Customer Churn Prediction in Telecommunication with Rotation Forest Method

Mümin Yıldız

Computer Engineering Department
Yıldız Technical University
Istanbul, Turkey
Email: muminyildiz@outlook.com

Songül Albayrak

Computer Engineering Department
Yıldız Technical University
Istanbul, Turkey
Email: songul@ce.yildiz.edu.tr

***Abstract*—The main task of customer churn prediction is to estimate subscribers who may want to leave from a company and provide solutions to prevent possible churnes. In recent years, estimating churners before they leave has became valuable in the environment of increased competition among companies. The research in this paper was done to estimate churners for companies in the telecommunication industry showing how prediction efficacy is increased by balancing the data with down sampling and classifying by the rotation forest method. The performance level of these techniques are compared with Antminer and C4.5 decision tree. The comparisons are done by using the dataset taken from American Telecommunication Company and accuracy, sensitivity and specificity are used for the performance criteria.**

***Keywords*–Customer Churn Prediction; Data Mining; Telecommunication; Rotation Forest; Antminer.**

I. Introduction

Customer churn has became highly important for companies because of increasing competition among companies, increased importance of marketing strategies and conscious behavior of costumers in the recent years. Customers can easily trend toward alternative services. Companies must develop various strategies to prevent these possible trends, depending on the services they provide.

During the estimation of possible churns, data from the previous churns might be used. An efficient churn predictive model benefits companies in many ways. Early identification of customers likely to leave may help to build cost effective ways in marketing strategies. Customer retention campaigns might be limited to selected customers but it should cover most of the customer. Incorrect predictions could result in a company losing profits because of the discounts offered to continuous subscribers. Therefore, the right predictions of the churn customers has became highly important for the companies.

The prominent role that the telecommunication sector has come to occupy worldwide makes it all the more important to develop prediction mechanisms along the lines of churn prediction. Few statistics show the importance of the customer retains in this sector. One of the remarkable studies shows that 1% increase in the customer retain campaigns may result in the 5% increase in the overall values of the companies [1]. In wireless network telecommunication industry, the monthly rate of customer churn is 2.2% and the annual rate of customer churn is 27% [2]. The yearly cost of customer churn is 4 billion dollars in Europe and America, and it is 10 billion dollars in the entire world [2]. We may suppose that 1.5 million customers would stay in the same company by increasing the correct prediction at the rate of 1%. This may yield to 54 million dollars benefit for the companies annually [3].

In the literature, many researches have been conducted to increase the prediction rates of costumer churns in the telecommunication industry. The scope of this researches covers creating new models, developing existing models, combining of existing models, attribute derivation and outlier analysis techniques.

Tsai and Lu [5] used two different hybrid models to develop a customer churn prediction model. The developed hybrid model is a combination of two artificial neural networks and the second hybrid model is a combination of self organizing maps and artificial neural networks. First models are used for data reduction and second models are used for actual classifier. Kechadi and Buckley [2] used attribute derivation process to increase the correct prediction rate. Bayesian Belief Network method is tried in a study which is conducted by Kisioglu and Topcu [1]. Verbeke et al. [6] increased the accuracy by using two different rules extraction method. This methods were AntMiner+ and ALBA. Bock and Poel [7] used two different rotation based ensemble classifiers. These are Rotation Forest and Adaboosts. Yeshwanth et al. [8] suggested a new hybrid model that combines C4.5 decision tree and genetic programming. Zhao et al. [3] used one class support vector machine to increase the performance. Ghorbani et al. [9]created a new hybrid model by combining neural network, tree models and fuzzy modeling.

Ant-Miner+ algorithm is working by using the ‘divide and conquer’ technique. Firstly, it starts with all of the training data. Then it creates the best rule, which includes a subset of training data and then the best rule is added to the list of previously discovered rules. After that the samples, which are covered by this rules are removed from the training data and everything starts again with the reduced training data-set. This iteration continues until when there is only a few remaining samples in training data. At this stage, a default rule is created which covers the remaining samples.

