Journal of Computers Vol. 30 No. 2, 2019, pp. 213-223
doi:10.3966/199115992019043002020

Improved Cost-sensitive Random Forest for Imbalanced Classification

Jie-Ming Yang[1], Cong Gao[1*], Zhao-Yang Qu[1], Zhong-Feng Kan[2], Ye Gao[2], Cheng Chang[2]

[1] School of Information Engineering, Northeast Electric Power University, Jilin 132012, China
yjmlzy@sina.com; gaoconglhn@163.com ; 1103878327@qq.com

[2] Jilin Electric Power Supply Company, Jilin 132000, China
kanzhongfeng@163.com 303241949@qq.com 519071152@qq.com

Received 25 July 2017; Revised 4 March 2018; Accepted 31 March 2018

Abstract. Class imbalance limits the performance of most traditional algorithms. Cost-sensitive algorithm has been introduced to handle imbalanced datasets. The cost sensitive random forest can deal with imbalanced data better. However, the cost function of cost-sensitive algorithm does not consider the actual distribution of the sample set and feature weight. And for the final prediction of random forest, it adopts equal vote but the base classifiers in random forests are not of equal accuracy. This paper proposes an improved cost-sensitive random forest algorithm called ICSRF, which constructs a cost function based on the actual distribution of imbalanced data set and introduces the weight distance, then takes weighted voting according to the performance of the base classifier that can improve the classification accuracy. The experiment results show that the ICSRF algorithm has higher accuracy rate and can effectively improve the classification performance of a few classes.

Keywords: cost-sensitive, imbalanced data, random forest, weight distance

1 Introduction

Imbalanced classification is an important topic in machine learning. Many real domains including telecommunication, medical diagnosis and network intrusion present that the number of instances of majority class outnumbers that of minority class [1]. While the traditional classification algorithms based on the category equilibrium hypothesis seek high classification accuracies, resulting in low predictive accuracy of the minority class. And the correct identification of the minority class is usually more important in imbalanced classification problems. For example, in medical diagnosis problems, the cost of misdiagnosing cancer patients as having no disease is far greater than the price that would be misdiagnosed as a cancer patient [2]. Therefore, the traditional classification algorithms are not suitable for imbalanced tasks. It's urgent for increasing the recognition rate of the minority class in data mining areas [3].

To handle the challenges imposed by imbalanced data sets, many solutions have been proposed, such as balancing the class distribution by resizing the IR, taking cost-sensitive learning into consideration in the traditional classification algorithms. The first approach is named resampling, which contains over-sampling and under-sampling [4]. Over-sampling replicates the examples of the minority class to balance the class distribution. But no new information is added and it extends the training time of classification model [5]. Also, it may lead to the over-fitting because of instances random replication. Under-sampling removes instances form the majority class to adjust the class proportion and may lead to a loss of important information. Therefore, there is a limit to address imbalance classification in terms of adjusting data categories. In the algorithm level approaches, researchers combine the Adaboost algorithm with cost-sensitive algorithm [6], which incorporates cost information into the weight update of formula.

* Corresponding Author

While the base classifier of Adaboost algorithm is created in the whole feature space, causing longer training time. Researchers combine random forest with cost-sensitive algorithm [7], which incorporates misclassification costs into the attributes splitting. But the actual distribution has not yet been considered in the cost function of traditional cost-sensitive algorithm based on random forest. In addition, it treats unimportant features as equally as the important features in the computational procedure of the cost function. It adopts Euclidean distance as the measure of the difference between two instances. And the importance of all the features in the feature space is different. Thus merely calculating Euclidean distance is unfair to important features when we construct cost function, because it could cause inaccurate cost of construction and it could hardly guarantee the performance of cost sensitive learning, and eventually it reduces the performance of classifier. Since each classification tree of random forest is built based on a different data subset and a subset of variables are randomly chosen at each split. It does not suffer from over-fitting. But it results in performance differences of the base classifiers in dealing with imbalanced data. Furthermore, the base classifiers of random forest may contain noises when the datasets are collected from real life. If adopting the primordial equal voting methods, it will have a wrong judgment and reduce the overall performance of the random forest that contains a large number of noise trees.

