Fuzzy Systems and Data Mining VI
A.J. Tallón-Ballesteros (Ed.)

doi:10.3233/FAIA200684

Classification of Unbalanced Data Based on RSM and Binomial Distribution

LI Rong [a,1] and ZHOU Wei-bai [a]
[a] *Guangzhou college of Commerce, China*

Abstract. In the case of extremely unbalanced data, the results of the traditional classification algorithm are very unbalanced, and most samples are often divided into the categories of majority samples, so the accuracy of judgment of the minority classes will be reduced. In this paper, we propose a classification algorithm for unbalanced data based on RSM and binomial undersampling. We use RSM's random part features rather than all each classifier to make each training classifier reduce the dimensions, and dimension reduction makes relatively minority class samples indirectly lift. Using the above characteristics of the RSM to reduce dimension can solve the problem that unbalanced data classification in the minority class samples is too little, and it can also find the important attribute of variables to make the model have the ability of explanation. Experiments show that our algorithm has high classification accuracy and model interpretation ability when classifying unbalanced data.

Keywords. Machine learning; unbalanced data; multi-classifier ensemble; Random subspace method

1. Introduction

When traditional classification algorithms encounter extreme imbalance data, the classification results are very inaccurate, and most samples are divided into the categories of most samples [1-2]. This method reduces the accuracy of judgment of minority classes. However, minority classes are often more valuable than majority classes in real life [3]. In order to deal with the problem about clearly classify minority data, it is very important to study data classification method of unbalanced data sets.

The method to deal with unbalanced data is mainly from two aspects of changing sample proportion of different types data and algorithms [4, 7-8]. The method of changing the proportion of different sample categories of data is mainly to reduce the imbalance degree of data by adding or deleting samples, such as oversampling, under-sampling and under-sampling of negative binomial distribution [5-6].The algorithm includes different categories for different costs, single category learning, or adjustment of probability valuation or threshold value when training the decision tree [9], such as random forest, support vector machine, decision tree, artificial neural network, and nearest neighbor method[9-11].

In previous studies, more researchers used the method of resampling. Although the method of resampling is helpful for solving unbalanced data, it also has some fatal

[1] Corresponding Author: LI Rong, associate professor, Room 1202, No. 8, Wenmei street, Haizhu District, Guangzhou, China; E-mail: 1047161101@qq.com

defects. For example, the method of random oversampling may lead to overfitting [5-6], while the method of random under-sampling may lose some samples containing important data [5, 10-12].To solve the problem of unbalanced data classification and reduce the defects caused by resampling, we propose an algorithm combining RSM and resampling. The algorithm uses RSM method to train each classifier based on its features of using random partial characteristics instead of all characteristics, which means training sub-classifier with a small number of variables at a time. After training the classifier each time, the dimension of it is lower, so that the number of minority class samples increases. In classifications of unbalanced data, one of the most important problems is that the sample size with a small proportion is too small. it can be solved by taking advantage of the feature of RSM. This feature can also be used to find out importance of variable attributes, to enable the model to have explanatory ability. We use 6 datasets from UCI machine learning database for simulation experiments. Experimental results show that our algorithm can effectively classify unbalanced data and obtain higher classification accuracy.

2. An unbalanced data classification algorithm combining RSM and binomial distribution sampling

Based on RSM, we combine sampling method of binomial distribution. First, our algorithm determines the number of samples needed for the majority sample after undersampling by using binomial distribution, which makes the proportions of the minority class and the majority class similar in new data set. Then use the new data set to train RSM classifier, and get a final classification result by majority vote.

2.1 Binomial distribution sampling

The probability mass function of binomial distribution is shown in formula (1).

$$f(y) = \binom{n}{y} p^{y} (1-p)^{n-y}, y = 0,1,\Lambda\ ,n \tag{1}$$

Where p is the probability of success. n is the total number of experiments. y is the number of successful experiments. Figure 1 shows the distribution of y when p=0.5 and n=10. we find that the probability of y's value nearby n/2 is the highest.

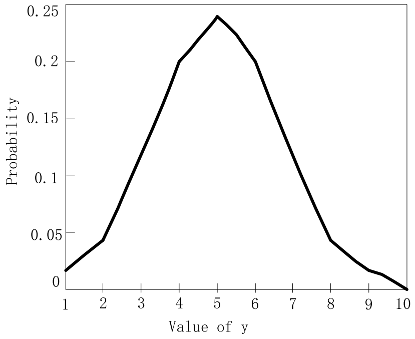

Figure 1. Binomial distribution (p=0.5, n=10).

