A Comparative Assessment of the Performance of Ensemble Learning in Customer Churn Prediction

Hossein Abbasimehr, Mostafa Setak, and Mohammad Tarokh
Department of Industrial Engineering, K.N. Toosi University of Technology, Iran

Abstract: *Customer churn is a main concern of most firms in all industries. The aim of customer churn prediction is detecting customers with high tendency to leave a company. Although, many modeling techniques have been used in the field of churn prediction, performance of ensemble methods has not been thoroughly investigated yet. Therefore, in this paper, we perform a comparative assessment of the performance of four popular ensemble methods, i.e., Bagging, Boosting, Stacking, and Voting based on four known base learners, i.e., C4.5 Decision Tree (DT), Artificial Neural Network (ANN), Support Vector Machine (SVM) and Reduced Incremental Pruning to Produce Error Reduction (RIPPER). Furthermore, we have investigated the effectiveness of two different sampling techniques, i.e., oversampling as a representative of basic sampling techniques and Synthetic Minority Over-sampling Technique (SMOTE) as a representative of advanced sampling techniques. Experimental results show that SMOTE doesn't increase predictive performance. In addition, the results show that the application of ensemble learning has brought a significant improvement for individual base learners in terms of three performance indicators i.e., AUC, sensitivity, and specificity. Particularly, in our experiments, Boosting resulted in the best result among all other methods. Among the four ensemble methods Boosting RIPPER and Boosting C4.5 are the two best methods. These results indicate that ensemble methods can be a best candidate for churn prediction tasks.*

Keywords: *Churn prediction, data mining, classification, ensemble learning.*

Received May 19, 2012; accepted January 6, 2013; published online March 13, 2014

1. Introduction

In recent years, Due to the saturated markets and competitive business environment, customer churn prediction has received an increasing attention. The aim of customer churn prediction is detecting customers with high tendency to attrite. Therefore, there is a need to a prediction model to accurately classify churners and non-churners customers so that the firm can use its marketing resources effectively to retain the churner customers.

Technically spoken, the purpose of churn prediction is to classify the customers into two types: Customers who churn (leave the company) and customer who continue doing their business with company [11]. Gaining a new customer costs 12 times more than keeping the existing one [23]; Therefore, a small improvement on the accuracy of churn prediction can result a big profit for a company [26].

Data mining tools are used to extract the valuable information and knowledge hidden in the vast amount of data [14]. Data mining techniques have been used widely in churn prediction context such as Support Vector Machines (SVM) [10, 31, 32], Decision Tree (DT) [15], Artificial Neural Network (ANN) [21, 25], Logistic regression [9, 18]. However, there is no overall best. Data mining techniques used in building churn prediction models.

Ensemble methods [14] or methods that use a combination of models can improve the application of single data mining techniques. For instance, Chergui *et al.* [7], using ensemble learning for handwriting recognition, the performance had increased compared to the individual classifier.

Lemmens and Croux [18] showed that ensemble methods such as bagging and boosting improve accuracy in churn prediction. However, other ensemble methods such as staking [30] and voting *et al*. [22] have not been tested in churn prediction context. In this paper, we aim to systematically compare the performance of four popular ensemble learning methods, including bagging, boosting, staking, and voting by using C4.5 DT, ANN , SVM and Reduced Incremental Pruning to Produce Error Reduction (RIPPER) classifiers [29] as the base learner in churn prediction context. Besides, the effectiveness of two different sampling techniques is investigated.

The reminder of this paper is organized as follows: In section 2, we present a literature review about churn prediction. Executed methods are described in section 3. In section 4, the data preprocessing, model building and results of a series of experiments are discussed. Conclusions are considered in section 5.

2. Literature Review

Customer churn is the tendency of a customer to stop his or her business with a company in a given time period [20]. Finding the churn drivers of customer churn or model building for customer churn prediction are the aims of researches in this field [8].

Table 1. Overview of churn prediction researches.