In this paper, we propose an improved cost-sensitive random forest algorithm which is built on the basis of random forest. Specifically, it structures cost function on the basis of the actual distribution of dataset. And it adopts weighted distance rather than Euclidean distance as the measure of the difference between two instances. In addition, our proposed algorithm uses weighted voting method. When the base classifiers in random forests are not of equal ability of classification, we assign more weights to the more competent classifiers for the final prediction. The AUC value of each decision tree determines the weight of the voting phase, and increases the overall performance of the classifier.

2 Related Work

2.1 Cost-sensitive Learning

Owing to the unbalanced distribution of categories, classification algorithms usually result in poor classification of a few classes due to over-training of most classes. Cost sensitive learning is an important method to solve imbalanced classification problem [8]. A higher misclassification cost is assigned for minority class in order to build a classifier that has the lowest cost. And it seeks the minimum total cost instead of the minimum error rate [9]. It can effectively improve the classification performance of minority class.

Cost-sensitive approaches are usually based on a cost matrix [10]. We define c_0 as the minority class, c_1 as the majority class, $F(i, j)$ as the cost of misclassifying an instance of the i category as an instance of the j category. The cost matrix is presented in Table 1.

Table 1. Cost matrix used by the base classifiers

class	c_0	c_1
c_0	$F(c_0,c_0)$…	$F(c_0,c_1)$
c_1	$F(c_1,c_0)$…	$F(c_1,c_1)$

Then constructing risk function according to Bayes theorem as shown in (1).

$$R(c_i|x) = \sum P(c_j|x)F(c_j, c_i)\,. \tag{1}$$

Where $P(c_j|x)$ is the posterior probability that it classified instance x as the category of c_j.

Instance x is classified as minimum of risk function. It has the following form.

$$c = \arg\min_{1\le i\le l}\left\{R(c_i|x)\right\}. \tag{2}$$

2.2 Decision Tree

The decision tree is an inductive learning algorithm on the base of examples and provides better fault tolerance [11]. Specifically, the decision tree algorithm firstly trains the samples and grows a decision tree. During the decision tree training process, instances at each interior node are split into subsets based on a feature. In the testing phase, the output classes of the test samples are determined at terminal nodes.

There are many kinds of decision tree algorithms, such as ID3, C4.5 [12]. The algorithm selects a splitting attribute according to certain selection principles. The C4.5 algorithm uses the information gain rate as the selection criteria. In the process of constructing decision tree, the algorithm splits node starting at the root node until all the leaf nodes are marked. We define S is the data set. Following an equation describe entropy.

$$E(S)=\sum_{i=1}^{n} p_i \log(p_i)\,. \tag{3}$$

Where P_i represents the proportion of the number of samples with the class label i.

Assume that we use A as a splitting property and A has X_A different attribute values. $E(S, A)$ is the expected entropy resulting from selection of attribute A. It can be calculated as Eq. (4)

$$E(S,A)=\sum_{v\in X_A}\frac{|S_v|}{|S|}E(S_v) \tag{4}$$

Where is the subset of dataset with the value, v in A, $E(S_v)$ is the entropy of the branch node of dataset S_v.

Information Gain (IG) is a widely used metric to estimate the features. The information gain refers to the difference of information entropy after the occurrence of a feature. The larger Information Gain of a feature is, the more important the feature is for categorization. Information gain is defined as (5)

$$Gain(S,A)=E(S)-E(S,A)\,. \tag{5}$$

Where S represents the dataset, A is the attribute.

Information gain rate is defined as (6)

$$Gain\_ratio(S,A)=\frac{Gain(S,A)}{split\_\inf o_A(D)}\,. \tag{6}$$

Where $split\_info(A)$ is the entropy of the training set D about A. And is defined in the following way

$$split\_\inf o_A(D)=-\sum_{j=1}^{v}\frac{|D_j|}{|D|}\log_2\frac{|D_j|}{D} \tag{7}$$

2.3 Cost-sensitive Algorithm Based on Random Forest

Random forest is an ensemble classification which constructs a number of decision trees. Each base classifier is built based on a different set of bootstrap instances of the dataset and randomly chooses a subset of attributes at each split which guarantee the diversity of the base classifier. Bagging and the random subspace method play a vital role in the problems of over-fitting and accuracy [13]. Thus, random forest combines the results of numerous decision trees. The category of the instance is determined by the way of the minority is subject to the majority, which is defined as (8).

$$H(x)=\arg\max_{y}\sum_{k} I(h_k(x)=y)\,. \tag{8}$$

Where $h_k(x)$ is decision tree model, y is the classification results of decision tree, $I(.)$ is indicator function.

