ORIGINAL RESEARCH

Analysis of imbalanced data set problem: The case of churn prediction for telecommunication

Chun Gui*

College of mathematics and computer science, Northwest University for Nationalities, Lanzhou, China

Received: August 4, 2016 **Accepted:** April 23, 2017 **Online Published:** June 27, 2017
DOI: 10.5430/air.v6n2p93 **URL:** https://doi.org/10.5430/air.v6n2p93

Abstract

Class-imbalanced datasets are common in the field of mobile Internet industry. We tested three kinds of feature selection techniques-Random Forest (RF), Relative Weight (RW) and Standardized Regression Coefficients (SRC); three kinds of balance methods-over-sampling (OS), under-sampling (US) and synthetic minority over-sampling (SMOTE); a widely used classification method-RF. The combined models are composed of feature selection techniques, balancing techniques and classification method. The original dataset which has 45 thousand records and 22 features were used to evaluate the performances of both feature selection and balancing techniques. The experimental results revealed that SRC combined with SMOTE technique attained the minimum value of Cost = 1085. Through the calculation of the Cost on all models, the most important features for minimum cost of telecommunication were identified. The application of these combined models will have the possibility to maximize the profit with the minimum expenditure for customer retention and help reduce customer churn rates.

Key Words: Churn prediction, Class-imbalanced dataset, Random forest, Synthetic minority over-sampling, Cost, Customer retention

1. Introduction

Classification with imbalanced datasets is important as it is excellent in many real-world data mining applications, such as fraud detection, medical diagnosis, and network intrusion and so on. On the ICDM in 2005, imbalanced type of data mining is listed as one of the top ten challenging problem in the field of data mining.[1] Decreasing the customer churn is one of the long term goals of any telecom company in China and around the world. Telecom Company produces the loss customers each month, especially now the market competition is becoming more and more intense, the cost is increasing to absorb new customers since the cost of obtaining a new customer is 4 to 6 times cost compared to retain an old customer. The success rate of marketing products to new customer is 15%, while for existing customer is 50%.[2] It indicates that customer retention in comparison to developing new ones is dramatically less expensive.

Many combined modeling methods were used to help improve prediction accuracy, planning for retention, to fundamentally understand the reason of customer churn and radically reduce customer churn rate. Traditional methods have been used to predict customers' churn and identify factors that most related with churn for many years. Keramati et al.[3] presented a compounded methodology which made a lot of improvement on the value of some of the evaluations indexes, such as Recall and Precision showed above 95% accuracy are easily achievable. Kisioglu and Topcu[4] proposed a model by Bayesian Belief Network to recognize the behaviors of customers with a tendency to churn, and obtained the

*Correspondence: Chun Gui; Email: guichun2103@163.com; Address: College of mathematics and computer science, Northwest University for Nationalities, Lanzhou, 730000, China.

most important variables related to customer churn. Hung et al.[5] studied customer churn in Taiwan market, indicated that neural network method gives better results than decision trees. Idris et al.[6] put forward methodology based on PSO, mRMR (Minimum Redundancy and Maximum Relevance), and Random Forest (RF) termed as Chr-PmRF, performs well for churn prediction.

2. Study motivation

Classification with imbalanced datasets is listed as one of the top ten challenging problem in the field of data mining in 2005 ICDM (International Conference on Data Mining series).[1] Churn prediction dataset is a difficulty to classifiers, which hypothesizes an almost balanced class distribution. The sample distribution of imbalance tend to make the traditional machine learning classification method is heavily tilted towards majority category in the classification process, so the classification performance degradation. Therefore, the customer churn is rising as a practical problem to be solved urgently. In this paper, the churn class is the minority one, and the non-churn class is the majority one.