Rotation forest method is a new generation ensemble learning algorithm. It is based on creating subsets by using principal component analysis method as a feature extraction technique [4]. In this research, it has been observed that rotation forest method gives better results than antminer+ method, which is used by Verbeke. To make comparison, the same data-set is used with Verbeke’s research and same evaluation criteria, such as accuracy, sensitivity and specificity ratios are examined. It is accepted that supposing a customer that will leave as would not leave and losing him is much more important than giving unnecessary promotions to customers who will not leave as would leave. For these reason sensitivity is seems to be more important than specificity.

The rest of the paper is organized in this following manner: Section 2 explains the rotation forest. Section 3 presents the data, data processing and evaluation criteria. In Section 4, results of rotation forest, Ant-Miner+ and C4.5 methods are compared. Finally, our conclusion is offered.

II. Rotation Forest

Rotation forest algorithm, which started to be used in literature in recent years and was put forth as the new generation on learning algorithm is based on forming a classifier ensemble by using principal component analysis, which is a feature extraction technique [4]. The basic working principle of rotation forest algorithm is similar to random forest and more than one trees are used. However, dataset that is used in the training of every decision tree in forest is determined by principal component analysis. At the phase of training of decision trees in the forest, the training dataset is divided into random subsets and features are extracted from each subset by using principal component analysis. The features that have the highest distinctiveness are determined after feature extraction. All components are considered to keep the variance of dataset same. For every classifier, the diversity is protected in classifiers ensemble by feature extraction. Basic steps of rotation forest algorithm are shown below [4].

Let X denotes training dataset, Y denotes class labels in dataset and F donates number of feature. If we suppose that training dataset includes n number of features and N number of customers, x training dataset is in from of a N by n matrix. Let Y be a vector and it shows the class label in the form of $[y_1, \ldots, y_N]$. Suppose these datasets is separated to K subsets times that is about same number with F and there are L classifiers (decision trees), which denoted by $[D_1, \ldots, D_L]$ in rotation forest according to operating principle of rotation forest algorithm. In this case, the training dataset is determined by processing steps below for every $D_{(i)}$ decision tree in rotation forest.

Step 1. F is split into K independent subsets randomly. Every independent subset should have M=n/ K features.

Step 2: Suppose that F_{ij} is the subset that consist of j feature, which was used in training of classifiers D_i and X_{ij} are the subsets that consists of features of F_{ij} in X dataset. In this case a new training dataset is determined as 75% train and 25% test of dataset by bootstrap method. After that covariance matrix C_{ij} is calculated applying principal component analysis to the newly created dataset.

Step 3: R_i transformation matrix is generated by equation (3) by using calculated covariance values.

$$R_i = \begin{bmatrix} a_{i,1}^{(1)}, a_{i,1}^{(2)}, ..., a_{i,1}^{(M_1)} & [0], & \cdots & [0] \\ [0] & a_{i,2}^{(1)}, a_{i,2}^{(2)}, ..., a_{i,2}^{(M_2)}, & \cdots & [0] \\ \vdots & \vdots & \ddots & \vdots \\ [0] & [0] & \cdots & a_{i,2}^{(1)}, a_{i,K}^{(2)}, ..., a_{i,2}^{(M_K)} \end{bmatrix}$$

Every column of R_i matrix is rearranged according to original feature and rotation matrix R_i^a is obtained. Consequently, XR_i^a dataset, which will be used in the training of D_i classifiers is obtained. This process steps are applied for every classifier in rotation forest. Classification results, which belongs to every decision tree in the forest by using transformed dataset takes a vote.

III. Methodology

This section describes the properties of datasets, preprocessing steps and evaluation criteria.

A. Dataset

Larose's dataset is obtained from wireless network telecommunication company, the dataset consists of 5000 customers information. Every customer has 21 feature and there is no missing data. The amount of churn customers is 14.3% of total customer for the coming tree months. More information about Larose dataset can be found on Larose (2005) [10]. The best 10 features of the Larose dataset that was selected by the information gain technique and their explanations are shown at Table 1.

TABLE I. BEST 10 FEATURES.