Random forest introduces cost-sensitive, misclassification cost is generally introduced in attribute splitting of decision trees. And split rule is not the traditional Gini index or information gain, but the reduction of misclassification cost. Then, selecting attribute that has the most rapid decline of misclassification costs. The reduction of misclassification cost is defined as the difference between selection attribute to split and not split, which is defined as follows.

$$\mathrm{Re}c = Mc - \sum_{i=0}^{n} Mc(A_i)\,. \tag{9}$$

Where *Rec* is the reduction of misclassification cost, *Mc* is the misclassification cost before the attribute splits, $\sum_{i=0}^{n} Mc(A_i)$ is the misclassification cost after selecting the attribute.

3 Improved Cost-sensitive Algorithm Based on Random Forest

Random forest does not suffer from over-fitting and introducing cost-sensitive algorithm that can deal with imbalanced data set. But the construction of cost function is not accurate, and it can not achieve the purpose of dealing with imbalanced data. The construction of the traditional cost function does not consider the actual distribution of the data set. And it adopts Euclidean distance as the measure of the difference between two instances, Euclidean distances is unfair to important features when we construct cost function. Therefore, the overall performance of the classifier is poor. Random forest utilizes both bootstrap bagging and a random subspace method which result in the performance differences when dealing with imbalanced data. While, the traditional random forest adopts equal voting methods, which affects the overall performance of the classifier. In this paper, we propose an improved cost-sensitive random algorithm forest (ICSRF).

The algorithm restructures cost function that considers the actual distribution and feature weight of the data set. The structure procedure of cost function will be detailed in Section 3.1. In the final prediction, considering a class imbalanced problem, our approach introduces weighted voting using the value of Area Under roc Curve(AUC). When the base classifier in the ensemble classification is not of equal accuracy. It is reasonable to assign more weights to the more competent classifiers. Outputs of ICSRF can be represented as follows:

$$H_c(x) = \arg\max_{y} \sum_{k} \alpha_k I(h_k(x) = y)\,. \tag{20}$$

where α_k represents the weights of the base classifier.

3.1 Construction of Cost Function

Our proposed algorithm constructs cost function according to the actual distribution of dataset. The weighted distance is introduced into the computational process of cost function. The details of the steps can be stated as follows:

Step 1. Computing the data centers between the majority, minority class and the entire dataset, respectively.

Computing the arithmetic average of each feature column. The data set is represented as the following matrix. *c* represents the class. Each row represents an instance and each column represents the feature of the data.

$$\begin{bmatrix} x_{11} & \cdots & x_{1m} & c \\ \vdots & \ddots & \vdots & c \\ x_{n1} & \cdots & x_{nm} & c \end{bmatrix}$$

The center of the majority class is calculated as follows:

$$A_1 = \frac{1}{n}\sum_{i=1}^{n} x_{i1}\,. \tag{31}$$

$$A_2 = \frac{1}{n}\sum_{i=1}^{n} x_{i2}\,. \tag{42}$$

……

$$A_m = \frac{1}{n}\sum_{i=1}^{n} x_{im} . \quad \textbf{(53)}$$

Getting the center of the majority class $(A_1, A_2, A_3, ...A_m)$. The center of the minority class is calculated in the same way.

Step 2. Calculating the weight distance from the majority class, the minority class center to the center of the whole dataset.

