Send Orders of Reprints at reprints@benthamscience.net

62 *The Open Electrical & Electronic Engineering Journal,* 2013, *7,* (Supple 1: M7) 62-70

Open Access

An Improved Random Forest Algorithm for Class-Imbalanced Data Classification and its Application in PAD Risk Factors Analysis

Dengju Yao[1,2,*], Jing Yang[1] and Xiaojuan Zhan[3]

[1]*College of Computer Science and Technology, Harbin Engineering University, 150001 Harbin Heilongjiang, China*

[2]*Software College, Harbin University of Science and Technology, 150040 Harbin Heilongjiang, China*

[3]*Department of Computer Science and Technology, Heilongjiang Institute of Technology, 150050 Harbin Heilongjiang, China*

Abstract: The classification problem is one of the important research subjects in the field of machine learning. However, most machine learning algorithms train a classifier based on the assumption that the number of training examples of classes is almost equal. When a classifier was trained on imbalanced data, the performance of the classifier declined clearly. For resolving the class-imbalanced problem, an improved random forest algorithm was proposed based on sampling with replacement. We extracted multiple example subsets randomly with replacement from majority class, and the example number of extracted example subsets is as the same with minority class example dataset. Then, multiple new training datasets were constructed by combining the each exacted majority example subset and minority class dataset respectively, and multiple random forest classifiers were training on these training dataset. For a prediction example, the class was determined by majority voting of multiple random forest classifiers. The experimental results on five groups UCI datasets and a real clinical dataset show that the proposed method could deal with the class-imbalanced data problem and the improved random forest algorithm outperformed original random forest and other methods in literatures.

Keywords: Class-imbalanced problem, random forest, sampling with replacement, classification, PAD, disease risk factors.

1. INTRODUCTION

The classification problem is one of the important research subjects in the field of machine learning. Currently, there are many kinds of machine learning algorithms which perform perfectly in common datasets. However, most machine learning algorithms train a classifier based on the assumption that the number of training examples of classes is almost equal. When we trained a classifier on imbalanced data, the performance of the classifier will suffer from clearly declining, where the examples of minority class will be prone to be classified as majority class. This phenomenon of uneven distribution of data sample classes is known as class-imbalanced problems.

Class-imbalanced problems widely exist in many fields such as financial fraud detection [1], oil prospecting [2], Anti-spam [3], text classification [4], especially in biomedical and bioinformatics researches [5]. Traditional classification algorithms target to optimize classification accuracy without fully considering the class distribution of examples, and the trained classifier is mostly derived from the majority class. This will result in very poor prediction performance of the minority class. In many cases, the user is more interested in minority class. Thus, addressing and solving imbalanced data problem is very critical for improving classification performance [6]

Random forest [7] is an ensemble classifier that consists of many decision trees and outputs the class that is the majority of the classes of all the individual trees. The method combines bootstrap and the node randomly split technical to train multiple trees, and the classification result is decided by majority voting. In recent years, the random forest algorithm has become a popular classification technique and research hot in the field of machine learning, and is widely used in a variety of problems such as classification, prediction, variable importance, feature selection and outlier detection [8]. Like most other machine learning algorithms, random forest also suffered from imbalanced problems. In this paper, an improved random forest algorithm suitable for processing imbalanced data is proposed. The improved method adopted sampling with replacement to independently and randomly extract example subset from majority class dataset. The number of examples in the extracted subset is substantially equal to the number of examples in minority class subset. Then the new class-balanced training dataset was created by combining the extracted majority class example subset and the minority class example subset. Next, random forest classifiers were trained on the new dataset, and the final classification result was resulted through multiple classifier voting.

The paper is followed as below: in Section 2, the related research works of imbalanced data classification are presented. The improved random forest algorithm and the method using for evaluating algorithm is presented in detail

*Address correspondence to this author at the College of Computer Science and Technology, Harbin Engineering University, 150001 Harbin Heilongjiang, China; Tel: 13796671858; Fax: 86397100;
E-mail: ydkvictory@163.com

1874-1290/13

in Section 3. In Section 4, we show our empirical results and evaluate the results. Finally, we conclude in Section 5.

2. RELATED WORKS

Recent research on the imbalanced data problem has been focused on several major groups of techniques.