Authors	Title	Techniques
AurélieLemmens and Christophe Croux [18]	Bagging and boosting classification trees to predict churn	Bagging, stochastic gradient boosting, binary logit model
KristofCoussement, Dirk Van den Poel [10]	Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques	Support vector machines, Logistic regression, Random forests
Jonathan Burez, Dirk Van den Poel [5]	Separating financial from commercial customer churn: A modeling step towards resolving the conflict between the sales and credit department	Random forests, survival analysis
Kristof Coussement, Dirk Van den Poel [9]	Integrating the voice of customers through call center emails into a decision support system for churn prediction	Logistic regression
Parag C. Pendharkar [21]	Genetic algorithm based neural network approaches for predicting churn in cellular wireless network services	Genetic algorithm based Neural Network approaches
Yaya Xie, Xiu Li, E.W.T. Ngai, Weiyun Ying [31]	Customer churn prediction using improved balanced random forests	ANN, DT, SVM, Random forests, Improved Balanced Random forests
Chih-Fong Tsai, Yu-Hsin Lu [25]	Customer churn prediction by hybrid neural networks	ANN, Self-Organizing Maps (SOM)
Nicolas Glady, Bart Baesens, Christophe Croux [13]	Modeling churn using customer life time value	Logistic regression, DTs and NN, cost-sensitive classifiers (AdaCost), Cost-sensitive decision tree
Kristof Coussement, Dries F. Benoit, Dirk Van den Poel [11]	Improved marketing decision making in a customer churn prediction context using generalized additive models	Generalized Additive Models (GAM), Logistic Regression
Bingquan Huang , B. Buckley , T.-M. Kechadi [15]	Multi-objective feature selection by using NSGA-II for customer churn prediction in telecommunications	Decision Tree C4.5
Adem Karahoca, Dilek Karahoca [16]	GSM churn management by using fuzzy c-means clustering and adaptive neuro fuzzy inference system	fuzzy C-means clustering and Adaptive Neuro Fuzzy Inference System (ANFIS)
Chih-Fong Tsai, Mao-Yuan Chen [25]	Variable selection by association rules for customer churn prediction of multimedia on demand	Association rules, DT, NN
Xiaobing Yu , ShunshengGuo , Jun Guo , Xiaorong Huang [32]	An extended support vector machine forecasting framework for customer churn in e-commerce	SVM, ANN, Extended Support Vector Machine (ESVM)
Wouter Verbeke , David Martens , Christophe Mues , Bart Baesens [27]	Building comprehensible customer churn prediction models with advanced rule induction techniques	Antminer+, Active Learning Based Approach (ALBA), RIPPER, SVM, Logit, C4.5
Hossein Abbasimehr, Mostafa Setak, Javad Soroor [1]	A framework for identification of high-value customers by including social network based variables for churn prediction using neuro-fuzzy techniques	K-Means, ANFIS, Locally Linear Neuro-Fuzzy (LLNF), ANN

We have reviewed recent papers that have focused on constructing a predictive model using data mining techniques for churn prediction. Table 1 summarized those researches.

3. Methods

In this section, a brief description of data mining methods used through this paper is given.

3.1. Bagging

Bagging (bootstrap aggregation) is a technique that aggregates results of n models that were built on the basis of K bootstrap sets. It is one of the earliest ensemble learning algorithms [3]. Given a set, D, of d tuples, the procedure for bagging is as follows. For iteration i (i = 1, 2,…, k), a training set, D_i, of d tuples is sampled (with replacement) from the original dataset. A classifier model, C_i is trained for each bootstrap sample D_i. Each classifier, C_i, assigns a class label for an unknown tuple, X that count as one vote. The bagged classifier C*, classify the tuple, X by taking a majority vote among the predictions made by each base classifier [14, 22].

3.2. Boosting

Boosting is one of the example of ensemble learning manipulates the training set [14, 22]. In contrast to bagging, boosting assigns a weight to each training tuple. A series of k classifiers is iteratively constructed. Once a classifier C_i is constructed, the weights of training tuples are changed so, that the subsequent classifier, C_{i+1}, focus on training tuples that were misclassified by C_i. The final ensemble classifier C*, combines the vote of each individual classifier, where each classifier's vote has a weight and it is a function of the classifier accuracy. There are several implementations of boosting algorithms which each one uses different methods for updating the weights of each tuple and combining the vote of classifiers.

Adaboost is a popular boosting algorithm that is proposed by Freund and Schapire [12]. Hence, in this paper, the Adaboost algorithm is chosen.

3.3. Staking

Staked generalization is another way of combining multiple classifiers [30]. In contrast to bagging and boosting, staking is used to combine models built by different algorithms.