2.1 Representation of telecom imbalanced data

In general, classifiers are used to maximum a global measure of accurate, which has nothing to do with the class distribution and generates a good performance on the majority class, attracting a few attentions to the minority class. As a result, minority classification produce more fallibility than majority one, as a large proportion of errors are focused on the minority class.[7]

Imbalanced datasets in telecom company have become a significant issue to churn prediction to common classifiers, which assume an almost balance class distribution.[8] This assumption is not consistent with the actual situation. Focus on telecom data collected in practice, we found that the number of churn and non-churn gap between the number of samples is very large. The sample distribution of imbalance tend to make the traditional machine learning classification method is heavily tilted towards multi sample categories in the classification process, so that the classification performance degradation. Therefore, the churn prediction of telecom data for imbalanced classification problem is a practical problem to be solved urgently. Standard classifiers for telecom data are developed to minimize minority samples (non-churn class) measure of error, which is independent on the class distribution and produces a bias to the non-churn class, drawing less attention to the churn class.

Binary classification, for example, describes a classifier to predict the results can be expressed by the confusion matrix in Table 1.

Table 1. Confusion matrix of four terms of measure

	Prediction outcome		Total
Actual value	TP	FP	P
	FN	TN	N
Total	P’	N’	

The error parts of the classifier predicted is represented by the FP and FN. We can see that the traditional classifiers are concerned with how to make the value of (FP + FN) as small as possible, in order to gain higher classification. However, traditional classifier for FP, FN two parts and the proportion of each in the wrong instance without any consideration and restrictions, that is to say, the traditional classification problem is actually based on the following assumptions: classifier make FN judgment and make FP judgment for the influence of the actual results are the same. Based on this assumption, there is no need to make restrains for two kinds of misjudgment proportion. Real production and operation, however, does not conform to the hypothesis. This is the focus to research: the influence of imbalanced data on telecom customer churn prediction.

Table 2. Cost matrix for telecommunication customer churn

	Prediction positive	Prediction negative
True positive	c_{00}	c_{01}
True negative	c_{10}	c_{11}
Total	True positive = TP + FN	True negative = FP + TN

2.2 The importance of imbalanced data set in telecom customer churn prediction

Normally, telecom companies use machine learning classifiers to estimate a customer will churn or not. Obviously, his is a binary classification problem. Considering from the point of machine learning, the classification results should be reduced to the minimum classification cost (the cost of algorithm is expressed by $C_{\text{algorithm}}$). However, the result in true operation should be to ensure the minimum loss for telecom company (the cost of true is expressed by C_{true}). On the premise of all the errors cost are equal to the telecommunication, obviously $C_{\text{algorithm}}$ and C_{true} are equal. In the actual situation, the cost of obtaining a fresh customer is 4 to 6 times than to retain an old ones, so the FP judgment, it means the loss caused by the judgment of FP is far more serious compared to the FN. Table 2 describes the different cost of telecommunication for churn customer and non-churn customer. The following equation expresses the actual cost for telecommunication.

$$C_{\text{ture}} = |FP| \times c_{10} + |FN| \times c_{01} \tag{1}$$

Table 3. Prediction models

No.	Name	Description	
		Feature selection	**Data processing**
1	RF-OD-RF	Random forest	Original data
2	RW-OD-RF	Relative weight	Original data
3	SRC-OD-RF	Standardized regression coefficients	Original data
4	RF-OS-RF	Random forest	Over-sampling
5	RW-OS-RF	Relative weight	Over-sampling
6	SRC-OS-RF	Standardized regression coefficients	Over-sampling
7	RF-US-RF	Random forest	Under-sampling
8	RW-US-RF	Relative weight	Under-sampling
9	SRC-US-RF	Standardized regression coefficients	Under-sampling
10	RF-SMOTE-RF	Random forest	Synthetic Minority Over-sampling Technique
11	RW-SMOTE-RF	Relative weight	Synthetic Minority Over-sampling Technique
12	SRC-SMOTE-RF	Standardized regression coefficients	Synthetic Minority Over-sampling Technique