The popular method to solve imbalanced data problem is random re-sampling technique which balances the number of training examples among classes [6]. Common random re-sampling techniques include the random over sampling (ROS) and the random under sampling (RUS). The former balances the number of training examples among classes by copying examples in a minority class, while the latter randomly selected an examples subset from a majority class in order to achieve the same purpose. These random re-sampling techniques are easy to apply and improve the performance of classifiers by compensating for the imbalanced class distribution. However, studies have shown that the random over sampling generally produce unwanted effects such as over-fitting and time overhead, while the random under sampling result in information loss by using only a examples subset of majority class. To overcome these imbalanced data problems of random re-sampling, SMOTE (synthetic minority oversampling technique) [9] change the distribution of the data by interpolation method. Though SMOTE also increases the number of minority class samples, it adopts artificially generated examples (e.g., creating new examples for the minority class that is inferred from existing examples [9]) rather than randomly copied examples. Thus SMOTE can avoid the problem of over-fitting, but it may introduce noise.

Other methods for solving the imbalanced data problem include developing new classification algorithms or improving the traditional classification algorithms to adapt the characters of imbalanced data classification. A typical method is the ensemble of under-sampled classifiers with bagging. We draw almost the same numbers of majority and minority examples for the sampled subset data by the sampling with replacement. Multiple classifiers are trained on the multiple sampled balanced subsets and applied to prediction the classes of example in imbalanced test dataset, the prediction result will be determined finally by majority voting. Sangyoon Oh [6] *et al*. proposed an ensemble learning algorithm with active example selection (EAES) for imbalanced biomedical data classification. The experimental results on six real-world imbalanced biomedical datasets show that EAES outperforms both the random under sampling and the ensemble with under sampling methods. Tsymbal A *et al*. [10] proposed a new genetic search algorithm for feature selection using integrated learning thinking. Although this method can get better learning performance, but the cost is of time-consuming. Another method is cost-sensitive learning algorithm [11], which has been proved to be equivalent to re-sampling methods. Additionally, Castillo *et al*. [12] researched the imbalanced problem in text classification. Chen *et al*. [13] studied on imbalanced problem in Medication detection. Yoon *et al*. [14] introduced imbalanced problem in bioinformatics.

On the other hand, Random Forest (RF) has become popular in machine learning and pattern recognition filed, and has been used widely in researches on classification, prediction, variable importance, feature selection and anomaly detection, etc. As a classifier integration method, RF have the features of classifying fast and training simple, so it is suitable for feature selection according to variable importance. Random forests is favored especially in the field of biomedical and bioinformatics, because it can identify interactions between multiple predictor variables. The research of David *et al*. [15] shows that the random forest is good at identifying the relevant features from high-dimensional data with weak main effects and low genetic possibilities. Mohammed *et al*. [16] used the random forest algorithm for disease prediction, and the results showed that random forest outperformed support vector machines, bagging and boosting technology with respect to the classification performance. Ding *et al*. [17] used the random forest algorithm for somatic mutation detection in the tumor normal pair sequence high-dimensional datasets and got higher prediction performance. Natalia *et al*. [18] succeeded in predicting Alzheimer's disease risk factors by using the random forest on a genome-wide association study datasets. Carolin *et al*. [19] proposed a conditional variable importance measure which was proved to be more reliably for reflecting the impact of the predictor variables on the response variable than the original marginal method. A. Verikas *et al*. [20] taken a new test on variable importance of random forest, and the experiments show that the variable importance ranking depends on the number of selected variables in the process of node splitting in random forests training. Gray and Fan [21] used genetic algorithm(GA) as weaker classifier to design classification forest, and the results show that original RF outperformed random forest based on GA with respect of classification accuracy. Wang *et al*. [22] from Tsinghua University in China proposed a face key-point positioning algorithm based on random forest classifier. Deng *et al*. [23] proposed a co-training-style random forest method to be used for lung nodule detection CT image data analysis.

Although random forest algorithm has succeeded in a large number of application filed, its' consistency and applicability still open to question with respect to classification performance, as research results in literature [20], especially in class-imbalanced data classification. Meanwhile, ROS and RUS either results in over-fitting or produces information loss. Against the above problems, we proposed an improved random forest algorithm based on SMOTE. The experiment results on UCI dataset show that the proposed method performed well on imbalanced data.

3. ALGORITHM DESIGHN AND EVALUATION

3.1. Algorithm Design

RF is an ensemble classifier that consists of a set of decision trees weak classifier $\{h(x, \theta k), k=1,...\}$, where $\{\theta k\}$ are random vector with independent and identically distribution. Given independent variable X, each decision tree classifier vote for determining the class of example X. If the decision tree is regarded as an expert in the classification task, random forests is a classifier that integrated many experts together to implement certain classifying tasks. Random forests algorithm combines "bagging" [24] and the random selection of features for tree node splitting to construct a collec-

Input: imbalanced example dataset, marked as D

Output: classifier F(x), x is prediction example

① split up D into positive example subset and negative example subset, marked as P and N respectively. If the D has more than two classes, the class with least examples was thought as minority class or negative example, and all others as majority class or positive example.