Staking has two steps: In the first step, different models are built by applying different algorithms on the original dataset. Output of each model is collected into a new set of data. For each instance in the original training set, the new dataset represents every model's prediction of that instance's class, along with its true classification. The original data and models constructed during this step are referred to as level-0 data and level-0 models respectively, according to Wolpert's [30] terminology. In the second step, a meta-model are used to derive a classifier from level-0 training data. Staking is not widely used in comparison with other ensemble methods such as bagging and boosting [29].

3.4. Voting

Another way of constructing an ensemble classifier is by voting among classifiers. Each independent classifier assigns a class label for each instance. Then, using a voting scheme, the class label of each instance is determined [22].

4. Empirical Analysis

The procedure used for churn prediction in this study is shown in Figure 1. The figure illustrates the techniques and algorithms used in this study. The main goal of our study is to compare four approaches to ensemble learning such as bagging, boosting, staking, and voting to investigate how they improve the performance of models to assist with customer churn prediction.

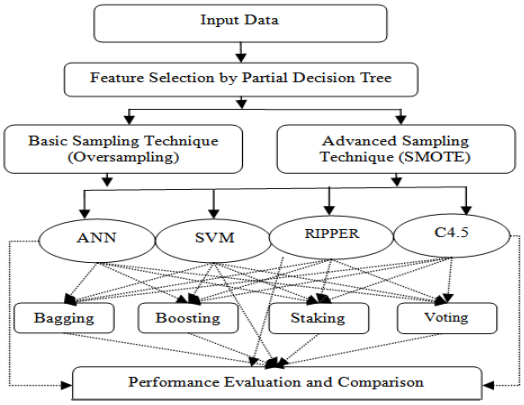

Figure 1. Research procedure.

Several algorithms can be used as base learners for ensemble learning purpose. In this paper, four common state-of-the-art algorithms including: ANN, C4.5 DT, RIPPER rule learner, and SVM are chosen. To investigate the performance of ensemble methods in customer churn prediction context, the following steps were performed; Firstly, a feature selection procedure was carried out. Secondly, we used two sampling methods for solving class imbalance problem. By doing so, two balanced datasets were obtained. Thirdly, each individual base learner was applied on the datasets. Because these algorithms are sensitive to parameter selection, the best parameter for each of them was found. After finding the best gained C4.5, RIPPER, ANN, and SVM classifiers, these classifiers were employed to build Bagging, Boosting, Staking, and majority voting ensembles. Finally, the models obtained using base learners and ensemble methods were compared in terms of evaluation criteria.

All algorithms were executed in Waikato Environment for Knowledge Analysis (WEKA) data mining software [29].

4.1. Dataset

All algorithms used in this paper are applied on a publicly available dataset downloaded from the UCI Repository of Machine Learning Databases at the University of California, Irvine[1]. The data set contains 20 variables about 5000 customers, along with a target field that indicates whether or not that customer churned (left the company). The proportion of churner in the dataset is 14.3%. For a full description of the dataset, one may refer to Larose [17]. We first split the data set into 67%/ 33% training/test set split. The proportion of churners was oversampled in order to give the classifier a better ability of prediction. Therefore, the proportion of churner and non-churner in training data set is 50%/ 50%. The test set was not oversampled to provide a realistic test set; the churn rate remained 14.3%. All models constructed during this work are evaluated on this test set.

4.2. Data Preprocessing

Data preprocessing is an essential phase in data mining. Low quality data will lead to low quality mining results. Data processing techniques, when applied before mining, can significantly improve the overall quality of the patterns mined and/or the time required for the actual mining. There are a number of data preprocessing techniques such as data cleaning, data transformation, data integration, data reduction [14]. In this paper, feature selection was performed to remove irrelevant attributes from dataset. Furthermore, sampling techniques were employed in order to make balance between positive and negative classes.

4.3. Feature Selection

We have used a novel data mining techniques, the Partial Decision Tree (PART) algorithm [29], for feature subset selection purpose. This algorithm combines the divide-and-conquer strategy for DT learning with the separate-and-conquer one for rule learning. A detailed description about the PART algorithm is given in [29]. Berger *et al*. [2] introduced feature selection by using PART algorithm. They have demonstrated that classifiers show comparable performance in their classification task when applied to the feature subset selected by using the PART algorithm. In this paper, a reduced subset of features was obtained by applying the PART algorithm on the dataset. At first, a set of decision rules is built by applying the PART on the training set. Each rule contains a number of features. All features contained in each rule are extracted. Finally, the set of reduced features is derived. The top nine features selected by algorithms are used in model building phase.