Table 4. Summary of features for the telecommunication customer data set

No.	Variable name	Description
1	USER.ACT.TYPE	The user types
2	INNET.MONTHS	The net month
3	TATAL.FEE	ARPU
4	TOTAL.FLUX	Flow
5	LOCAL.TIMES	The total time of local call
6	TOLL.TIMES	The total time of long distance call
7	ROAM.TIMES	The total time of roaming call
8	ZHUJIAO.TIMES	The total time of calling
9	TOTAL.FEE.RATE	The growth of total cost
10	TOTAL.FLUX.RATE	The growth of total flow
11	FLUX.BHD	An indicator of flow
12	TOLL.TIMES.RATE	The growth of long distance call
13	ROAM.TIMES.RATE	The growth of roaming call
14	ZHUJIAO.TIMES.RATE	The growth of calling
15	OWE.MONTH	Overdue month
16	OWE.FEE	Owe fees
17	CALL.DURA.BHD	An indicator of calling
18	LOCAL.TIMES.RATE	The growth of local call
19	IS.LOST	Whether the loss.1: loss, 0: not loss
20	MONTH.ID	Month ID
21	PROV.ID	Province ID
22	USER.ID	Customer ID

3. Methodology

3.1 Model

In this study, three kinds of feature selection methods-RF,[9] Relative Weight (RW)[10, 11] and Standardized Regression Coefficients (SRC), and three kinds of imbalanced dataset techniques: Random Over-sampling, Random Under-sampling and synthetic minority over-sampling (SMOTE), are used to create prediction models. RF is used to test the models' performance. As a result, we obtain 12 different types of prediction models, and each model containing seven experimental processes for feature selection. The names of models are in Table 3.

3.2 Dataset

The used dataset contains 450,000 customer records. The feature of dataset is clearly described in Table 4. Obviously the class variable is IS.LOST. Looking into data set, we saw that there are 100,000 records are labeled as IS.LOST = 1 and the rest 350,000 records, are labeled as IS.LOST = 0. By testing, 57,944 records were deleted, because they contained null values. The experiment is based on the customers whose ARPU is ranked in top 20% in whole data set. So (450,000-57,944)* 20% = 78,411 records are selected for experiment. In these 78,411 records, there are 10,182 records with the class label IS.LOST = 1 and the rest 68,229 records with the class label IS.LOST = 0. Imbalanced ratio is 10,182:68,229 = 1:6.7. In order to maintain the experimental results in an appropriate random environment, for each algorithm we optionally select 80% of dataset as training set and the rest 20% as test set. By the aforesaid process we prepared 12 combined models to implement and compare experiments. The data preprocess mainly analyze the variable importance (short for VI) of explanatory variables with the class variable IS.LOST. As the variables of MONTH_ID, PROV_ID, USER_ID have no association with class variable, these three explanatory variables can be removed firstly. Then we computed three kinds of correlate values for each two variables- pearson coefficient, spearman coefficient and kendall coefficient. To be fair, the mean value of these three coefficients is computed. Feature group has a correlation coefficient greater than 0.5 are LOCAL.TIMES.RATE and TOLL.TIMES.RATE, CALL.DURA.BHD and ZHUJIAO.TIMES, we select TOLL.TIMES.RATE instead of LOCAL.TIMES.RATE, ZHUJIAO.TIMES instead of CALL.DURA.BHD. The followed analysis is based on remaining 16 explanatory variables and class variable IS.LOST.

3.3 Feature selection

The purpose of this study is to compare the performance of twelve prediction combined models within the telecommunication context. In data preprocessing stage, three kinds of feature selection method are used. They are respectively RF, Relative Weight and Standardized Regression Coefficients.

3.4 Data sampling techniques

It is common to use data sampling techniques to conquer the class imbalanced issue. There are two kinds of methods be pursued- either Instance Selection random from original data set[12] or Instance Generation with the imbalanced data set.[13] In experiment we used Instance Selection: Random Under-sampling and Random Over-sampling, and Instance Generation: SMOTE.

3.5 Classification algorithm

RF is an ensemble learning algorithm for classification tasks by constructing a multi decision trees at training stage and outputting the mode of the classification or regression (mean prediction) of the individual trees. Random forest corrects for decision trees' habit of over fitting to their training set. Leo Breiman and Adele Cutler developed the method for inducing a random forest.[14] Another algorithm combined Breiman's "bagging" ideology and the features random selection, introduced by Ho[15] and Amit and Geman[16] in order to construct a collection of decision trees with controlled variance. As the RF classifier provides a solid foundation in terms of the versatility, robustness and performance, we used RF to perform classification and prediction.

