2022 4th International Conference on Advances in Computer Technology, Information Science and Communications (CTISC) | 978-1-6654-5872-6/22/$31.00 ©2022 IEEE | DOI: 10.1109/CTISC54888.2022.9849803

A Combination Method of Resampling and Random Forest for Imbalanced Data Classification

LIU Zheng
State Key Laboratory of Mathematical Engineering and Advanced Computing
Information & Engineering University
Zhengzhou, China
liuz1005@foxmail.com

QIU Han
State Key Laboratory of Mathematical Engineering and Advanced Computing
Information & Engineering University
Zhengzhou, China
qiuhan410@aliyun.com

ZHU Junhu
State Key Laboratory of Mathematical Engineering and Advanced Computing
Information & Engineering University
Zhengzhou, China
1641873027@qq.com

***Abstract*—In the research of imbalanced data classification, the resampling effect of the existing resampling and random forest combination technology is greatly affected by the characteristic dimension of the training set, resulting in the unsatisfactory classification effect of the model. To solve this problem, a combination method of resampling and Random Forest for imbalanced data classification is proposed. The core idea is to resample the original data set multiple times by using the feature subset of each subtree in the random forest. By reducing the number of features used in a single resampling, the influence of feature dimension on resampling effect is reduced; Different feature subsets are used for resampling in different subtrees to improve the diversity of base classifiers. Based on SMOTE, B-SMOTE and ADASYN resampling techniques, three different optimization algorithms PSRF, PBSRF and PADRF are realized respectively. Taking the geometric mean and recall rate as evaluation indexes, the experimental results on 10 groups of KEEL data show that compared with the algorithm before optimization, the geometric mean values of the three algorithms are increased by 0.85%, 2.93% and 0.79% respectively, and the recall rates are increased by 1.44%, 3.26% and 1.07% respectively.**

Keywords—machine learning; imbalanced data; resampling; Random Forest

I. INTRODUCTION

In credit card fraud detection, network intrusion detection, medical diagnosis and other practical fields, the number of abnormal events is far less than the number of normal events, and the samples used to build the classification model are imbalanced[1]. In this case, traditional machine learning algorithms prefer to judge abnormal (minority) samples as normal (majority) samples, resulting in low detection rate of abnormal events. The research shows that the combination of resampling and random forest technology can effectively improve the recognition ability of classification model for minority samples in imbalanced data.

Resampling technology is to balance the proportion of samples in the data set by increasing the number of minority samples or reducing the number of majority samples, including under sampling, oversampling and mixed sampling. Under sampling easily leads to the loss of characteristic information of majority samples. The essence of mixed sampling technology is the mixture of oversampling and under sampling[2]. Therefore, this paper is mainly based on oversampling technology. Oversampling refers to increasing the number of minority samples and improving the distribution of samples through artificial synthesis. Existing oversampling technologies include SMOTE[3], Borderline-SMOTE[4], ADASYN[5], etc. Among them, the SMOTE algorithm proposed by Chawla is a classical oversampling method. Its core idea is to calculate the Euclidean distance between sample points and synthesize minority samples by interpolation between minority samples. On this basis, researchers proposed some improved algorithms, including RB-SMOTE[6], BL-SMOTE[7], SMOTE-RSB[8], etc. Through the research on the distribution of sample points, these algorithms strengthen the oversampling of minority samples that are easy to lead to misclassification, and improve the recognition ability of minority samples at the same sampling rate. However, when the feature dimension of the data set continues to increase, the Euclidean distance difference between the query point and other points will gradually decrease. At this time, the adjacent query algorithm will lose its significance[9,10], and the sampling effect of the over sampling method based on this idea will also be affected, which will seriously affect the classification performance of the model.