ICSRF uses the information gain to measure the importance of each feature in different categories. The information gain is defined as Eq. (14). Then, all features produce weight vector in the majority class $w(w_1, w_2, ..., w_m)$ and weight vector in the minority class $w'(w_{1'}, w_{2'}, ..., w_{m'})$. Add the respective weight vectors when calculate the weight distance from the majority class, the minority class center to the center of the whole dataset. The weight distance is defined as Eq. (15)

$$IG(x_k, c_i) = \sum_{c\in\{c_i,\overline{c_i}\}} \sum_{x\in\{x_k,\overline{x_k}\}} P(x,c)\log\frac{P(x,c)}{P(x)P(c)} . \quad \textbf{(64)}$$

Where P_c is probability of the datasets in category c, A_{ij} is probability of the datasets containing the term. $P(x,c)$ is probability of the datasets in category c that contain the term x.

$$d_i = \sqrt{\sum_{j=1}^{m} w_i (x_{ij} - \overline{x})^2} . \quad \textbf{(75)}$$

Where w_i is the weight of the feature in the minority class. A_{ij} is the center of the majority class. When calculating the weighted distance between the minority class and the center of the dataset, we adopt weight vector in the minority class and the center of the minority class. $\overline{A}$ is the center of the entire dataset.

Step 3. Definition γ coefficient

For the imbalanced data N, which contains the majority class c_1, the minority class c_0. And c_1 has N_1 instances, c_0 has N_0 instances. γ coefficient is defined as follows:

$$\gamma_i = \frac{\sum_{j=1}^{2} N_j}{N_i} . \quad \textbf{(86)}$$

Step 4. Construction of cost function

$$F(c_i, c_j) = \begin{cases} \gamma_i * \dfrac{d_i'}{d_j''}, d_i' < d_j''; \\ \gamma_i * \dfrac{d_i'}{d_j''}, d_j'' < d_i'; \\ 0, i = j; \\ 1, d_i' = d_j'' \end{cases} . \quad \textbf{(97)}$$

where d_i' represents the weight distance between the class c_i center and the center of the whole dataset. d_j'' represents the weight distance between the class c_j center and the center of the whole dataset.

3.2 Algorithm Description

Input: T: a training set

Output: The cost of the majority class and minority class. The AUC weight of each decision tree. The performance of the random forest.

Phase I calculating cost

Step1: calculating the misclassification cost. (as shown in the Eq. (11) to Eq. (17)).

Phase II cost-sensitive random forest

Step2: k of training subsets are obtained based on a different set of bootstrap samples of the data.

Step3: for each training subset.

a: randomly choosing a subset of features.

b: calculating the reduction of misclassification cost in the training subset (as shown in the Eq. (9)).

c: selecting attribute that has the maximal to the node split. Generating decision trees without pruning

Step4: OOB samples are used to estimate performance of the base classifier and get the AUC of each decision tree.

Step5: assigning the AUC weight to each decision tree.

Step6: using random forest with weighted voting for the test instances (as shown in the Eq. (8)).

4 Experimental Section

4.1 Datasets and Performance Measures

In order to evaluate the performance of ICSRF algorithm, we performed experiments using eight class datasets. Because the data sets have more than two classes, we choose the class with fewer data as the minority class and the rest as the majority class [14]. Table 2 shows the relevant information for each data set.

Table 2. The information of six data sets

dataset	example	feature	Minority class	Majority class	IR
breast-cancer	286	10	85	201	29.7
glass	214	10	70	144	32.7
balance-scale	625	5	49	576	7.8
heart-h	294	14	106	188	36.1
waveform	5000	21	1650	3350	33
diabetes	768	9	268	500	34.9

The traditional classification algorithm frequently uses the classification accuracy as the index to measure the validity of the algorithm. While people have more interest in minority class in the imbalanced data. The classification accuracy may results in misleading conclusions because of the favoritism to majority class. In order to evaluate the performance of the classification algorithm more meaningfully, a representation of classification performance can be formulated by a confusion matrix illustrated in Table 3. In this paper, we use the performance measures F-measure, AUC Recall and TPrate to measure the performance of our algorithm.

Table 3. Confusion matrix for performance evaluation

	Hypothesis positive	Hypothesis negative
Actual positive	TP	FP
Actual negative	FN	TN

4.2 Experimental Results

A comparison of our method with decision tree algorithm, random forest and the traditional cost sensitive random forest are made based on six imbalanced datasets. In this experiment, the misclassification cost of the traditional cost sensitive random forest is determined by the IR of the data set. For example, in the data set of breast-cancer, The ratio of majority class to minority class is 2.4, the misclassification cost of the minority class is 2.4 and the majority class is 1. The experimental results are shown as Fig. 1, Fig. 2, Fig. 3 and Fig. 4.

