners. The RUS randomly chooses the chunk of samples from majority class, equal in numbers to minority class and establishes a balanced dataset. Such a dataset exhibits an equal distribution between minority and majority classes. After accomplishing the balanced data distribution, KNN and RF are employed to evaluate the impact on performance as shown in Table 2.

**Table 2.** The performance of RF and KNN on undersampled data.

<table>
<thead>
<tr>
<th></th>
<th>AUC Improvement</th>
<th>Accuracy Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>RF</td>
<td>0.0282</td>
<td>0.9222 to 0.6344</td>
</tr>
<tr>
<td>KNN</td>
<td>0.0395</td>
<td>0.9261 to 0.5825</td>
</tr>
</tbody>
</table>

There is a fractional improvement in AUC of 0.0282 in case of RF and 0.0395 in case of KNN, compared to classification performed on original dataset. The accuracy is deteriorated from 0.9222 to 0.6344 in case of RF and 0.9261 to 0.5825 in case of KNN. Although accuracy suffers after applying RUS but that does not reflect the real deterioration in performance of respective classifier. This is because if none of the churners is truly predicted, with all the true predictions as non-churners, the overall accuracy still reaches more then 0.80, which shows accuracy as not a suitable performance measure in this case. Thus, AUC is more relevant performance measure in our case, for precisely representing the correctly predicted churners and non-churners. Further, the improvement achieved in AUC as a result of applying undersampling shows that data balancing improves the performance.

### 4.3. Performance based on PSO based Undersampling

A PSO based undersampling method [23] evolves an optimal sample with an iterative approach by involving RF and KNN for evaluating the fitness of the every candidate sample. Multiple samples are developed by choosing instances from minority and majority classes and their fitness is evaluated with RF and KNN. The sample evolved through a process of PSO based undersampling hints for a better performance, because now the data comprises of evenly balanced instances of churners and
non-churners which are optimized for classification. This consequently extends improved learning to used classifiers.

**Table 3.** The performance of RF and KNN on PSO based sampled data.

Both the classifiers show improvements in AUC as shown in table 3. RF and KNN achieve 0.7101 and 0.6165 AUC values respectively, which show improvement in performance. The chosen subsample not only evenly represents the instances of both the classes but the balanced data distribution extends improved training level as well. This is accomplished because of KNN and RF classifiers involved in the fitness evaluation of PSO based undersampling. Therefore, the PSO based undersampling serves the idea of curtailing the size of the data along with achieving balanced data distribution for improved performance.

**4.4. Performance Analysis after Dimensionality Reduction**

The improvement in performance through undersampling encourages selecting the important features of optimally selected PSO undersampled dataset. Feature reduction can further squeeze the dataset to most meaningful and discriminating features, by eliminating noise, irrelevance and correlation of the dataset. This potentially can improve the performance.

**4.4.1. Fisher’s Ratio based Feature Selection vs PCA based Feature Selection**

PCA encompasses the interclass distance and yields the artificial principal components, which have maximum variance coverage in the initial components. First 40 principle components show highest performance of 0.6804 AUC for RF as reported in table 4.

**Table 4.** The performance of RF, on PCA reduced and PSO undersampled dataset.

The RF has suffered here in regards of performance. The AUC is deteriorated to 0.6804 compared to 0.7101 achieved in the case of PSO based optimal sampling as shown in Table 3. Such deterioration in AUC shows that although the optimized PSO based undersampling has positive impact on prediction performance but PCA’s generated components do not have potential discriminating information about churners and non-churners to boost the performance. PCA transforms the dataset into principal components to cover the maximum variance present in the data, but PCA does not
consider any dependency of data with class labels. The variance based principal components ignore the suppressed presence of the churners in the dataset. The maximum variance coverage can lead to account for the majority of the instances from non-churners only, because of their dominating presence in the dataset. Therefore, the principal components do not extend the required information for training classifiers to attain improved performance.

Fisher’s Ratio is another measure that identifies the discriminating power of each feature in the dataset. The performance levels attained by various Fisher’s ratio based selected feature subsets are shown in Table 5 and Table 6.

**Table 5.** The performance of KNN, on Fisher’s Ratio based reduced feature sets.

**Table 6.** The performance of RF, on Fisher’s Ratio based reduced feature sets.

The set of 50 exhaustively searched features produce 0.6181 AUC for KNN, which is highest as shown in Table 5. While in the case of RF, the set of only 40 exhaustively searched features produce the highest of 0.7185 AUC. Fisher’s ratio based feature selection exhibit higher performance compared to PCA, because Fisher’s ratio considers each feature in terms of its data distribution for both the classes and then computes respective significance levels. Hence, Fisher’s ratio improved feature selection is attributed to considering the class distribution of the data compared to PCA’s blind feature reduction based on variance in the data.

**4.4.2. Fisher’s Ratio based Feature Selection vs F-Score based Feature Selection**

Fisher’s ratio and F-score based feature selection are based on ranking and weighting the individual feature. Fisher’s ratio measures the discriminating power of the features by considering the means and variances of both the binary classes for the whole dataset, while the F-score method finds the discrimination of the two sets of real numbers, and it is calculated for each feature. The higher F-score value features are more discriminating. The set of 40 exhaustively searched features, on the basis of higher F-score value produces 0.6247 AUC for KNN as shown in Table 7.