Feature Name	Feature Description	Value
international_plan	International call usage	Yes/No
total_day_minutes	Daily total talk time	Minutes
number_customer_service_calls	Number of call to customer service	
voice_mail_plan	Voice mail usage	Yes/No
total_eve_minutes	Total talk time in evening	Minutes
state	Living place	
total_day_charge	Daily Total spent credits	
number_vmail_messages	Number of voice messages	
total_intl_calls	Total number of international call	
total_intl_charge	Total spent credits of international calls	

B. Data Preprocessing

Classification tends to be in favor of majority classes when there exists unbalance in the dataset. Distribution of the used dataset has 14.3% churn customers and 85.7% non churn customers. For this reason, down sampling process is applied to dataset. In the down sampling process subsets are generated that will have x times churn customer and 2x times non churn customer. To generate subsets, firstly 20 empty sets are created and all churn customers are added to these sets. Non churn customers are selected as randomly. The important point in here is one non churn customer can be chosen more than one time for a subset. By these methods all subsets are created and they have 2121 customers. 707 of them are churn (33.3%) and 1414 of them are non churn (66,7%) customers. This process increased the sensitivity rate significantly. Comparisons are made with the average ratios of these 20 subset.

C. Evaluation Criteria

Classification results of the used techniques were compared by using class confusion matrix that is shown at Table 3.

The cost per customer is decreasing while the total number of the customer is increasing for serving companies in the telecommunications sector. So, instead of supposing a customer that will leave as would not leave and losing him(FN), unnecessary promotions might be given to a customer by supposing the customer that will not leave as would leave (FP) (Keeping FN low is more important than keeping FP low. That mean sensitivity is more important than specificity.)

TABLE II. CLASSIFICATIN RESULTS.

	Technique	Accuracy (%)	Sensitivity (%)	Specificty (%)
Original Dataset	Rotation Forest	**95.68**	**73.4**	99.49
	AntMiner+ [6]	90.85	37.09	**99.71**
	C4.5 [6]	93.59	64.93	98.34
Down Sampling	Rotation Forest(Subsets Average)	92.49	**84.57**	96.46
Oversampling	AntMiner+ [6]	**93.15**	65.76	**97.72**
	C4.5 [6]	91.66	80.82	93.45

TABLE III. CLASS CONFUSION MATRIX.

		PREDICTED		TOTAL
		Churn Customer	Non Churn Customer	
ACTUAL	Churn Customer	TP True Positive	FN False Negative	Actual Positive Number
	Non Churn Customer	FP False Positive	TN True Negative	Actual Negative Number
TOTAL		Predicted Positive Number	Predicted Negative Number	Total Customer Number

Sensitivity and specificity rates are observed because of primary focus is to find churn customers. Accuracy rates do not show the truth when there is unbalance between classes but this is also shown in the table. Sensitivity shows the rate of correctly estimated churn customers and specificity shows the rate of correctly estimated non churn customers. Accuracy sensitivity and specificity ratios are shown in equation (1) - (3).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$Sensitiviy = \frac{TP}{TP+FN} \quad (2)$$

$$Specificity = \frac{TN}{TN+FP} \quad (3)$$

Sensitivity is more important than specificity because finding churn customer more important. The best classification results are shown in bold.

IV. RESULTS

Rotation forest method is compared with antminer+, which is defended by Verbeke to evaluate the results. Principal component analysis is used for feature extraction, and the C4.5 decision tree is used as classifier in rotation forest technique. Additionally, to compare with classical methods, results of C4.5 decision tree, which was used by Verbeke is shown in Table 2 with their own findings. Antminer+ and C4.5 methods were calculated in the original dataset and in the oversampled dataset by Verbeke [6].

The classification results by using rotation forest method in entire dataset and average of classification results by using rotation forest method in twenty subsets, which were created by down sampling method are shown in Table 2. The best results are shown in bold. This results are compared with antminer+ and C4.5 decision tree algorithms.

According to original dataset rotation forest gave the best result with 73% of success considering sensitivity rate. This method is 8.47% more successful than C4.5 decision tree method and 36.31% than antminer+. Rotation forest algorithm is 0.22% lower than antminer+ in specificity rate. The difference is negligible. Rotation forest is better by 1.15% than C4.5 in specificity rate. Rotation forest has achieved 95.68%, C4.5 has achieved 93.59% and antminer+ has achieved 90.85% success in accuracy rate. According to this results rotation forest is the best technique.