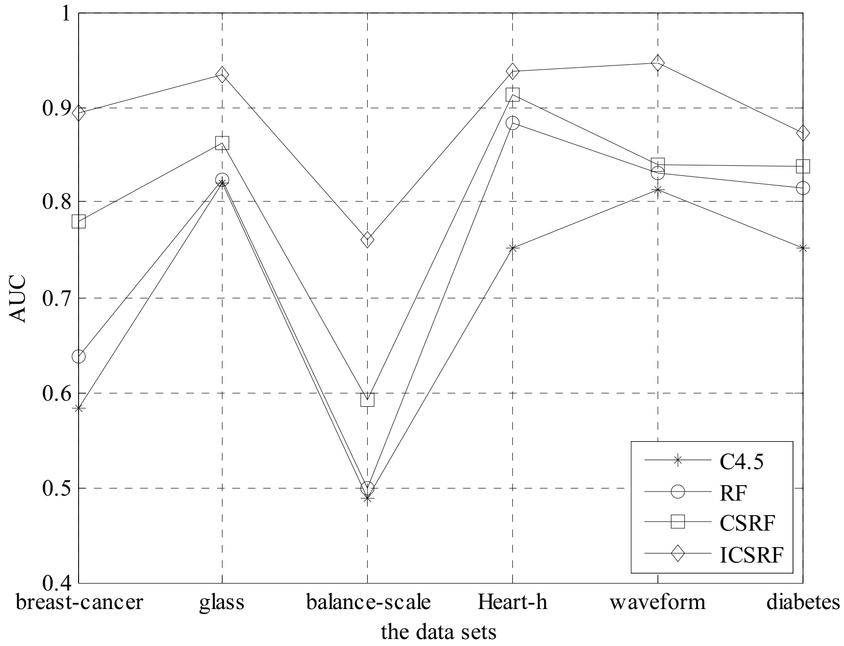

Fig. 1. The AUC information of C4.5, RF, CSRF and ICSRF

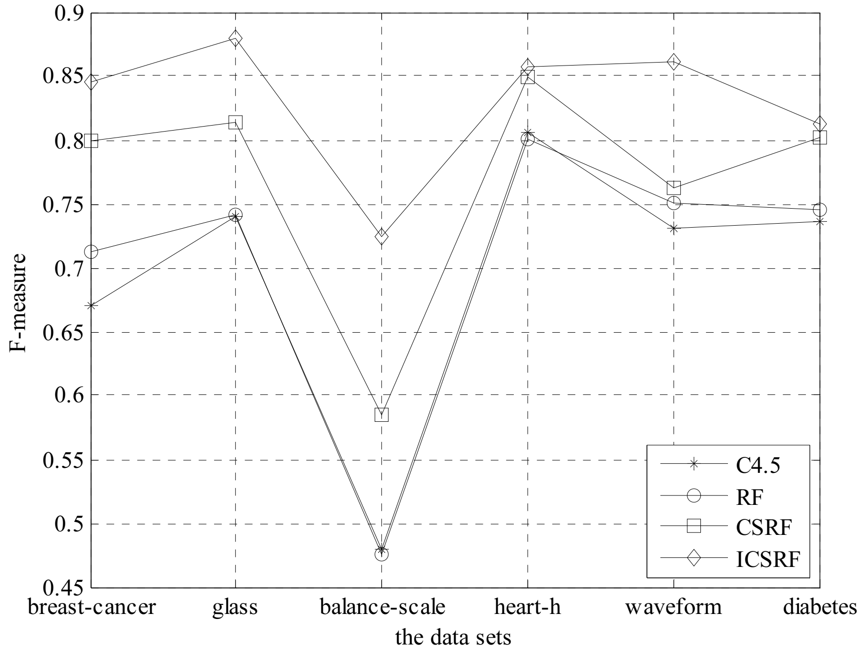

Fig. 2. The F-measure information of C4.5, RF, CSRF and ICSRF

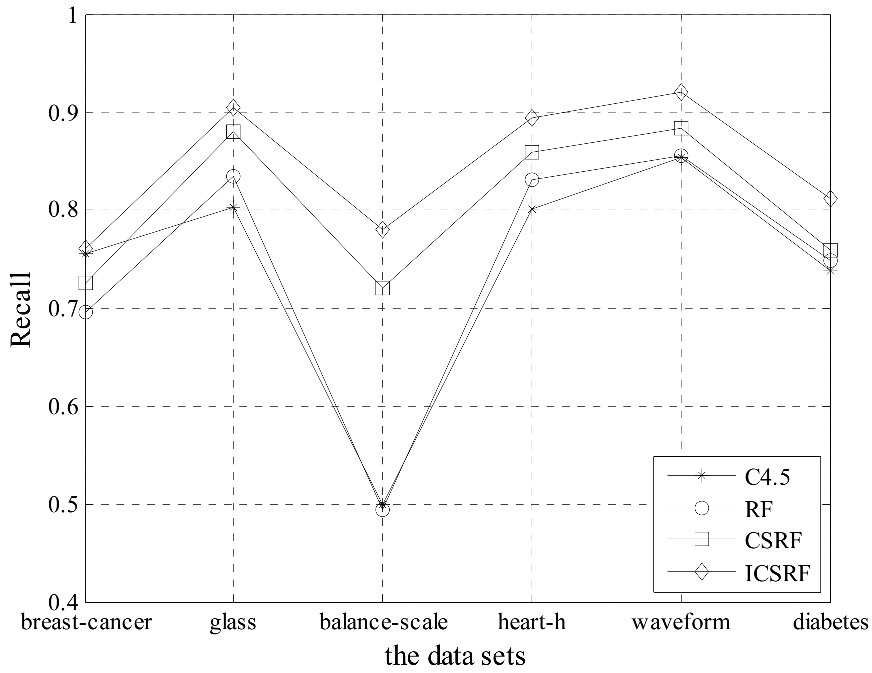

Fig. 3. The Recall information of C4.5, RF, CSRF and ICSRF

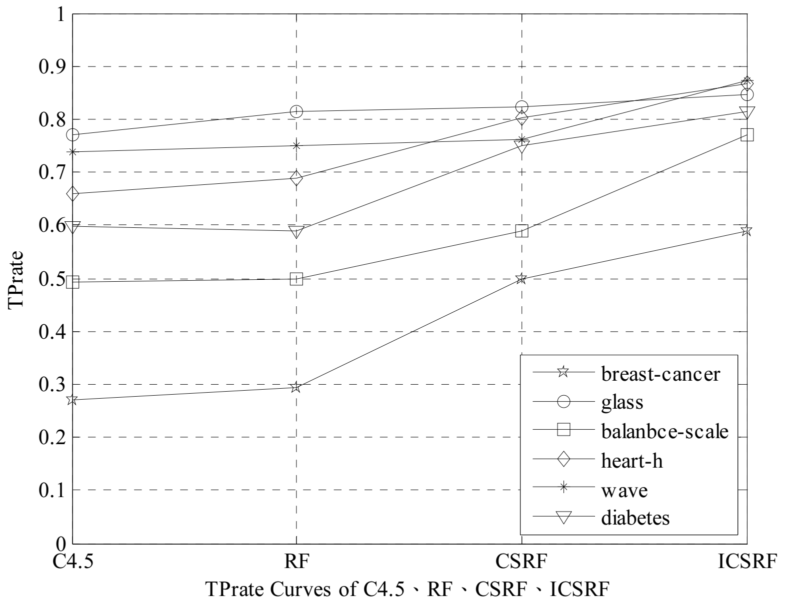

Fig. 4. TPrate Curves of C4.5, RF, CSRF, ICSRF

Fig. 1 and Fig. 2 show the performance measures of AUC, F-measure and Recall, which use four classification algorithms based on six imbalanced datasets. It can be seen that ICSRF algorithm outperforms that based on the other algorithms. The node of decision tree algorithm is made up of attributes, the C4.5 algorithm uses the information gain rate to attributes splitting, which is more intuitive and easy to implement, but is prone to over-fitting. In addition, the classifier is more complex when the data is high-dimensional and does not consider the imbalanced data. So the performance of minority class is poor. Random forest has certain advantage over the decision tree. It utilizes both bagging and a random subspace method and it does not suffer from over-fitting. But it causes that the base classifiers are not of equal accuracy in imbalanced data. If adopting the primordial equal voting methods, it is unfair to the more competent classifiers for the final prediction and reduces the performance of the classifier. Furthermore the recognition rate of the minority class is low. Random forest does not achieve better results when using small amount of data. Therefore, the performances are not improved a lot by comparing with the decision tree. The traditional cost sensitive Random forest can deal with imbalanced data better and the AUC measure result is superior to that based on decision tree algorithm, Random forest. But it ignores the importance of construction of cost function in the classification. It sets