Table 5. The distribution of four data sets

Processing method	IS.LOST=0	IS.LOST=1	Total	Imbalanced ratio
Original data set (OD)	68229	10182	78411	6.7:1
Over-sampling (OS)	68229	61092	129321	1.1:1
Under-sampling (US)	10182	10182	20364	1:1
SMOTE	61092	61092	122185	1:1

3.6 Evaluation criteria

The traditional criteria for churn prediction are found to have limited value in imbalanced data set. In order to evaluate the performance of 12 combined models, we used some appropriate indexes for imbalanced data set. These indexes are computed by the confusion matrixes (see Table 1). The performance metrics used in evaluating and comparing combined models are: Cost, Precision, Recall, F-score and Accuracy. Cost is the most important criteria for it determines the real investment and the value of investment. Accuracy is the commonly used evaluation standard of classification, which reflects classifier's performance for data set as a whole, but can't correctly reflect the imbalanced data classification performance. In view of the imbalanced data set, commonly used criteria are: cost, precision, recall, F-score, and Accuracy.

Table 6. The results of RF classification performance measures for all models

Model criteria	RF-OR -RF	RW-OR -RF	SRC-OR -RF	RF-OS -RF	RW-OS -RF	SRC-OS -RF	RF-US -RF	RW-US -RF	SRC-US -RF	RF-SMOTE -RF	RW-SMOTE -RF	SRC-SMOTE -RF
Cost	2395	2420	2337	3385	3326	2576	2849	2902	3128	5201	6415	**1085**
Precision (YES)	0.7030	0.6860	0.6902	0.9474	0.9489	**0.9598**	0.7694	0.7611	0.7502	0.9290	0.9100	**0.9298**
Recall (YES)	0.2193	0.2095	0.1828	1.0000	1.0000	**1.0000**	0.7939	0.7892	0.7912	0.9318	**0.9449**	0.9335
F-score (YES)	0.3332	0.3154	0.2891	0.9730	0.9735	**0.9792**	0.7710	0.7550	0.7702	0.9304	0.9266	**0.9316**
Preci-sion (NO)	0.8943	0.8943	0.8883	1.0000	1.0000	**1.0000**	0.7821	0.7857	0.7693	0.9310	0.9420	**0.9828**
Recall (NO)	0.9892	0.9882	**0.9920**	0.9506	0.9513	0.9627	0.7582	0.7572	0.7255	0.9281	0.9057	**0.9856**
F-score (NO)	0.9389	0.9389	0.9353	0.9747	0.9748	0.9808	0.7611	0.7712	0.7467	0.9296	0.9230	***0.9842***
Accuracy of original data set	0.8867	0.8867	0.8867	0.9761	0.9761	**0.9761**	0.7577	0.7577	0.7577	0.9319	0.9319	**0.9319**
Accuracy after feature selection	0.8875	0.8875	0.8814	0.9737	0.9742	**0.9800**	0.7656	0.7731	0.7590	0.9300	0.9248	**0.9717**

4. Experiment

4.1 Experimental data set

The distributions of the four data sets are in Table 5. It is obviously that we obtained balanced data set after imbalanced technology.

4.2 Experimental results

The results of 12 models on 9 performance measures are in Table 6. Figures 1 to 4 present the Cost, Precision, Recall, F-score, and Accuracy of the combined models.