4.4. Handling Class Imbalance

Real customer churn datasets have extremely skewed class distribution. For example the dataset used in this

[1]http://www.ics.uci.edu/~mlearn/MLRepository.html

study has a skewed class distribution; such that the class distribution of churners to non-churners is 14.3:85.7.

There are several data mining problems related to rarity along with some methods to address them [4]. Sampling is one of the most widely used techniques for dealing with rarity. Sampling methods are divided into two categories: Basic sampling method and advanced sampling method. The basic sampling methods include under-sampling and oversampling. Under-sampling eliminates majority-class examples while over-sampling, in its simplest form, duplicates minority-class examples. Both of these sampling techniques decrease the overall level of class imbalance, thereby making the rare class less rare [28]. The advanced sampling methods may use intelligence when eliminating/duplicating examples. In this study, some techniques from both basic and advance sampling methods have been used. Oversampling is used as a representative of basic sampling technique, and SMOTE [6] is used as a representative of advanced sampling technique. Synthetic Minority Over-sampling Technique (SMOTE) is an oversampling approach in which the minority class is oversampled by creating synthetic examples rather than by oversampling with replacement [6]. With oversampling the churners, a predictive model can gain a better capability of discerning discriminating patterns.

4.5. Evaluation Criteria

In this study, the Area Under receiver Curve (AUC), sensitivity, and specificity [14] are used to quantify the accuracy of the predictive models.

If True Positives (TP), False Positives (FP), True Negatives (TN) and False Negatives (FN) are the TP, FP, TN and FN in the confusion matrix, then the sensitivity is (TP/(TP+FN)): The proportion of positive cases which are predicted to be positive. The specificity is (TN/((TN+FP)): The proportion of negative cases which are predicted to be negative [14].

To assess the accuracy of a classifier independent of any threshold, ROC analysis can be used. The horizontal axis and the vertical axis of an ROC curve are defined by Equations 1 and 2 respectively [14].

$$x=1-\ specificity\ (t) \quad (1)$$

$$y=sensitivity\ (t) \quad (2)$$

To measure the accuracy of a model, the AUC can be measured [4, 14].

4.6. Model Building

As mentioned before, our main motivation is to investigate the performance of ensemble methods in customer churn prediction. Besides, we aim to explore the effectiveness of two different sampling techniques. To address these research objectives, four state-of-the-art base learners including C4.5, RIPPER, ANN, and SVM have been applied on two balanced datasets obtained by using two different sampling techniques. Because these algorithms are sensitive to parameter selection, at first, the best parameter was found for each of them.

For ANN algorithm, 16 models were trained to discover a model which outperforms all others in terms of performance criteria. A three layer neural networks was chosen. The number of neurons was set to 5. In addition, the learning rate from 0.1 to 0.4 with step 0.1 and the momentum from 0.1 to 0.4 with step 0.1 were tested.

For C4.5 algorithm, the parameter, confidence factor, which is the confidence threshold for pruning, in each time was set equal to 0.25, 0.125, 0.0625, 0.031, 0.015, and 0.001.

For RIPPER algorithm, the parameter, minimum total weight of instances in a rule, in each time was set to equal to 2, 4, 6, 8, 10, and 12.

For SVM, RBF kernel was chosen since, it had shown good performance [19]. The RBF kernel function requires two parameters to be set; C and γ, with C the penalty parameter for the error term and γ as the kernel parameter. Parameter selection plays an important role in the predictive performance of SVM [10]. In this study, γ was chosen from 0.1 to 10 and C was selected from 0.01 to 100. In each step, the next parameter value is set to be 10 times larger than the previous one. After finding the best gained C4.5, RIPPER, ANN, and SVM classifiers, these classifiers were used to build Bagging, Boosting, Staking, and majority voting ensembles. All runs of experiments were repeated for 1 to 10 steps of each ensemble methods.

4.7. Results and Analyses

We start with investigating the effectiveness of two different sampling techniques. As mentioned in section 4.4, in this paper, oversampling as a representative of basic sampling technique and SMOTE [6] as a representative of advanced sampling technique were used. Therefore, two balanced datasets were obtained by using two sampling techniques. In other words, the number of churners and non-churners is the same in the two datasets. The results of utilizing the four base learners on the two datasets are summarized in Table 2 in terms of AUC, Sensitivity (Sens), and Specificity (Spec).