the misclassification cost just according to the proportion of the majority class and the minority class, which are not consistent with the actual distribution of the sample. The effect of cost sensitive learning cannot be guaranteed. The proposed ICSRF restructures cost function that considers the actual distribution and feature weight of the data set. It makes greater contribution of cost sensitive learning. And we introduce the weighted voting method for the final prediction. The performance of the classifier is greatly improved by this method.

Fig. 1, Fig. 2 and Fig. 3 show the proposed method ICSRF is significantly superior to other algorithms. But the classification performance of ICSRF to minority classes can not be observed intuitively. We can observe the accuracy of the four classification algorithms for the minority class and the difference between the algorithms. Fig. 4 indicates the accuracy curves of the minority class and uses four classification algorithms based on six imbalanced datasets. True positives rate (TPrate) represents the probability that the minority samples are correctly divided into the minority categories.

We can see from Fig. 3 that ICSRF improves the accuracy of minority class. In smaller degree of imbalance, the ICSRF algorithm has not an efficient enhancement on TPrate. But in the balance-scale and breast-cancer which has a greater imbalance rate, the accuracy of minority class is more pronounced. It shows that the algorithm can deal with imbalanced data better. AUC reflects the overall performance of the classifier and the performance of AUC achieves a certain increase as shown in the Fig. 1. And it can be see that ICSRF algorithm sacrifices the accuracy of the majority class, but obviously improves the accuracy of minority class.

Table 4 to Table 10 show results of the time and the classification accuracy of different categories on six different data sets.