Random Forest (RF)[11] is an ensemble learning method based on bagging. By randomly selecting samples and features, a random forest model composed of multiple different decision tree models is constructed. However, the imbalanced number of samples is not considered in the sample selection of random forest, which leads to its weak ability to identify minority samples. Researchers use resampling technology to optimize the Random Forest. Literature[12] proposed resampling the sub data sets sampled by bootstrap of each sub tree of random forest, tried three different resampling algorithms, and proposed the SMOTEBagging algorithm; Literature[13] introduced oversampling factor to make the size of training sets between subtrees not completely consistent and improve the diversity of base classifiers; Literature[14] fused cascaded up sampling and down sampling of sub data sets to improve the diversity of base classifier; Literature[15] improved over sampling and random forest respectively through weighting strategy. The above researches improve the processing ability of random forest model for imbalanced data through different means, and they rely on all features for data resampling in the resampling stage. When the feature dimension of imbalanced data set is high, its classification performance will still be seriously affected by the feature dimension.

To solve this problem, a combination method of resampling and Random Forest for imbalanced data classification is proposed in this paper. For the random forest with n subtrees,

978-1-6654-5872-6/22/$31.00 ©2022 IEEE

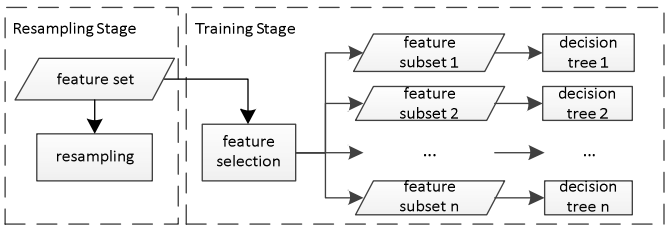
firstly, *n* feature subsets are constructed by random selection. Secondly, the original data is resampled and bootstrap sampled by using the feature subsets, and *n* balance training subsets are constructed. Then train the decision tree model. By reducing the number of features used in single resampling, this method reduces the influence of high-dimensional features on resampling effect; By using different feature subsets in different subtrees for resampling and model training, the diversity of base classifiers is improved. Based on SMOTE, Borderline-SMOTE and ADASYN resampling algorithms, three different optimization algorithms, PSRF, PBSRF and PADRF, are implemented. The geometric mean and recall rate are used as evaluation indexes to compare with the pre optimization algorithm. The experimental results on 10 groups of KEEL data show that the geometric mean of the three algorithms are improved by 0.85%, 2.93% and 0.79% respectively, and the recall rate is improved by 1.44%, 3.26% and 1.07% respectively.

II. A COMBINATION METHOD OF RESAMPLING AND RANDOM FOREST

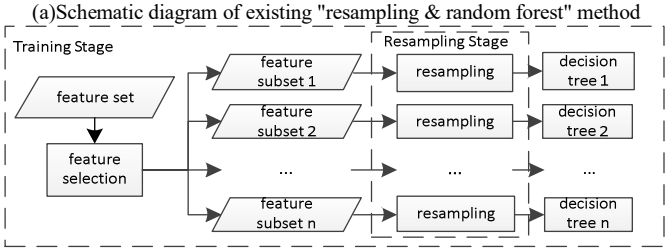
A. Mechanism Analysis

In view of the problems of random forest model in dealing with imbalanced data classification, researchers combine resampling with random forest technology, and discuss the sample distribution, which improves the recognition ability of the classification model for minority samples to a certain extent.

The random forest model constructs a decision tree model based on the feature subset composed of several features, which has strong high-dimensional data processing ability. However, as shown in Fig. 1(a), the existing "resampling & random forest" method use all features for data resampling in the resampling stage, and the effect is greatly affected by the feature dimension of the data set, which fails to give full play to the advantages of random forest model in high-dimensional data processing.



(a)Schematic diagram of existing "resampling & random forest" method



(b) Schematic diagram of this method

Figure 1. Method diagram

In the model training stage, the classification performance of a single subtree will not be affected by the unselected features. Therefore, this paper proposes a new random forest optimization method for imbalanced data. The core idea is to select the feature subset, and then use the feature subset to resample the original data set. The schematic diagram of the method is shown in Fig. 1(b). This method reduces the number of features used in single resampling and slows down the influence of high-dimensional features on the resampling effect. Moreover, different feature subsets are used in different subtrees for resampling and model training, which increases the diversity of basic classifiers and improves the classification performance of the model.