② $ratio \leftarrow \left\lfloor \frac{|N|}{|P|} \right\rfloor$

③ extract *ratio* example subsets using randomly sampling with replacement (Bagging) from majority class example dataset, the example number of each subset is equal to the size of minority class example dataset, each subset marked as N_i

④ for i form 1 to ratio

⑤ construction training set $T_i \leftarrow P + N_i$

⑥ train random forest classifier C_i, $C_i(x)$ *represent* the prediction result for example x

⑦ end for

⑧ $F(x) = \operatorname{sgn} \sum_{i=1}^{ratio} C_i(x)$

Fig. (1). Improved random forest algorithm for imbalanced data.

tion of decision trees with controlled variation. Random forests can be constructed as follows [7]:

Step 1: For a given training dataset, extract a new sample set by N times repeated random sampling using bootstrap method. For example, from the data (x_1, y_1),...,(x_n, y_n) to build a sample (x_1^*, y_1^*)....(x_N^*, y_N^*) . Samples which are not being extracted consist of out-of-bag data (OOB).

Step 2: Build a decision tree or regression tree based on sample set resulted from step 1;

Step 3: Repeat step1-2, result in many trees, composing a forest.

Step 4: Let every tree in the forest to vote for x_i.

Step 5: Calculate the sum of votes for every class, the class with highest number of votes is the classification label for x_i.

Step 6: The percentage of incorrect classification is the classing error ratio of random forest.

In order to solve the imbalance problem where the number of positive examples are far greater than the number of negative examples in training dataset and in order to compensate loss of information in random under sampling at the same time, this paper adopted a new sampling method to improved original random forest. Inspired by literature [25, 26], we extracted multiple example subsets randomly with replacement from majority class, and the example number of extracted example subsets is as the same with minority class example dataset. Then, multiple new training datasets were constructed by combining the each exacted majority example subset and minority class dataset respectively, and multiple random forest classifiers were training on these training dataset. For a prediction example, the class was determined by majority voting of multiple random forest classifiers. The algorithm detail is described as Fig. (**1**).

3.2. Algorithm Evaluation

Accuracy, sensitivity and *specificity* are three commonly used measurements for evaluating classifier performance, which are computed based on the confusion matrix [29]. A confusion matrix is a matrix usually used to represent the relationships between real class attributes and that of predicted classes. In a two-class prediction problem, the upper left cell denotes the number of samples classified as true while they are true (TP), and lower right cell denotes the number of samples classified as false while they were false (TN). The other two cells represent the number of samples misclassified. Specifically, the lower left cell represents the number of samples classified as false while they were true (FN), and the upper right cell represents the number of samples classified as true while they actually were false (FP). When the confusion matrixes were obtained, the accuracy, sensitivity and specificity could be calculated using the following formulas respectively.

The *accuracy* of classifiers is the percentage of correctness of prediction among the test sets. It is defined in (1). The *sensitivity* is referred as the true positive rate, and the *specificity* as the true negative rate. Both sensitivity and specificity are used for measuring the factors that affect the performance, and are computed using (2) and (3), respectively.

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

$$sensitivity = \frac{TP}{TP + FN} \tag{2}$$

$$specificity = \frac{TN}{TN + FP} \tag{3}$$

The *sensitivity* is equivalent to recall in pattern recognition, and precision in pattern recognition is slightly different with *accuracy*. The precision is defined in (4).

$$precision = \frac{TP}{FP + TP} \tag{4}$$

In class-imbalanced problem, because the TN and FP is far greater than TP and FN [23], so

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \approx \frac{TN}{TN + FP} = sp \tag{5}$$

Therefore, accuracy is generally approximated to specificity. We usually adopt sensitivity and specificity to evaluate the classification performance without concerning accuracy in imbalanced data classification.

4. EXPERIMENT AND DISCUSSION

In this section, we evaluated the improved random forest algorithm on UCI datasets and used it for PAD risk factors analysis on real clinical dataset, respectively.

4.1. Test on UCI Datasets

In order to verify the effectiveness of the proposed method and strategies, the paper adopted 5 groups UCI [30]

Table 1. Summary of UCI Datasets

Dataset	F	S	Min/Max	T	R
cmc	10	1473	333/1140	2	3.42
haberman	4	306	81/225	2	2.78
ionosphere	35	351	126/225	b	1.79
letter	17	20000	789/19211	A	24.35
pima	9	768	268/500	1	1.87

datasets which have been used in the literature [25, 31] as experiments data. These 5 groups datasets include cmc, haberman, ionosphere, letter and pima. All attributes of these datasets are real numbers, and sample classes are imbalanced (minority class as positive examples and the rest as negative example). The characteristics of these 5 datasets are summarized in Table **1**, where F respect to the number of features in dataset, S respect to the size of dataset, Min and Max respect the examples number of minority class and majority class in dataset respectively, T respect to class label of minority class, and R respect to the ratio of majority class to minority class.