Table 4. Comparison results of breast-cancer data set

algorithm	running time	accuracy of majority class	accuracy of minority class
C4.5	0.09	96	27.1
RF	0.37	86.6	29.4
CSRF	0.43	83.4	49.1
ICSRF	0.49	81.6	59

Table 5. Comparison results of glass data set

algorithm	running time	accuracy of majority class	accuracy of minority class
C4.5	0.04	81.4	77.1
RF	0.21	80.1	81.2
CSRF	0.35	74.9	82.2
ICSRF	0.39	70.5	84.7

Table 6. Comparison results of balance-scale data set

algorithm	running time	accuracy of majority class	accuracy of minority class
C4.5	0.05	98	49.3
RF	0.42	92	50
CSRF	0.48	90.2	58.9
ICSRF	0.54	80.6	77

Table 7. Comparison results of heart-h data set

algorithm	running time	accuracy of majority class	accuracy of minority class
C4.5	0.02	89.4	64
RF	0.28	86.7	68.9
CSRF	0.75	80.9	80.3
ICSRF	0.78	78.5	86.7

Table 9. Comparison results of waveform data set

algorithm	running time	accuracy of majority class	accuracy of minority class
C4.5	0.14	87.3	73.8
RF	3.57	79.1	75.1
CSRF	3.68	75.6	76.3
ICSRF	3.95	72.2	87.3

Table 10. Comparison results of diabetes data set

algorithm	running time	accuracy of majority class	accuracy of minority class
C4.5	0.02	81.3	59.7
RF	0.237	83.4	59
CSRF	0.287	81.2	74.9
ICSRF	0.289	78.7	81.4

The tables indicate that the decision tree has the shortest modeling time, but it has the worst performance for the minority class. RF, CSRF, ICSRF have little difference on running time and the accuracy of majority class decreases slightly. However, the classification accuracy of the minority classes in CSRF and ICSRF models has been greatly improved, and ICSRF has the best performance for minority classes. Table 4 and Table 6 show that the accuracy of breast cancer datasets and balanced scale datasets, which is better than other data sets. Because the imbalanced degrees of the former are greater and it can indicate ICSRF algorithm can effectively deal with imbalanced data set.