B. Framework Design

Based on the above analysis, a combination method of resampling and Random Forest for imbalanced data classification is proposed in this paper. Since the main contribution of the method is to improve the processing capacity of the model for high-dimensional imbalanced data by reducing the number of features used in single resampling, there are no special requirements for the resampling technology itself. In practical application, the appropriate resampling technology can be selected in combination with the actual situation of the data set.

The Random Forest optimization method for imbalanced data includes training stage and testing stage. In the training stage, firstly, *n* different feature subsets are constructed by sampling without replacement, and then these feature subsets are used to resample and bootstrap the training set respectively to obtain *n* balanced training subsets. Then, the *n* balanced training subsets are trained by decision tree model, and a random forest model composed of *n* decision tree models is obtained. In the test stage, first input the test data into *n* decision tree models, then use each decision tree model to judge the test data respectively, and then give the final classification results according to the voting results. It should be noted that although this method increases the number of resampling, the actual running time will not increase significantly because the random forest subtrees are relatively independent and can run in parallel.

C. Algorithm Description

The combination method of resampling and Random Forest for imbalanced data classification proposed in this paper has no specific requirements for the specific resampling technology. In the implementation of the algorithm, PSRF (Partial Features SMOTE Random Forest) algorithm is proposed based on SMOTE resampling technology, PBSRF (Partial Features Borderline-SMOTE Random Forest) algorithm is proposed based on Borderline-SMOTE resampling technology, and PADRF (Partial Features ADASYN Random Forest) algorithm is proposed based on ADASYN resampling technology. Since the core idea and algorithm framework of the three algorithms are consistent, and the main difference lies in the selection of resampling technology, this section only describes the specific steps of PSRF algorithm. The specific steps of PSRF algorithm are shown in algorithm 1.

Algorithm 1: PSRF

Input:

Number of random forest subtrees: *n*,

Number of dataset features: k,

Training set: $S = \{(x_1, y_1), (x_2, y_2), \cdots, (x_m, y_m)\}$,

Test set: $S' = \{(x'_1, y'_1), (x'_2, y'_2), \cdots, (x'_{m'}, y'_{m'})\}$.

Output:

Classification model: $H(x)$, Test result

Training stage:

1. Building feature subsets F_i ($i=1, 2, \ldots, n$)

By selecting d features from the feature set of training data without putting back sampling, the feature subset corresponding to random forest subtree is constructed. Formula 1 shows the common values of d:

$$d = \log_2 k + 1 \quad or \quad d = \mathrm{sqrt}(k) \tag{1}$$

2. Constructing balance training subset S_i ($i=1, 2, \ldots, n$)

2.1 Using the d features in F_i, the Euclidean distance from each minority sample to the other minority samples in the training set S_i is calculated, and its K proximity is sorted;

2.2 According to the requirements of sampling rate, m samples are randomly selected from the K neighborhood of minority samples to complete the construction of new samples. Each newly generated sample is located between two minority samples. The calculation method is to multiply the distance dif_j between the two minority samples by the random number between 0-1. The calculation formula is as follows:

$$x_{new} = x + rand(0,1) \times dif_j \tag{2}$$

2.3 For the balance training set obtained in 2.2, the balance training subset S_i is obtained by bootstrap resampling.

3. Building decision tree model H_i ($i=1, 2, \ldots, n$)

3.1 At the decision tree node, select the feature based on the minimum Gini index obtained from formula 4 from the feature subset F_i as the split feature, where D represents the current data set, D_l and D_r represent the left subset and right subset of D divided by feature t respectively;

$$Gini(D) = 1 - \sum_{k=1}^{d} \left(\frac{C_k}{|D|}\right) \tag{3}$$

$$GiniSplit(t) = Gini(D) - \frac{|D_l|}{|D|} Gini(D_l) - \frac{|D_r|}{|D|} Gini(D_r) \tag{4}$$

3.2 Do not prune until the tree grows to the maximum, and get the decision tree model H_i.

Testing stage:

Select sample x from test sample set S' to obtain $H_i(x_m)$. According to the majority voting algorithm, the final classification result is $f(x_m) = MajorityVote(\{H_i(x_m)\}_{i=1}^{n})$

III. EXPERIMENT

The algorithm in this paper is implemented based on Python 3.8.10. The random forest subtree is implemented based on the decision tree algorithm in scikit-learn library.