We run original random forest (originalRF) algorithm and our proposed improved random forest (ImprovedRF) on all five datasets respectively. The original RF adopted randomForest package in R software, and the improved RF algorithm is implemented using R language based on the original RF. The hardware environment for the experiments includes computer with Intel Core (TM) 2 Duo CPU E4600@2.40GHz and 3.75G memory, The software environment used the Microsoft Windows 7 operating system used and R software.

In order to minimize the bias associated with the random sampling of the training, we used a 5-fold cross-validation method. With cross-validation, some of the data is removed before training begins. Then when training is done, the data that was removed can be used to test the performance of the learned model on "new" data [27, 28]. In our method, we randomly divided the original dataset into 5 disjoint subsets (folds), with each fold containing approximately the same number of records. The sampling is stratified by the class labels to ensure that the subset class proportions are roughly the same as those in the whole dataset. For each training subset, a random forest classifier is constructed using the 4 of the 5 folds and tested on the fifth one to obtain a cross-validation estimate of its prediction accuracy. The 5 times prediction accuracy are then averaged to provide an estimate for the classier accuracy constructed from all the data. Finally, we run each algorithm 10 times, and the performance of the each classifier (precision and recall) is evaluated by the average of 10 times experimental results.

On 5 groups UCI dataset, We compared the method AdaBoost, UderSampl, Hsampl, AsymBoosT, BalanceCascade, LibID (once) and LibID (repeat), original RF and our proposed improved RF. The experimental results of the above first seven algorithms come from the literature [25, 31]. The performances of 9 classifiers are summarized in Table **2**.

As shown in Table **2**, only on letter dataset, our improved RF algorithm is slightly outperformed by AsymBoosT and BalanceCascade with respect to total performance, but on all other datasets, whether original RF or the improved RF significantly outperformed all other algorithms. Additionally, the improved RF algorithm outperformed original RF algorithm on all datasets expect letter. Random forest is an ensemble classifier and suitable for using for classification problem on weak classification dataset, but letter is a strong classification dataset [25]. AdaBoost algorithm has performed almost perfectly in letter dataset. However, maybe because the imbalanced ratio of classes in letter is too large, the re-sampling inevitably produces some information loss, and then reduced performance of the improved RF. It is for this reason that the original RF produces the classification result almost equivalent to AdaBoost. Overall, our proposed improved random forest algorithm has shown a larger advantage in resolving class-imbalanced problem, especially on weaker classification datasets. The experimental results also show that recall has been improved without reducing precision, and the recall is particularly important in some field such as biomedical and bioinformatics.

4.2. Analysis of Risk Factors for PAD

Peripheral arterial disease (PAD) is a common manifestation of systemic atherosclerosis [32]. It has become the most important issue to reduce the incidence rate of diabetes and prevent patients from diabetic complications [33]. The risk factors leading to diabetic complications are complicated, and accurately identifying the important risk factors of the complications from vast amounts of medical data is the key to reducing the incidence rate of diabetes complications. The complex interactions between many factors lead to difficulty for this research. Previous disease risk factors studies often adopt statistical approaches. However, the size of the dataset is so huge and the number of variables is so much that traditional statistical approaches can not accurately and effectively explain the results. As previously mentioned, random forest is a versatile classification algorithm suited for the analysis of large datasets, which is popular because it has high-prediction accuracy and provide information on importance of variables for classification [34]. Here, we employ random forest algorithm to research risk factors for Peripheral arterial disease and to construct PAD prediction model. For this purpose, we have collected clinical dataset from the First Affiliated Hospital of Harbin Medical University in China between 2006 and 2010, which includes clinical records of 2765 diabetic patients and 73 record items for each patient. Table **3** provides an overview of the features in the