5 Conclusion

Imbalanced classification problems have brought challenges to the traditional algorithms in machine learning [15]. Most traditional algorithms are not suitable for imbalanced tasks because such models are based on the hypothesis of class-balanced. Generally, cost sensitive is combined with classification algorithms [16]. And the misclassification cost of the traditional cost sensitive is determined by the proportion of the majority class and the minority class. While the actual cost is difficult to estimate accurately. Thus the cost sensitive learning does not have the maximum impact. Our method restructures cost function based on the actual distribution of the data. It adopts weighted distance in the process of construction of cost function. Furthermore ICSRF algorithm assigns more weights to the more competent classifiers for the final prediction. Because the base classifiers in random forests are not of equal accuracy. AUC is a performance index for imbalanced data classification and determines the weight of the base classifiers in the voting phase. Our results conclusively show that ICSRF method is superior to other algorithms.

Acknowledgements

This research is supported by National Natural Science Foundation of China under Grant No. 51437003 and the science and technology projects in Jilin Province under Grant No. 20160623004TC.

References

[1] L.-Z. Peng, H. L. Zhang, B. Yang, Y.-H. Chen, A new approach for imbalanced data classification based on data gravitation, Information Sciences An International Journal 288(C)(2014) 347-373.

[2] H. Yin, Y.-P. Hu, A cost-sensitive algorithm based on random forest, Engineering Journal of Wuhan University 47(5)(2014) 707-711.

[3] J. Yang, Y. Liu, X. Zhu, Z. Liu, X. Zhang, A new feature selection based on comprehensive measurement both in inter-category and intra-category for text categorization, Information Processing & Management 48(4)(2012) 741-754.

[4] B. Krawczyk, M. Woźniak, G. Schaefer, Cost-sensitive decision tree ensembles for effective imbalanced classification, Applied Soft Computing 14(1)(2014) 554-562.

[5] L. Yin, Y. Ge, K. Xiao, X. Wang, X. Quan, Feature selection for high-dimensional imbalanced data, Neurocomputing 105(3)(2013) 3-11.

[6] Y. Sun, M.S. Kamel, A.K.C. Wong, Y. Wang, Cost-sensitive boosting for classification of imbalanced data, Pattern Recognition 40(12)(2007) 3358-3378.

[7] F. Herrera, On the use of MapReduce for imbalanced big data using Random Forest, Information Sciences An International Journal 285(C)(2014) 112-137.

[8] Y.L. Chen, C.C. Wu, K. Tang, Time-constrained cost-sensitive decision tree induction, Information Sciences 354(C)(2016) 140-152.

[9] S. Kulluk, A cost-sensitive classification algorithm, Knowledge-Based Systems 95(C)(2016) 99-113.

[10] F. Chang, J. Guo, W. Xu, K. Yao, A feature selection method to handle imbalanced data in text classification, Journal of Digital Information Management 13(3)(2015) 169-175.

[11] K. Kim, A hybrid classification algorithm by subspace partitioning through semi-supervised decision tree, Pattern Recognition 60(2016) 157-163.

[12] P. Xu, S. Lin, Internet traffic classification using C4.5 decision tree, Journal of Software 20(10)(2009) 407-414.

[13] S. Jang, S.H. Park, J.G. Baek, Real-time contrasts control chart using random forests with weighted voting, Expert Systems with Applications 71(2017) 358-369.

[14] S. Maldonado, R. Weber, F. Famili, Feature selection for high-dimensional class-imbalanced data sets using Support Vector Machines, Information Sciences 286(2014) 228-246.

[15] P. Soda, A multi-objective optimisation approach for class imbalance learning, Pattern Recognition 44(8)(2011) 1801-1810.

[16] Z. Xing, M. Wang, Z. Yang, J. Ning, Cost-sensitive learning for social network analysis and network measurement, Computers & Electrical Engineering 61(2017) 67-79.