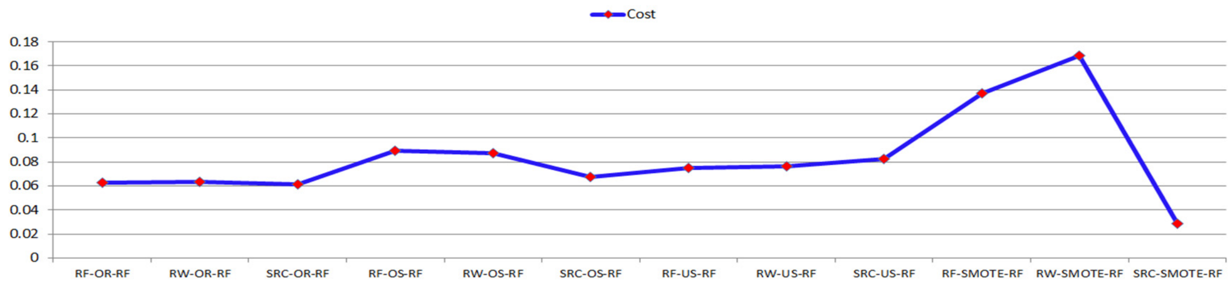

Figure 1. The cost of the combined models

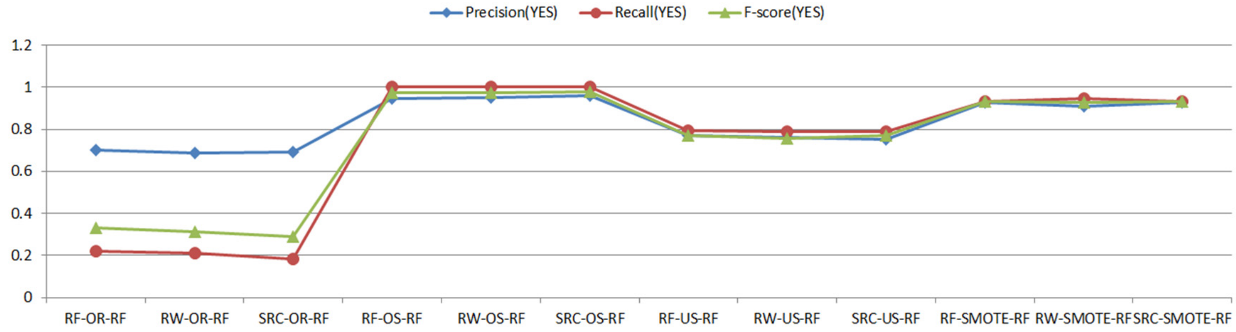

Figure 2. The precision (YES), recall (YES) and F-score (YES) of the combined models

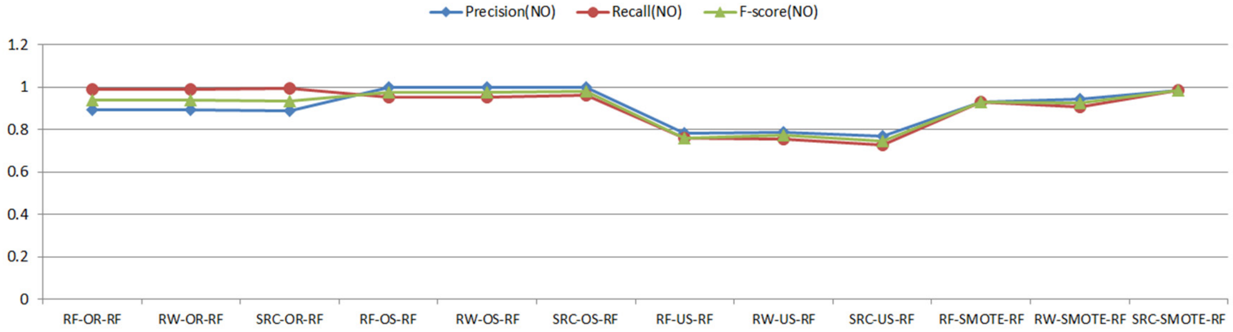

Figure 3. The precision (NO), recall (NO) and F-score (NO) of the combined models

Figure 1 is the line chart of Cost for each combined model. SRC-SMOTE-RF model is the best one for indicator of Cost. Low Cost illustrates good performance of model in classification and prediction. At the same time, low cost can maximize income for telecommunication. Figure 2 is the line chart of Precision, Recall and F-score of class label is IS.LOST = “YES” for each combined model. These three values are the bigger the better. The best value appeared in the combined model of Over-sampling and SMOTE. Figure 3 is the line chart of Precision, Recall and F-score of class label is IS.LOST = “NO” for each combined model. These three values are the bigger the better. All the values decreased slightly, this is because the minority class number increases relatively after data-balancing processing, which weakened the overwhelm affection of the majority class on the performance of the model. We judged SMOTE combined model is the best one. Compared with Figure 2 and Figure 3, the better improved performance appeared in the class of label = “YES”. This is just the results what we hope to get, because the purpose of customer churn’s study prediction is to find the possible loss customer.