Table 2. Performance of 9 Classifiers on 5 UCI Datasets

Data(|p|/|N|)	Classifier	Precision	Recall
cmc (333/1140)	AdaBoost	0.40	0.39
	UderSampl	0.33	0.63
	Hsampl	0.37	0.48
	AsymBoosT	0.39	0.42
	BalanceCascade	0.35	0.59
	LibID (once)	0.48	0.74
	LibID (repeat)	0.50	0.67
	originalRF	0.82	0.93
	ImprovedRF	0.99	0.78
haberman (81/225)	AdaBoost	0.35	0.36
	UderSampl	0.36	0.60
	Hsampl	0.36	0.47
	AsymBoosT	0.34	0.39
	BalanceCascade	0.36	0.57
	LibID (once)	0.54	0.80
	LibID (repeat)	0.59	0.84
	originalRF	0.76	0.90
	ImprovedRF	0.96	0.81
ionosphere (126/225)	AdaBoost	0.95	0.88
	UderSampl	0.92	0.89
	Hsampl	0.94	0.86
	AsymBoosT	0.95	0.88
	BalanceCascade	0.93	0.89
	LibID (once)	0.94	0.89
	LibID (repeat)	0.94	0.91
	originalRF	0.94	0.87
	ImprovedRF	0.95	1
pima (268/500)	AdaBoost	0.63	0.60
	UderSampl	0.58	0.73
	Hsampl	0.62	0.65
	AsymBoosT	0.63	0.61
	BalanceCascade	0.60	0.71
	LibID (once)	0.78	0.76
	LibID (repeat)	0.77	0.81
	originalRF	0.80	0.85
	ImprovedRF	1	0.85

Table 2. contd….

letter (789/19211)	AdaBoost	0.99	0.98
	UderSampl	0.83	0.99
	Hsampl	0.92	0.99
	AsymBoosT	0.99	0.98
	BalanceCascade	0.96	0.99
	LibID (once)	0.88	0.99
	LibID (repeat)	0.85	0.98
	originalRF	0.999	0.949
	ImprovedRF	0.87	1

Table 3. Overview of Variable Group (and Number of Variables) that Appear in the Data Mining Dataset

Group	Variables
Basic information (20)	Age, Sex, Marital status, Duration of diabetes, Hypertension history, Diabetic family history, Hypertension family history, Other medical history, Weight, Height, Waistline, Hip circumference, Body mass index (BMI), History of smoking, History of drinking, Education, Job, Number of children, Systolic blood pressure(SBP), Diastolic blood pressure(DBP)
Blood lipid (9)	Total cholesterol (TC), Triglyceride (TG), Very low density lipoprotein (VLDL), Low density lipoprotein (LDL), High density lipoprotein (HDL), Apolipoprotein A (Apo-a), Apolipoprotein B (Apo-b), Cholinesterase (CHE), apoA-I/apoB
Blood glucose (6)	Fasting plasma glucose (FPG), 30-minuets postprandial blood glucose (BG30), BG60, BG120, BG180, hemoglobin Alc (HbAlc)
Insulin releasing test (11)	Fasting C-peptide (FCP), 30- minutes postprandial c- peptide (CP30), CP60, CP120, CP180, Fasting insulin(FINS), 30-minutes postprandial insulin (INS30), INS60, INS120, INS180, Homeostasis model assessment insulin resistance (HOMA-IR)
Liver function (8)	Aspartate aminotransferase(AST), Alanine aminotransferase(ALT), γ-glutamyltransferase (GGT), Albumin (ALB),Total bilirubin (TBIL),Conjugated bilirubin(DBIL),Unconjugated Bilirubin (IBIL), Alkaline Phosphatase (ALP), Presence ofr non-alcoholic fatty liver
Kedney function (4)	Blood urea nitrogen (BUN), Serum creatinine (SCR), Uric acid (UA), Urine microdosis protein (UMP)
Blood coagulation item (15)	Fibrinogen (Fib), Prothrombin time (PT), serum D-dimer (DD), Thrombin time (TT), Activated partial thromboplastin time (APTT), International normalized ratio (INR), hemocysteine (HCY), Fibrinogen (Fib), Prothrombin time (PT)

Table 4. The Summary of Variable Characteristics in Diabetes Mellitus Clinical Dataset

Name	Category	Range/Mean	Name	Category	Range/Mean	Name	Category	Range/Mean
Sex	nominal	Female, Male	BG30	numeric	11.61	TG	numeric	2.398
Age	numeric	52.56	BG60	numeric	14.87	HDL.C	numeric	1.187
HOD	numeric	6.579	BG120	numeric	14.48	LDL.C	numeric	3.455
HBPH	nominal	N,L,M,S	BG180	numeric	11.03	VLDL.C	numeric	0.4626
FH	nominal	H,D,B,N	FINS	numeric	11.76	Apo.A	numeric	1.283
WH	nominal	N,L,M,S	INS30	numeric	27.64	Apo.B	numeric	0.9428
CH	nominal	N,L,M,S	INS60	numeric	42.94	BUN	numeric	6.081
Height	numeric	168.6	INS120	numeric	45.09	Cr	numeric	77.66
Weight	numeric	72.94	INS180	numeric	31.68	UA	numeric	308.98
Waist	numeric	89.3	FCP	numeric	1.610	ALT	numeric	31.42
Hip	numeric	94.16	CP30	numeric	2.885	AST	numeric	24.82
SP	numeric	138.0	CP60	numeric	4.175	GGT	numeric	44.29
DP	numeric	82.18	CP120	numeric	5.375	Fib	numeric	3.569

Table 4. contd….