For the down sampled data, average of classification results of 20 subsets that were classified by rotation forest and were balanced by down sampling has achieved 84.57% sensitivity rate in balanced dataset. Classification result of antminer+ algorithm has calculated 65.76% and classification result of C4.5 algorithm has calculated 80.82% in sensitivity rate with oversampled data. Antminer+ has achieved best result with 97.72% in specificity rate. The result of rotation forest algorithm is 96.46% and it is worse than antminer+ algorithm. However, the sensitivity rate is more important than the specificity rate for customer churn prediction. So success achieved with rotation forest is better than antminer+ and C4.5 decision tree. Antminer+ has achieved 92.49% in accuracy rate. Rotation forest has achieved 92.49%. Antminer+ is better by a small difference of 0.66 than rotation forest in accuracy rate.

Rotation forest algorithm is the best method for this dataset because of accuracy does not reflect the truth as explained in Section 3.C and sensitivity rate is more important than specificity rate. Moreover, balancing data process is a very important factor to find correct churn customer. Sensitivity is increased 11% by using rotation forest method after data balancing. The data balancing process is giving realistic and reliable results although the accuracy rate decreases.

Lets compare rotation forest and antminer+ methods financially. Suppose that a company has one million customer and 15% of them will leave. Rotation forest method will find churn customers 20% more correct according to the antminer+ algorithm. That mean is to keep more 30,000 customer in the company. Lets assume that the average bill in America is $ 100, the telecommunication company will win three million dollars yearly by the rotation forest method by this way.

V. CONCLUSION

According to the results, rotation forest is better than C4.5 decision tree and antminer+ because increasing of true prediction of churn customer rate is more important. The difference of between rotation forest and antminer+ algorithms is 36.31% in original dataset for sensitivity rate. Balancing data is increased all sensitivity rates. According to this results rotation forest method is the best algorithm and 18.81% more successful than antminer+ in terms of sensitivity.

References

[1] P. Kisioglu and I. Y. Topcu, ”Applying Bayesian belief network approach to customer churn analysis: a case study on the telecom industry of Turkey.” Expert Systems with Applications 38, 2010, pp. 7151-7157.

[2] B. Huang, M. T. Kechadi, and B. Buckley, ”Customer churn prediction in telecommunications.” Expert Systems with Applications, 39(1), 2012, pp. 1414-1425

[3] Y. Zhao, B. Li, and X. Li, ”Customer churn prediction using improved one-class support vector machine.” Lecture Notes in Artificial Intelligence, 3584, 2005, pp. 300—306

[4] J. J. Rodrigez, L. l. Kuncheva, and J. A. Carlos, ”Rotation Forest; A New Classifier Ensemble Method.” IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(10), 2006, pp. 1619–1630.

[5] C. F. Tsai and Y. H. Lu, ”Customer churn prediction by hybrid neural networks.” Expert Systems Application, 36(10), 2009, pp. 12547—12553, doi:

[6] W. Verbeke, D. Martens, C. Mues, and B. Baesens, ”Building comprehensible customer churn prediction models with advanced rule induction techniques.” Expert Systems with Applications, 38, 2011, pp. 2354—2364.

[7] K. W. Bock and D. V. Poel, ”An empirical evaluation of rotation-based ensemble classifiers for customer churn prediction.” Expert Systems with Applications, 38(10), 2011 pp. 12293-–12301.

[8] V. Yeshwanth, V. V. Raj, and M. Saravanam, “Evolutionary churn prediction in mobile networks using hybrid learning”, Proc. of XXIV Florida Artificial Intelligence Research Society Conference, 2011, pp. 471– 476.

[9] A. Ghorbani, F. Taghiyareh, and C. Lucas, ”The application of the locally linear model tree on customer churn prediction.” Proceedings of the International Conference of Soft Computing and Pattern Recognition (SOCPAR’09), Malaysia, 2009, pp. 472—477.

[10] D. Larose,(2005). Discovering knowledge in data:An introduction to data mining. New Jersey, USA: Wiley