Figure 4 is the line chart of accuracy with different data-balancing technology and feature selection. We found there is a little improvement of accuracy after feature selection than original data set. Especially for SRC-SMOTE-RF combined model, there is obvious perfection on the criteria of accuracy. The feature selection result of SRC-SMOTE-RF is $\{1, 3, 4, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16\}$.

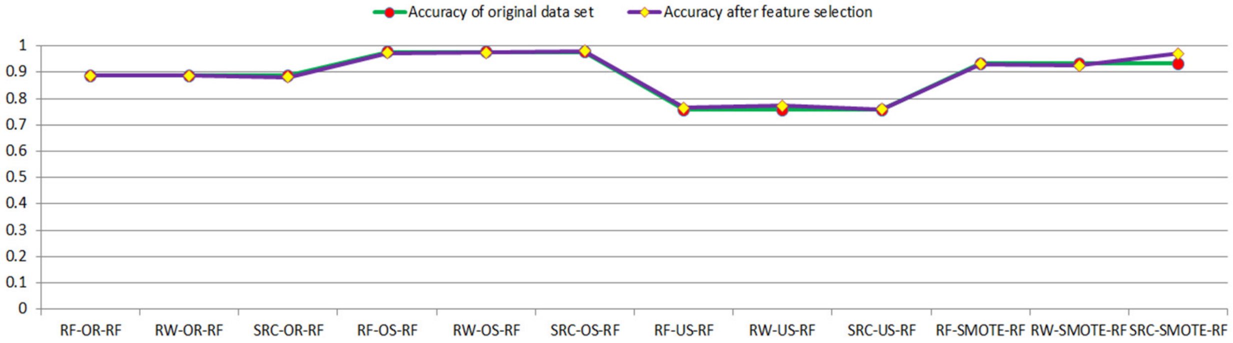

Figure 4. The accuracy of original data set and after feature selection data set

4.3 Experimental discussion

The best results come from Over-sampling technology. There are many values reached 100%, and almost all of the values for models RF-OS-RF and RW-OS-RF show digressive trend. This reflects serious over fitting phenomenon. The reason for this result is the inherent disadvantage of Over-sampling technology. In model SRC-OS-RF, the best values of result come from the attribute set of $\{1, 3, 4, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16\}$. This not only reflects the combination of SRC attribute selection with Over-sampling technology can effectively avoid over fitting, but also gets good performance in prediction.

The integral prediction performance for dataset after Under-sampling technology is the worst. The reason for this result is the inherent disadvantage of Under-sampling technology. When the Under-sampling technology is used, we lost much effective information for original dataset.

The integral prediction performance for dataset after SMOTE technology is slightly inferior to Over-sampling technology, but the values of F-score and Cost are the optimal in all of twelve combined models. In three of SMOTE technology models: RF-SMOTE-RF, RW-SMOTE-RF, SRC-SMOTE-RF. SRC-SMOTE-RF is the best one, at the same time, SRC-SMOTE-RF is the optimal in all of twelve combined models. It overcomes the disadvantage of over fitting, maximize the improvement of the accuracy, and minimize the cost. The performance of the whole SRC-SMOTE-RF model is stable, and we find that attributes group of $\{1, 3, 4, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16\}$ is most effective for customer churn. The order of influence from high to low are USER.ACT.TYPE, ZHUJIAO.TIME, OWE.FEE, TOTAL.FEE.RATE, FLUX.BHD, ROAM.TIMES.RATE, OWE.MONTH, TOLL.TIMES.RATE, TOTAL.FLUX, TOLL.TIMES, TOTAL.FLUX.RATE, ZHUJIAO.TIMES.RATE, ROAM.TIMES, TOTAL.FEE, the removed attributes are LOCAL.TIMES, INNET.MONTHS.