A. Data Set

In order to verify the effectiveness of this algorithm, ten imbalanced data sets are selected from KEEL. The details are shown in Table I.

TABLE I. EXPERIMENTAL DATA SET

The Name of Data Set	Number of Examples	Imbalance Rate	Number of Attributes
ecoli3	336	8.6	7
ecoli4	336	15.8	7
glass6	214	6.4	9
newthyroid2	215	5.1	5
page-blocks0	5472	8.8	10
segment0	2308	6.0	19
vehicle0	846	3.3	18
vowel0	988	10.0	13
yesat3	1484	8.1	8
yeast5	1484	32.7	8

B. Evaluation Metrics

This paper focuses on the classification of imbalanced data, and uses the recall rate of minority samples and geometric mean as the performance evaluation index of classification algorithm. Its definition is based on the confusion matrix in Table II.

TABLE II. TABLE TYPE STYLES

	Predicted Positive	Predicted Negative
Positive	TP	FN
Negative	FP	TN

TP represents the number of correctly classified minority classes, FN represents the number of incorrectly classified minority classes, FP represents the number of incorrectly classified majority classes, and TN represents the number of correctly classified majority classes.

Recall rate refers to the proportion of the number of correctly predicted samples in the total number of samples in this category. In the problem of imbalanced data classification, we pay more attention to the recall rate of minority samples. Its calculation formula is:

$$x_{Rec} = \frac{TP}{TP + FN} \tag{5}$$

The geometric mean is the comprehensive performance of the recall rate of majority samples and the recall rate of minority samples, which has good robustness. Its calculation formula is:

$$x_{G-mean} = \sqrt{\frac{TP}{TP+FN} \times \frac{TN}{TN+FP}} \quad (6)$$

C. Experiment

This section illustrates the effectiveness of this method by comparing the performance differences caused by different combination methods when using the same oversampling algorithm. Therefore, we use the resampling and random forest combination method proposed in this paper to realize PSRF, PBSRF and PADRF algorithms. Among them, PSRF is based on SMOTE algorithm, PBSRF is based on Borderline-SMOTE algorithm, and PADRF is based on ADASYN algorithm. We compare PSRF, PBSRF and PADRF proposed in this paper with RF, SMOTE&RF, Borderline-SMOTE&RF and ADASYN& RF algorithms. Among them, RF refers to the original random forest model without resampling, SMOTE&RF refers to oversampling with SMOTE before random forest classification, Borderline-SMOTE&RF refers to oversampling with Borderline-SMOTE before random forest classification, and ADASYN&RF refers to oversampling with ADASYN before random forest classification. The experimental results are shown in Table III and Table IV.

TABLE III. COMPARISON OF G-MEAN VALUE OF EACH ALGORITHM

Name	Algorithm						
	RF	***SMOTE &RF***	***PSRF***	***Borderline-SMOTE&RF***	***PBSRF***	***ADASYN &RF***	***PADRF***
ecoli3	0.7039	0.7802	**0.8127**	0.7192	**0.8153**	0.8075	**0.8286**
ecoli4	0.8433	0.9356	**0.9403**	0.7372	**0.9092**	0.8650	**0.9325**
glass6	0.8421	**0.8850**	0.8846	0.8371	**0.8626**	0.9015	**0.9042**
newthyroid2	0.9542	0.9486	**0.9647**	0.9767	**0.9796**	0.9796	0.9796
page-blocks0	0.9086	0.9352	**0.9380**	0.9405	**0.9432**	0.9414	**0.9433**
segment0	0.9873	0.9916	0.9916	0.9919	**0.9921**	0.9919	**0.9921**
vehicle0	0.9650	**0.9661**	0.9628	0.9686	**0.9728**	0.9663	**0.9704**
vowel0	0.9704	**0.9876**	0.9824	0.9809	**0.9938**	0.9736	**0.9764**
yeast3	0.7364	0.8377	**0.8633**	0.8480	**0.8666**	0.8563	**0.8757**
yeast5	0.5892	0.8909	**0.9029**	**0.8516**	0.8091	**0.9037**	0.8628
AVERAGE	0.8500	0.9158	**0.9243**	0.8851	**0.9144**	0.9187	**0.9266**
OPTIMIZATION EFFECT		**0.85%**		**2.93%**		**0.79%**	