Name	Category	Range/Mean	Name	Category	Range/Mean	Name	Category	Range/Mean
HR	numeric	79.45	CP180	numeric	4.555	HOMA	numeric	3.673
BMI	numeric	25.55	HbAlc	numeric	8.746	CAD	nominal	No, Yes
FPG	numeric	6.942	TC	numeric	5.35			

Note: N, L, M, S, H, D and B represents No, Low degree, Moderate degree, High degree, Severe degree, Hypertension, Diabetes, and Both, respectively.

Table 5. The Results of Original RF and Improved RF on Diabetes Mellitus Clinical Dataset

Algorithm	Precision	Recall
originalRF	0.7422	0.3506
ImprovedRF	0.7494	0.7688

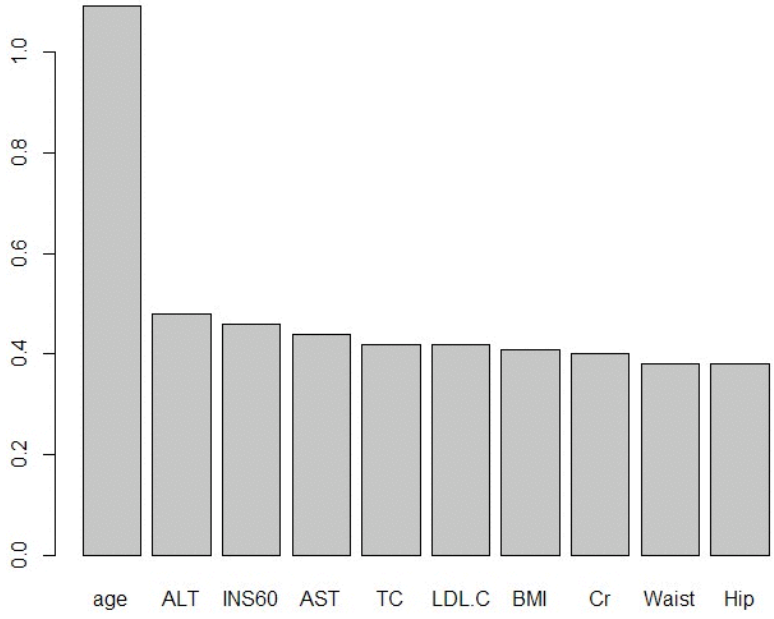

Fig. (2). The importance of top-10 features.

dataset. The data only include patients with type II diabetes excluding patients with type I, gestational and secondary diabetes mellitus.

As a real clinical dataset, there are some noisy, irrelevant and redundant information. Before running our ImprovedRF algorithm on this dataset, data cleaning were used to preprocess the data. We filled vacancy values, discarded outliers, smoothed noise to improve data reliability. As a result, clearly irrelevant attributes were removed from the dataset by the advices of diabetes experts. Finally, the dataset remained 1910 records of type 2 patients, including 1266 males and 644 females, and 46 features which may be relevant to the targeted variable of PAD with type 2 diabetes mellitus. The features include 5 nominal features, 41 numeric features and a class feature which describe the patient is or not with Peripheral arterial disease. The detail information of these features is summarized in the Table **4**. In the dataset, there are 1368 positive samples represented the patients with Peripheral arterial disease and 542 negative samples represented the people without Peripheral arterial disease. This is apparent a class-imbalanced dataset and the imbalance ration is 2.524:1. Next, we used the improved random forest proposed in this paper for exploring risk factors for Peripheral arterial disease.

On the diabetes mellitus clinical dataset after preprocessing, we run original random forest and our improved random forest respectively. The results are presented in Table **5**.

As shown in Table **5**, the ImprovedRF algorithm could improve recall significantly while the precision remained almost same with originalRF. The result showed that the proposed method could effectively deal with class-imbalanced problem and could improve prediction accuracy of minority class example.

Additionally, random forest algorithm can give importance score of variables according to their ability to predict target variable. The score value is larger, the variable is more important for target variable. This variable importance score can help doctors to explore disease risk factors when the algorithm is used on biomedical or bioinformatics data. Fig. (**2**) represented the importance of top-10 features in diabetes mellitus clinical dataset for predicting PAD.

