Computers and Electrical Engineering 38 (2012) 1808–1819

Contents lists available at SciVerse ScienceDirect

Computers and Electrical Engineering

journal homepage: www.elsevier.com/locate/compeleceng

# 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,*]

[a] Department of Computer and Information Sciences, Pakistan Institute of Engineering and Applied Sciences, Nilore, Islamabad 45650, Pakistan
[b] Department of Computer Sciences and Information Technology, University of Poonch, Rawalakot 12350, AJK, Pakistan

ARTICLE INFO

*Article history:*
Received 24 October 2011
Received in revised form 3 September 2012
Accepted 3 September 2012
Available online 25 September 2012

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. Prediction 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.

© 2012 Elsevier Ltd. All rights reserved.

## 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

☆ Reviews processed and approved for publication by Editor-in-Chief Dr. Manu Malek.

* Corresponding author. Tel.: +92 51 2207381x3159.
*E-mail address:* asif@pieas.edu.pk (A. Khan).

0045-7906/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.compeleceng.2012.09.001

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) [2], 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 [8–10], Logistic Regression [10,11], Genetic Programming [12,26], Neural Network [13–16], 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] and KNN [11], 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 Fig. 1 highlights various steps involved in Chr-PmRF.

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

Download English Version:

https://daneshyari.com/en/article/454074

Download Persian Version:

https://daneshyari.com/article/454074

Daneshyari.com