By comparing the experimental results of RF algorithm and other algorithms, it can be found that both the existing "resampling & random forest" algorithm and the three optimization algorithms proposed in this paper can significantly improve the performance of the classification model in the imbalanced data classification problem.

Compared with SMOTE&RF algorithm, the classification effect of PSRF algorithm proposed in this paper is better. Specifically, the g-mean of PSRF is better than SMOTE&RF on 6 data sets, and the recall on 5 data sets is better than SMOTE&RF algorithm. Overall, compared with SMOTE&RF, the g-mean of PSRF algorithm proposed in this paper is improved by 0.85% and the recall is improved by 1.44%.

Compared with Borderline-SMOTE&RF algorithm, PBSRF algorithm proposed in this paper has better classification effect. Specifically, the g-mean of PBSRF algorithm is better than Borderline-SMOTE&RF algorithm on 9 data sets, The recall on 7 data sets is better than Borderline-SMOTE&RF algorithm. Overall, compared with Borderline-SMOTE&RF algorithm, the g-mean of PBSRF algorithm proposed in this paper is increased by 2.93% and the recall is increased by 3.26%.

Compared with ADASYN&RF algorithm, PADRF algorithm proposed in this paper has better classification effect. Specifically, the g-mean of PADRF algorithm is better than ADASYN&RF algorithm on 8 data sets, the recall on 5 data sets is better than ADASYN&RF algorithm. Compared with ADASYN&RF algorithm, the g-mean of PADRF algorithm proposed in this paper increases by 0.79% and the recall rate increases by 1.07%.

TABLE IV. COMPARISON OF RECALL VALUE OF EACH ALGORITHM

Name	Algorithm						
	RF	***SMOTE &RF***	***PSRF***	***Borderline-SMOTE&RF***	***PBSRF***	***ADASYN &RF***	***PADRF***
ecoli3	0.5143	0.6571	**0.7143**	0.5714	**0.7143**	0.7143	**0.7429**
ecoli4	0.7500	0.9000	0.9000	0.7000	**0.8500**	0.8000	**0.9000**
glass6	0.7333	0.8000	0.8000	0.7333	**0.7667**	0.8333	0.8333
newthyroid2	0.9143	0.9143	**0.9429**	0.9714	0.9714	0.9714	0.9714
page-blocks0	0.8339	0.8946	**0.8982**	0.9054	**0.9089**	0.9089	**0.9125**
segment0	0.9758	0.9848	0.9848	0.9848	0.9848	0.9848	0.9848
vehicle0	0.9450	**0.9600**	0.9550	0.9650	**0.9750**	0.9650	**0.9700**
vowel0	0.9444	**0.9778**	0.9667	0.9666	**0.9889**	0.9556	0.9556
yeast3	0.5515	0.7152	**0.7636**	0.7394	**0.7697**	0.7515	**0.7879**
yeast5	0.3556	0.8000	**0.8222**	**0.7333**	0.6667	**0.8222**	0.7556
AVERAGE	0.7518	0.8604	**0.8748**	0.8270	**0.8596**	0.8707	**0.8814**
OPTIMIZATION EFFECT		**1.44%**		**3.26%**		**1.07%**	