5. Conclusions

This study has presented a comparison among twelve combined models for imbalanced data set. It has contrasted different evaluation standards such as Cost, Precision, Recall, F-score and Accuracy for imbalanced classification. The experiment is divided into three stages. Firstly, three kinds of feature selection methods are used to remove the attribute with low importance, they are RF, Relative Weight and Standardized Regression Coefficients. Secondly, two methods of Instance Selection and Instance Generation with three kinds of technology- Random Over-sampling, Random Under-sampling, SMOTE are used to complete processing for imbalanced data set. Lastly, RF algorithm is used as base, which is a well-known decision tree ensemble eminent for its versatility, robustness and performance.

Acknowledgements

This work is partially supported by National Natural Science Foundation of China (No. 61562075). The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

References

[1] He HE, Garcia EA. Learning from Imbalanced Data, IEEE Transactions on Knowledge and Data Engineering In Knowledge and Data Engineering. 2009; 21(9): 1263-84. https://doi.org/10.1109/TKDE.2008.239

[2] Liu J. The outlook for 2013 mobile internet companymunication service trends. China's information industry network-people's post and telecommunications. 2013: 1-6.

[3] Keramati A, Jafari-Marandi R, Aliannejadi M, et al. Improved churn prediction in telecommunication industry using data mining techniques. Applied Soft Computing. 2014; 24: 994-1012. https://doi.org/10.1016/j.asoc.2014.08.041

[4] Kisioglu P, Topcu YI. Applying Bayesian Belief Network approach to customer churn analysis: A case study on the telecom industry of Turkey. Expert Systems with Applications. 2011; 38: 7151-7. https://doi.org/10.1016/j.eswa.2010.12.045

[5] Hung S, Yen D, Wang H. Applying data mining to telecom churn management. Expert Systems with Applications. 2006; 31: 515-24. https://doi.org/10.1016/j.eswa.2005.09.080

[6] Idris A, Rizwan M, Khan A. Churn prediction in telecom using Random Forest and PSO based data balancing in combination with various feature selection strategies. Computers and Electrical Engineering. 2012; 38: 1808-19. https://doi.org/10.1016/j.compeleceng.2012.09.001

[7] Japkowicz N, Stephen S. The class imbalance problem: A systematic study. Intelligent Data Analysis Journal. 2002; 6(5): 429-50.

[8] Weiss GM. Mining with rarity: A unifying framework. SIGKDD Explorations. 2004; 6(1): 7-19. https://doi.org/10.1145/1007730.1007734

[9] Breiman L. Random forests. Mach. Learn. 2001; 45(1): 5-32. https://doi.org/10.1023/A:1010933404324

[10] Johnson JW. Detemining the statistical significance of relative weights. Psychological Methods. 2009; 4(1): 387-99.

[11] James M. Multivariate relative importance: extending relative weight analysis to multivariate criterion spaces. Journal of Applied. 2008; 93(2): 329-45.

[12] García S, Derrac J, Cano J, et al. Prototype selection for nearest neighbor classification: taxonomy and empirical study. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2012; 34(3): 417-35. PMid:21768651. https://doi.org/10.1109/TPAMI.2011.142

[13] Triguero I, Derrac J, García S, et al. A taxonomy and experimental study on prototype generation for nearest neighbor classification. IEEE Transactions on Systems, Man, and Cybernetics-Part C: Applications and Reviews. 2012; 42(1): 86-100. https://doi.org/10.1109/TSMCC.2010.2103939

[14] Fortran original by Leo Breiman and Adele Cutler, R port by Andy Liaw and Matthew Wiener. Breiman and Cutler's Random Forests for Classification and Regression. Random Forest citation info, GPL-3. 2015: 1-29.

[15] Ho TK. The Random Subspace Method for Constructing Decision Forests. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1998; 20(8): 832-44. https://doi.org/10.1109/34.709601

[16] Amit Y, Geman D. Shape quantization and recognition with randomized trees. Neural Computation. 1997; 9(7): 1545-88. https://doi.org/10.1162/neco.1997.9.7.1545