For C4.5, The best model (AUC=0.983, Sens=1, Spec=0.958) in terms of all performance indicators is achieved when applying it on the dataset obtained by oversampling with the CF equal to 0.25.

For RIPPER algorithm, the best model is obtained in terms of performance indicators when applying it on the dataset obtained by oversampling with the parameter, minimum total weight of instances in a rule was set to 6 (AUC=0.938, Sens=0.884, Spec=0.977).

For ANN, the best model (AUC=0.94, Sens=0.871, Spec=0.981) is gained when applying it on the dataset obtained by oversampling with the learning rate was equal to 0.3 and the momentum was equal to 0.2.

For SVM classifier, the best model is achieved when applied on the dataset obtained by oversampling with parameters C=100 and gamma=10 (AUC=0.969, Sens=0.978, Spec=0.96).

When looking at the results of all individual base learners, we observe significant improvement in classifiers performance when they are applied on dataset obtained by oversampling techniques. Indeed, the advanced oversampling technique, SMOTE doesn't increase predictive performance. This is in line with finding of Burez and Van [4] who noted that the advanced sampling technique CUBE does not increase predictive performance.

Table 2. Base learners results.

Technique	Sampling technique	AUC	Sens	Spec	Parameter value
C4.5	Oversampling	0.983	1	0.958	CF=0.25
C4.5	SMOTE	0.886	0.786	0.963	CF=0.25
Ripper	Oversampling	97.7	0.977	0.977	min total weight of instances in a rule =6
Ripper	SMOTE	88.3	0.773	0.985	min total weight of instances in a rule =8
ANN	Oversampling	0.94	0.87	0.981	Learning rate= 0.3, momentum= 0.2
ANN	SMOTE	0.92	0.844	0.936	Learning rate= 0.2, momentum= 0.2
SVM	Oversampling	0.969	0.978	0.96	RBF Gamma=10, C=100
SVM	SMOTE	0.876	0.835	0.917	RBF Gamma=1, C=100

After finding the best gained C4.5, RIPPER, ANN, and SVM classifiers, these classifiers were utilized to build Bagging, Boosting, Staking, and majority voting ensembles. All runs of experiments were repeated for 1 to 10 steps of each ensemble methods. The results are shown in Table 3.

As Tables 2, 3 and Figures 2, 3, 4, 5 show, Bagging C4.5 (AUC=0.999, Sens=1, Spec=0.984) outperforms C4.5. In addition, Bagging RIPPER (AUC=0.997, 0.942, .98) surpasses RIPPER. But, ANN performs better than Bagging ANN (AUC=0.941, Sens=0.862, Spec=0.972). Bagging SVM (AUC=0.987, Sens= 0.987, Spec=0.96) outperforms SVM.

In general, except for ANN, application of Bagging on each base learner leads to significant improvement.

Boosting ANN (AUC=0.966, Sens=0.875, Spec= 0.984) outperforms ANN; this result is in contrast to Bagging result and shows that Boosting is powerful than Bagging in churn prediction setting. Boosting C4.5 (AUC=1, Sens=1, Spec=0.986) surpasses C4.5.

Boosting RIPPER (AUC=1, Sens=1, Spec=0.988) performs better than RIPPER.

Boosting SVM (AUC= 0.966, Sens=1, Spec=0.962) also, outperforms SVM; in this case the sensitivity of model is improved. By knowing that misclassifying a churner as non-churner costs more than misclassifying a non-churner as churner, sensitivity of a churn model is the most important criterion.

Table 3. Ensemble learning results.

Ensemble method	Base learner	AUC	Sens	Specs
Bagging	**C4.5**	0.999	1	0.984
	RIPPER	0.997	0.942	0.98
	ANN	0.941	0.862	0.972
	SVM	0.987	0.987	0.96
Boosting	**C4.5**	1	1	0.986
	RIPPER	1	1	0.988
	ANN	0.966	0.875	0.984
	SVM	0.966	1	0.962
Staking	**C4.5**	0.992	0.973	0.962
	RIPPER	0.988	0.978	0.99
	ANN	0.998	0.978	0.991
	SVM	0.978	0.982	0.974
Voting	**ALL base learners**	0.998	0.978	0.984

In sum, Boosting resulted in improvement for all classifiers used in this study. It is a best candidate for churn prediction attempt. Besides, the application of Boosting has brought the biggest improvement for RIPPER Algorithm.