Fig. (**2**) shows that age is the most important factor in PAD prevalence in patients. Considering variable importance score from the random forest, age has a score (1.09) which are more twice larger than the second variable ALT with importance score (0.48). This is consistent with previous studies on PAD risk factors for PAD. ALT and INS60 levels were the second and the third major factors for PAD respectively. Because some individuals in this study were diabetic patients and have a high incidence of non-alcoholic fatty liver (NAFLD), we speculate that elevated ALT levels are an important PAD risk factor in diabetic individuals. Recent studies have shown that alanine aminotransferase (ALT) is a sign of liver damage and is related to endothelial dysfunction and angiosclerosis. This result is consistent with our findings. We have also made the important discovery that postprandial insulin levels (INS30) are also important risk factors for PAD. These results have been confirmed in both *in vivo* and *in vitro* studies. Insulin is a potent growth factor that augments collagen synthesis and stimulates arterial smooth muscle cell proliferation which is an atherogenic process. Overall, the results in this paper are highly consistent with previous studies, it presents that the proposed improved random forest algorithm is reasonable and suited for disease risk factors analysis, especially for solving class-imbalanced problem.

5. CONCLUSION

In order to resolve sample class-imbalanced problems in classification, an improved random forest was proposed based on SMOTE. The proposed method was implemented on R software and tested on five groups UCI datasets. The experimental results showed that our proposed algorithm could improve recall while keeping precision. The proposed method was also used on real diabetes mellitus clinical dataset and used for risk factors analysis for peripheral arterial disease, and the results were highly consistent with previous research. However, classification problem on class-imbalanced data is an important research subjects in the field of machine learning. The method proposed in this paper has only been tested on some classic datasets, experiments on more datasets especially real datasets is the future research directions.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflicts of interest.

ACKNOWLEDGEMENT

The authors are grateful to the support of the National Natural Science Foundation of China under Grant No. 61073043, 61073041, and 61100008, the Natural Science Foundation of Heilongjiang Province under grant number No. F200901 and F201023, Harbin Special Funds for Technological Innovation Research under grant number 2010RFXXG002, 2011RFXXG015, and Fundamental Research Funds for the Central Universities of China under grant number HEUCF100602. Thanks to Doctor Yang for helpful comments, suggestion and criticisms.

REFERENCES

[1] J. Stolfo, W. Fan, W. Lee, A. Prodromidis, and K. Chan, "Cost-based modeling for fraud and intrusion detection: results from the jam project", In: *DARPA Information Survivability Conference and Exposition*, 2000, pp. 130-144.

[2] M. Kubat, R. Holte, and S. Matwin, "Machine learning for the detection of oil spills in satellite radar images", *Mach. Learn*., vol. 30, pp. 195-215, Feb1998.

[3] T. Fawcett. "In vivo spam filtering: A challenge problem for data mining", *ACM SIGKDD Explor.*, vol. 5, pp. 140-148, Dec 2003.

[4] Y. Xu, J. T. Lee, and B. Wang, "A study of feature selection for text categorization on imbalanced data", *J. Comput. Res. Dev.*, vol. 43, pp. 58-62, 2006.

[5] C. L. Wang, C. Ding, and R. F. Meraz, "PSol: A positive sample only learning algorithm for finding non-coding RNA genes", *Bioinformatics*, vol. 22, pp. 2590-2596, Aug 2006.

[6] O. Sangyoon, M. S. Lee, and B. T. Zhang, "Ensemble learning with active example selection for imbalanced biomedical data classification", *IEEE ACM Trans. Comput. BI*, vol. 8, pp. 316-325, Mar 2011.

[7] L.Breiman, "Random forests", *Mach. Learn.*, vol. 45, pp. 5-32, 2001.

[8] C. Strobl, A. L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis, "Conditional variable importance for random forests", *BMC Bioinforma.*, vol. 9, pp. 307, Jul 2008.

[9] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique", *J. Artif. Intell. Res.*, vol. 16, pp. 321-357, 2002.

[10] A. Tsymbal, M. Pechenizkiy, and P. Cunningham, "Sequential genetic search for ensemble feature selection", In: *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence*, 2005, pp. 877-882.

[11] A. Zadrozny, J. Langford, and N. Abe, "Cost- sensitive learning by cost-proportionate example weighting", In: *Proc of the 3rd Internal Conference on Data Mining*, 2003, pp. 435- 442.

[12] M. D. Castillo, and J. I. Serrano, "A multi-strategy approach for digital text categorization from imbalanced documents", *ACM S Igkdd Explor. Newslett.*, 2004, vol. 6, pp. 70-79.

[13] J. X. Chen, T. H. Cheng and L. F. Chan, "An application of classification analysis for skewed class distribution in therapeutic drug monitoring -the case of vancomycin", In: *Proceedings of the IDEAS Workshop on Medical Information Systems*: *The Digital Hospital*, 2004, pp. 35-39.

[14] K. Yoon and S. Kwek, "An unsupervised learning approach to resolving the data imbalanced issue in supervised learning problems in functional genomics", *Neural Comput. Appl.*, vol. 16, pp. 295-306, 2007.