**Table 7.** The performance of KNN, on F-Score based reduced feature sets.

In the case of RF, the highest 0.7214 AUC is produced by a set of only 30 exhaustively searched features as shown in Table 8.
Table 8. The performance of RF, on F-Score based reduced feature sets.

The AUC improvement is although marginal compared to Fisher’s ratio, but the numbers of selected features are reduced from 40 to 30, which significantly reduces the computational cost. F-score produces better set of features with high discriminating power. F-score is more adaptive for establishing better separability through selected features. F-score’s better ability to select the features with more discriminating potential compared to Fisher’s ratio lies, in minimizing the intraclass distance and maximizing the interclass distance of instances.

4.4.3. F Score based Feature Selection vs mRMR based Feature Selection

mRMR targets to select the feature subset, by maximizing the mutual dissimilarity within the class and minimizing the marginal similarity with the class labels. mRMR based feature selection, performed on PSO sampled dataset produces highest of 0.7511 AUC for RF with only 39 exhaustively searched features. The RF has shown significant improvement in performance, with only 39 features as shown in table 9.

Table 9. The performance of KNN, on mRMR based reduced feature sets.

Such a significant improvement in terms of AUC with fewer of the features and instances show the effectiveness of selecting the appropriate sampling and feature selection methods. The KNN classifier has shown deteriorated performance when provided with mRMR based reduced feature sets as shown in table 10, where highest of the performances is 0.6224 AUC with 20 features.

Table 10. The performance of RF, on mRMR based reduced feature sets.

This deterioration is because of the fact that mRMR optimized the proximity amongst the instances by minimizing the redundancy between features and maximizing their correlation with class labels. This process directly affects the KNN’s proximity based classification criteria where the class label is assigned to the target instance on the basis of its closest proximity with K instances from training vector. F-score and Fisher’s ratio rank the features based on mutual information without considering relationships among features while mRMR is the different feature selection method in terms that it selects the features having strong correlation with class variable and mutually different
form each other. This is the reason mRMR method substantially reduces the features set to 39 features and improves the AUC performance to the highest score of 0.7511.

4.4.4. **Performance Comparison of proposed Chr-PmRF with other existing approaches**

The proposed Chr-PmRF approach shows promising performance amongst the various combinations explored in this work, as shown in Table 11 and Figure 4.

**Table 11.** The performance comparison of highest AUC scores attained by RF and KNN

The comparison of Chr-PmRF is also made with other existing approaches as well, as given in Table 12.

**Table 12.** Performance comparison of Chr-PmRF with existing approaches

We have compared our results with the Robert BusaFekete et al. [25] approach. They have used an Adaboost based model 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 uses the base classifier, constructs the 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 12 show that the 0.7258 AUC and 0.7158 AUC are produced [27], using tree and stump based learners with Adaboost, respectively. Stochastic Gradient Boosting algorithm [29], has used the grouping modalities technique to deal with categorical values and applied boosting with decision trees as classification model. This approach scores 0.7282 AUC for churn prediction. In our work, a combination of PSO based optimized undersampling, mRMR and RF have emerged as a better classification approach compared to other two referred approaches [27] and [29]. Our Chr-PmRF has produced 0.7511 AUC that is competitive to 0.7282 AUC [29] and 0.7258 [27] as shown in Table 12. Although the AUC improvement is nominal but Chr-PmRF, essentially operates in a more systematic manner, where dataset is comprehensively processed by applying PSO based optimized sampling approach in combination with an adaptive mRMR based feature selection strategy. The effective preprocessing significantly impacts the computational cost, in this data mining problem of churn prediction. The set of only 39 features from PSO based optimal sample are chosen using mRMR, which ultimately produces 0.7611 AUC with RF using 10 folds cross validation. RF has
shown better performance in our experimentation compared to KNN, for being ensemble of decision trees and majority voting involved in making final decision.

5. Conclusions

This work validates the claim as regards classification that appropriate preprocessing and establishing the proper data distribution is vital for classification. The PSO based optimal sampling approach not only undersamples the data but optimizes the samples selection on the basis of AUC measure, to attain better classification performance. The discriminating power of the optimally selected samples is further explored by employing appropriate feature selection strategies. Where mRMR returns a set of 39 features which help RF to attain highest AUC performance. The proposed Chr-PmRF approach is a promising contribution where PSO based sampling, mRMR based feature selection and RF classifier effectively handles the problems of customer churn prediction in telecommunication. Thus, the proposed Chr-PmRF approach can be useful for the competitive telecommunication industry, where the size of the data is exceptionally large and involves high computational cost for churn prediction.

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

References


**Adnan Idris** received his MS 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 I.T, Islamabad in 2002. Further he has 7 years research & teaching experience at university level. Currently he is doing PhD 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.S. and Ph.D. degrees in Computer Systems Engineering from GIK Institute of Engineering Sciences and Technology Topi, Pakistan, in 2003 and 2006, respectively. He has spent two-years as Post-Doc Research Fellow at Department of Mechatronics, GIST South Korea. He is currently 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.

**Muhammad Rizwan** has completed his BS(CIS) degree from Pakistan Institute of Engineering and Applied sciences, Islamabad. His research interest includes computer programming, Machine Learning and Pattern Recognition.