In staking, in each run we used the best gained C4.5, RIPPER, ANN and SVM as the base learners and one of them was the meta-learner. Staking with C4.5 as the meta-learner (AUC=0.992, Sens=0.973, Spec=0.962) doesn't differ significantly in terms of predictive performance when compared to C4.5; it classifies fewer churner correctly than C4.5; in turn it classifies more non-churner correctly than C4.5. Furthermore, Staking with RIPPER as the meta-learner (AUC= 0.988, Sens=0.978, 0.991) outperforms RIPPER. Staking with ANN as the meta-learner (AUC=0.998, Sens=0.978, Spec=0.991) performs better than ANN. Staking with SVM as the meta-learner (AUC=0.978, Sens=0.982, Spec=0.974) surpasses SVM. In sum, application of staking when the base learners were used improved the performance.

Voting (AUC=0.998, Sens=0.978, Spec=0.984) outperforms the base learner RIPPER, ANN, and SVM in terms of all performance indicators. Furthermore, it performs better than C4.5 in terms of AUC and Specificity, but it has a lower sensitivity than base learner C4.5.

By considering the experimental results we can conclude that: In general, the application of ensemble methods improves the performance of churn prediction models (see Tables 2, 3 and Figures 2, 3, 4, 5). Boosting has the best result among all other methods. Among the four ensemble methods Boosting RIPPER (AUC=1, Sens=1, Spec=0.988) and Boosting C4.5 (AUC=1, Sens=1, Spec=0.986) are the two best methods.

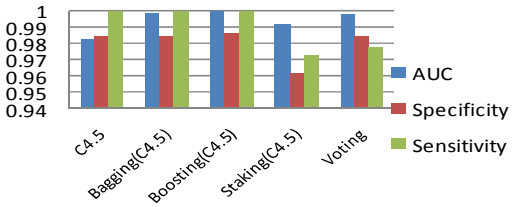

Figure 2. Predictive performance in terms of AUC, Sensitivity, and Specificity with C4.5.

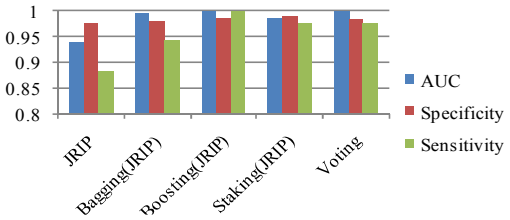

Figure 3. Predictive performance in terms of AUC, Sensitivity, and Specificity with JRIP.

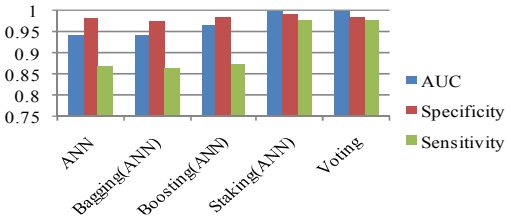

Figure 4. Predictive performance in terms of AUC, Sensitivity, and Specificity with ANN.

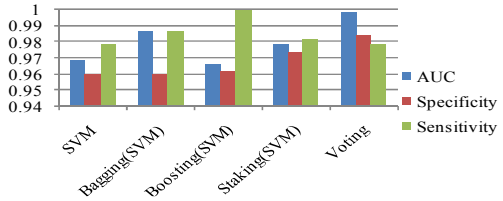

Figure 5. Predictive performance in terms of AUC, Sensitivity, and Specificity with SVM.

5. Conclusions

Many data mining techniques such as C4.5 DT, SVM, ANN, RIPPER rule learner have been used in churn prediction. One way for improving the performance of these data mining techniques is using ensemble methods. In this paper, we systematically compared the performance of four ensemble methods including Bagging, Boosting, Staking, and Voting using four commonly used base learners, i.e., C4.5, ANN, SVM, and RIPPER as base learners. Besides, the effectiveness of two different sampling techniques including basic sampling (oversampling) and advanced sampling (SMOTE) techniques was investigated. Experimental results show that SMOTE doesn't increase predictive performance. Furthermore, the application of ensemble learning has brought significant improvement for individual base learners in terms of three performance indicators i.e., AUC, sensitivity, and specificity. Particularly, in our experiments, Boosting has the best result among all other methods. Among the four ensemble methods, Boosting RIPPER and Boosting C4.5 are the two best methods. These results indicate that ensemble methods are dominated than base learners in churn prediction and are the best candidate for churn prediction tasks.