[15] M. R. David, A. M. Alison, A. M. Brett, McKinney, E. James, J. Crowe and H. M. Jason, "Feature Selection using a Random Forests Classifier for the Integrated Analysis of Multiple Data Types", In: *2006 IEEE Symposium on Computational Intelligence and Bioinformatics and Computational Biology*, 2006, pp. 1-8.

[16] K. Mohammed, C. Sounak and P. Mihail, "Predicting disease risks from highly imbalanced data using random forest", *BMC Med. Informa. Decis.*, pp. 11-20, Jul 2011.

[17] J. R. Ding, A. Bashashati, A. Roth, K. Tse, T. Zeng, M. Hirst, M. A. Marra, A. Condon, S. Aparicio, and S. P. Shah, "Feature-based classifiers for somatic mutation detection in tumor normal paired sequencing data", *Bioinformatics*, vol. 28, no. 2, pp. 167-175, Jan, 2012

[18] N. Briones, and V. Dinu, "Data mining of high density genomic variant data for prediction of Alzheimer's disease risk", *BMC Med. Genet.*, pp. 1-12, Jan 2012.

[19] C. Strobl, A. L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis, "Conditional variable importance for random forests", *BMC Bioinforma.*, vol. 9, pp. 307-319, Jul 2008

[20] A. Verikas, A. Gelzinis, and M.Bacauskiene, "Mining data with random forests: A survey and results of new tests", *Pattern Recognit.*, vol. 44, no. 2, pp. 330-349, Feb 2011

[21] J. B. Gray, and G. Fan, "Classification tree analysis using TARGET", *Comput. Stat. Data Anal.*, vol. 52, pp. 1362-1372, Jan 2008.

[22] L. T. Wang, X. Q. Dingm, and C. Fang, "Accurate localizaion of facial feature points based on random forest classifier", *J. Tsinghua Uni.*, vol. 49, no. 4, pp. 543-546, Mar 2009.

[23] C. Deng, and M. Z. Guo, "A new co-training-style random forest for computer aided diagnosis", *J. Intell. Inf. Syst.*, vol 36, pp. 253-281, Nov 2011

[24] L.Breiman, "Bagging predictors", *Mach. Learn*., vol. 24, no. 2, pp. 123-140, Aug 1996

[25] Q. Zou, M. Z. Guo, Y. Liu, and J. Wang Jun, "A Classification Method for Class-Imbalanced Data and Its Application on Bioinformatics", *J. Comput. Res. Dev.*, vol. 47, no. 8, pp. 1407-1414, Aug 2010.

[26] X. Li, L .X. Wang, and S. Y. Jiang, "Ensemble learning based feature selection for imbalanced problems", *J. Shandong Uni.* (*Eng. Sci*)., vol. 41, no. 3, pp. 7-22, Mar 2011.

[27] N. Meinshausen, "Quantile Regression Forests", *J. Mach. Learn., Res.*, vol. 7, pp. 983-999, Dec 2006.

[28] D. Delen, G. Walker, and A. Kadam, "Predicting breast cancer survivability: a comparison of three data mining methods", *Artif. Intell. Med.*, vol. 34, pp. 113-127, Jul 2004.

[29] J. H. Warner, Q. W. Liang, M. Sarkar, P. E. Mendes, and H. J. Roethig, "Adaptive regression modeling of biomarkers of potential harm in a population of U.S. adult cigarette smokers and nonsmokers", *BMC Med. Res. Methodol.*, pp. 1-10, Mar 2010.

[30] Available at: http://archive.ics.uci.edu/ml/.

[31] X. Y. Liu, J. X. Wu, and Z. H. Zhou, "A cascade-based classification method for class-imbalanced data", *J. Nanjing Uni. (Nat. Sci.)*, vol. 42, no. 2, pp. 148-155, Feb 2006.

[32] W. R. Hiatt, "Medical treatment of peripheral arterial disease and claudication", *N. Engl. J. Med.,* vol. 344, no. 21, pp.1608-1621, May 2001.

[33] D. Mukherjee, "Peripheral and cerebrovascular atherosclerotic disease in diabetes mellitus", *Best Pract. Res. Clin. Endocrinol. Metab.*, vol. 23, no. 3, pp. 335-345, Jun 2009.

[34] W. G. Touw, J. R. Bayjanov, L. Overmars, L. Backus, J. Beokhorst, and S. A. Hijum, "Data mining in the Life Science with Random Forest: a walk in the park or lost in the jungle", *Brief Bioinforma.*, pp. 1-12, Jun 2012.

Received: February 224, 2012 Revised: April 29, 2012 Accepted: December 12, 2012