IV. CONCLUSION

In the existing research on the classification of imbalanced data sets, the interference of the feature dimension of data sets on the resampling technology is ignored, and the method of randomly selecting features from subtrees in random forest model can better deal with this problem. Therefore, this paper proposes a combination method of resampling and Random Forest for imbalanced data classification. The core idea is to resample the original data set multiple times by using the feature subset of each subtree in the random forest. By reducing the number of features used in single resampling, the influence of feature dimension on resampling effect is reduced. Moreover, different feature subsets are used for resampling in different subtrees to improve the diversity of base classifiers. Based on SMOTE, Borderline-SMOTE and ADASYN resampling technology, three different optimization algorithms of PSRF, PBSRF and PADRF are proposed respectively. The effectiveness of this method is verified by comparing with the pre optimization algorithm on 10 groups of KEEL data. In the next step, we plan to study the feature correlation and optimize the feature selection method of the random forest subtree.

REFERENCES

[1] WANG Le, HAN Meng, LI Xiaojuan, ZHANG Ni, CHENG Haodong. Review of Classification Methods for Unbalanced Data Sets[J]. Computer Engineering and Applications, 2021, 57(22):11.

[2] LIU Dingxiang, QIAO Shaojie, ZHANG Yongqing, et al. A Survey on Data Sampling Methods in Imbalance Classification[J]. Journal of Chongqing University of Technology(Natural Science) , 2019, 33(7):102-112.

[3] Chawla N V , Bowyer K W , Hall L O , et al. SMOTE: Synthetic Minority Over-sampling Technique[J]. Journal of Artificial Intelligence Research, 2002, 16(1):321-357.

[4] Han H , Wang W , Mao B . Borderline-SMOTE: A New Over-sampling Method in Imbalanced Data Sets Learning[C]// Advances in Intelligent Computing, International Conference on Intelligent Computing, ICIC 2005, Hefei, China, August 23-26, 2005, Proceedings, Part I. 2005.

[5] He H , Yang B , Garcia E A , et al. ADASYN: Adaptive synthetic sampling approach for imbalanced learning[C]// Neural Networks, 2008. IJCNN 2008. (IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on. IEEE, 2008.

[6] YANG Yi, LU Chengbo, XU Genhai. A Refined Borderline-SMOTE Method for Imbalanced Data[J]. Journal of Fudan University(Natural Science), 2017 ,56(5):8.

[7] ZHANG Chenning, LI Guocheng. Imbalanced Data Classification Based on BL-SMOTE and Random Forest[J]. Journal of Beijing Information Science & Technology University, 2019, 34(2):6.

[8] Ramentol E , Yailé D Caballero, Bello R , et al. SMOTE-RSB*[J]. Knowledge and Information Systems, 2012.

[9] Hughes G F . On the mean accuracy of statistical pattern recognizers[J]. IEEE Transactions on Information Theory, 2006.

[10] Beyer K , Goldstein J , Ramakrishnan R , et al. When Is "Nearest Neighbor" Meaningful?[C]// International Conference on Database Theory. Springer, Berlin, Heidelberg, 1999.

[11] Breiman L . Random Forests[J]. Machine Learning, 2001.

[12] Wang S , Xin Y . Diversity analysis on imbalanced data sets by using ensemble models[C]// IEEE Symposium on Computational Intelligence & Data Mining. IEEE, 2009.

[13] ZHENG Jianhua, LIU Shuangyin, HE Chaobo, et al. lmproved Random Forest Classification Algorithm for lmbalance Data Based on Hybrid Sampling Strategy[J].Journal of Chongqing University of Technology (Natural Science) ,2019,33 (7):113 一 123.

[14] ZHENG Jianhua,LI Xiaomin,LIU Shuangyin,and LI Di. Improved Random Forest Imbalance Data Classification Algorithm Combining Cascaded Upsampling and Downsampling [J]. Computer Science, 2021, 48(7):10

[15] ZHANG Jiawei,GUO Linming,YANG Xiaomei. Improved Oversampling and Random Forest Algorithm for Imbalanced Data [J]. Computer Engineering and Applications, 2020, 56(11):7