References

[1] Abbasimehr H., Setak M., and Soroor J., "A Framework for Identification of High-Value Customers by Including Social Network Based Variables for Churn Prediction Using Neuro-Fuzzy Techniques," *International Journal of Production Research*, vol. 51, no. 4, pp. 1279-1294, 2013.

[2] Berger H., Merkl D., and Dittenbach M., "Exploiting Partial Decision Trees for Feature Subset Selection in e-Mail Categorization," *in Proceedings of the ACM Symposium on Applied Computing*, New York, USA, pp. 1105-1109, 2006.

[3] Breiman L., "Bagging Predictors," *Machine Learning*, vol. 24, no. 2, pp. 123-140, 1996.

[4] Burez J. and Van D., "Handling Class Imbalance in Customer Churn Prediction," *Expert Systems with Applications*, vol. 36, no. 3, pp. 4626-4636, 2009.

[5] Burez J. and Van D., "Separating Financial from Commercial Customer Churn: A Modeling Step Towards Resolving the Conflict between the Sales and Credit Department," *Expert Systems with Applications*, vol. 35, no. 1, pp. 497-514, 2008.

[6] Chawla V., Bowyer W., Hall O., and Kegelmeyer P., "SMOTE: Synthetic Minority Over-Sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, no. 1, pp. 321-357, 2002.

[7] Chergui L., Mammar K., and Salim C., "Cobmining Neural Networks for Arabic Handwriting Recognition," *the International Arab Journal of Information Technology*, vol. 9, no. 6, pp. 588-595, 2012.

[8] Coussement K. and Van D., "Improving Customer Attrition Prediction by Integrating Emotions from Client/ Company Interaction Emails and Evaluating Multiple Classifiers," *Expert Systems with Applications*, vol. 36, no. 3, pp. 6127-6134, 2009.

[9] Coussement K. and Van D., "Integrating the Voice of Customers Through Call Center Emails Into a Decision Support System for Churn Prediction," *Information & Management*, vol. 45, no. 3, pp. 164-174, 2008.

[10] Coussement K. and Van-Poel D., "Churn Prediction in Subscription Services: An Application of Support Vector Machines While Comparing Two Parameter-Selection Techniques," *Expert Systems with Applications* vol. 34, no. 1, pp. 313-327, 2008.

[11] Coussement K., Benoit D., Van-Poel D., "Improved Marketing Decision Making in a Customer Churn Prediction Context using Generalized Additive Models," *Expert Systems with Applications*, vol. 37, no. 3, pp. 2132-2143, 2010.

[12] Freund Y. and Schapire R., "Experiments with a New Boosting Algorithm," *in Proceedings of the*

13th International Conference on Machine Learning, Bari, Italy, pp. 148-156, 1996.
[13] Glady N., Baesens B., and Croux C., "Modeling Churn Using Customer Lifetime Value," *European Journal of Operational Research*, vol. 197, no. 1, pp. 402-411, 2009.
[14] Han J. and Kamber M., *Data Mining Concepts and Techniques*, Morgan Kaufmann, San Francisco, USA, 2006.
[15] Huang B., Buckley B., and Kechadi T., "Multi-Objective Feature Selection by Using NSGA-II for Customer in Telecommunications," *Expert Systems with Applications*, vol. 37, no. 5, pp. 3638-3646, 2010.
[16] Karahoca A. and Karahoca D., "GSM Churn Management by Using Fuzzy C-Means Clustering and Adaptive Neuro Fuzzy Inference System," *Expert Systems with Applications*, vol. 38, no. 3, pp.1814-1822, 2011.
[17] Larose D., *Discovering Knowledge in Data: An Introduction to Data Mining*, Wiley, New Jersey, USA, 2005.
[18] Lemmens A. and Croux C., "Bagging and Boosting Classification Trees to Predict Churn," *Journal of Marketing Research*, vol. 43, no. 2, pp. 276-286, 2006.
[19] Martens D., Van T., and Baesens B., "Decompositional Rule Extraction from Support Vector Machines by Active Learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 2, pp. 178-191, 2009.
[20] Neslin A., Gupta S., Kamakura W., Lu J., and Mason C., "Defection Detection: Measuring and Understanding the Predictive Accuracy of Customer Churn Models," *Journal of Marketing Research*, vol. 43, no. 2, pp. 204-211, 2006.
[21] Pendharkar P., "Genetic Algorithm Based Neural Network Approaches for Predicting Churn in Cellular Wireless Network Services," *Expert Systems with Applications*, vol. 36, no. 3, pp. 6714-6720, 2009.
[22] Tan N., Steinbach M., and Kumar V., *Introduction to Data Mining*, Pearson Education, USA, 2005.
[23] Torkzadeh G., Chang J., and Hansen W., "Identifying Issues in Customer Relationship Management at Merck-Medco," *Decision Support Systems*, vol. 42, no. 2, pp. 1116-1130, 2006.
[24] Tsai C. and Chen M., "Variable Selection by Association Rules for Customer Churn Prediction of Multimedia on Demand," *Expert Systems with Applications*, vol. 37, no. 3, pp. 2006-2015, 2010.
[25] Tsai C. and Lu Y., "Customer Churn Prediction by Hybrid Neural Networks," *Expert Systems with Applications*, vol. 36, no. 10, pp. 12547-12553, 2009.
[26] Van D. and Larivière B., "Customer Attrition Analysis for Financial Services using Proportional Hazard Models," *European Journal of Operational Research*, vol. 157, no. 1, pp. 196-217, 2004.
[27] Verbeke W., Martens D., Mues C., and Baesens B., "Building Comprehensible Customer Churn Prediction Models with Advanced Rule Induction Techniques," *Expert Systems with Applications*, vol. 38, no. 3, pp. 2354-2364, 2011.
[28] Weiss M., "Mining with Rarity: A Unifying Framework," *SIGKDD Explorations*, vol. 6, no. 1, pp. 7-19, 2004.
[29] Witten H. and Frank E., *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, San Francisco, USA, 2005.
[30] Wolpert H., "Stacked Generalization," *Neural Networks*, vol. 5, no. 2, pp. 241-259, 1992.
[31] Xie X. and Ngai W., "Customer Churn Prediction using Improved Balanced Random Forests," *Expert Systems with Applications*, vol. 36, no. 3, pp. 5445-5449, 2009.
[32] Yu X., Guo S., Guo J., and Huang X., "An Extended Support Vector Machine Forecasting Framework for Customer Churn in E-Commerce," *Expert Systems with Applications*, vol. 38, no. 3, pp. 1425-1430, 2010.

Hossein Abbasimehr received his BSc degree in Information Technology engineering from Shahid Madani Azarbaijan University, Iran in 2009, and his MSc degree in e-commerce from k.N.Toosi University of Technology Tehran, Iran in 2011. He is currently a PhD student at k.N.Toosi University of Technology and his research interests include, mainly, data mining, business intelligence, and customer relationship management.

Mostafa Setak is an assistant professor in the Department of Industrial Engineering at K.N.Toosi University of technology, Tehran, Iran. He received his BSc degree in Industrial Engineering from Sharif University of Technology, Tehran, Iran in 1993. His earned his MSc degree in Industrial Engineering from Iran University of Science & Technology, Tehran, Iran in 1996. He obtained his PhD in Industrial Engineering from TarbiatModares University, Tehran, Iran, 2006. He is currently supervising graduate students in the area of Supply Chain Management, and Information Technology. He served as a system engineer, and consulted to several organizations in the area of strategic planning and Information Technology. He also possessed managerial positions such as head of the Planning and Business Development Department, and Chairperson of the Board for industrial companies.

Mohammad Tarokh is an associate professor in industrial engineering department, K.N.Toosi University of Technology, Tehran, Iran. He received his Bachelor degree in applied mathematics in 1985 from Sharif University of Technology, Tehran, Iran. He earned his Master of Science degree in 1989 in computer science from Dundee University, UK. He obtained his PhD in Computer Application in Industrial Engineering from Bradford University, UK, 1993. His research interests are in the analytical modeling of economics of information systems, business intelligence, competitive intelligence, strategic intelligence, and the impact of IT on firm strategies. His teaching interests include modeling and evaluation of computer system, supply chain management, customer relationship management, information technology, queuing systems and extended enterprise.