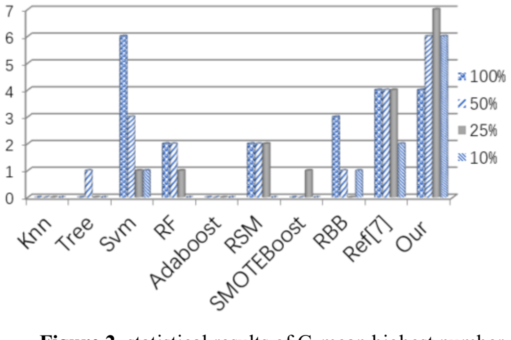

Figure 2. statistical results of G-mean highest number

We set n as 2 times the sample size of the original minority class samples and p as 0.5. The cumulative distribution function of binomial distribution was obtained, as shown in (2).

$$F(y)=\sum_{k=0}^{[y]}\binom{n}{k}p^{k}(1-p)^{n-k} \quad (2)$$

Where, *F(y)*is cumulative distribution function. k is the number of successes.

We randomly select a value between 0 and 1 as a probability value, and find out the y value corresponding to probability value. We take this y value as the number of samples after undersampling of the majority class and a new value of minority sample sets n-y. After binomial distribution sampling, we can get a new sample which has a similar size with minority class and majority class. We classify this new sample using the RSM classifier and get the final classification result.

2.2 RSM algorithm

RSM algorithm is not like random forest or Adaboost using resampling method to get different training data sets, each time RSM randomly selects a few features in characteristic space and uses these characteristics forming training data set to train classifier. This method improves the stability and efficiency of classifier, and overcomes the problem of small sample in high dimensions. RSM has the advantages of reducing dimensions by using multiple classifiers and feature selection at the same time, so it is very suitable for small sample data classification.

Our method also takes advantages of RSM's feature that uses small variables to build classification models each time to calculate the variable importance of attributes of data set. The calculation formula is shown in (3).

$$w_i=\frac{\sum_{t=1}^{NumT}V_t}{n_i} \quad (3)$$

Where w_i is the variable importance of the i^{th} attribute; V_t is a classification accuracy rate calculated by test set. The test set is formed by the t-th base classifier which do not use majority class samples after previous under-sampling. n_i is the number of times that the i$^{\text{th}}$ attribute is selected from all basic classifiers. Accumulate the accuracy of the i$^{\text{th}}$ attribute, divided by the number of times. It also uses as variable importance of the attribute.

2.3 Specific algorithm

The steps of our algorithm are as follows:

Input: training data set D in S dimension of N samples;

D is divided into minority class samples D^{pos} and majority class samples D^{neg};

```
    $D^{pos}$ and $D^{neg}$ are the sample numbers of minority class and majority class samples respectively;
    Set the basic classifier to be used as WeakLearn;
    Set the number of training times $NumT$ and the dimension of sampling $NumF$ ($0 < NumF < S$);
For $t = 1,2,\Lambda ,NumT$
    1. Use binomial distribution（$n = 2N^{pos}$, $p$ =0.5）to decide $N_t^{neg}$ and $N_t^{pos}$
    2. Resampling $N_t^{neg}$ majority class samples $D_t^{neg}$ by random under-sampling
    3. Sampling $N_t^{pos}$ minority class samples $D_t^{pos}$ by random sampling, which was extracted and not put back (If $N_t^{pos} > N^{pos}$, $D_t^{pos} = D^{pos}$+ randomly selects ($N_t^{pos}$ - $N^{pos}$)minority class samples)
    4. Use $D_t^{neg}$ and $D_t^{pos}$ to form a new training sample $D'$
    5. Select the dimension of random sampling from S-dimension training samples
    6. Copy dataset $D'$ and retain the data selected in the previous step
    7. Train classifier **WeakLearn**
    8. Use the original test data set to get the classification category $L_t$
    9. Randomly select $2 \cdot N^{pos}$ samples from the unused $D^{neg}$ as the test set, and calculate the classification accuracy $V_t$ as the variable importance weight of the $NumF$ attributes of the round
Output : 1. The result of $NumT \cdot L_t$ is calculated to the classification of the final result according to the majority vote.
         2. Use $NumT \cdot V_t$ to calculate the variable importance of each attribute
```

3. Experimental results and analysis

3.1 Experimental data

We use 6 sets of unbalanced data from the UCI machine learning database, as shown in Table 1.

Table.1 UCI data sets

data sets	total	majority class	minority class	Proportion of minority	Number of attributes
Pima Indians Diabetes	768	500	268	34.9	8
Blood Transfusion Service Center	748	570	178	23.8	4
Ionosphere	351	225	126	35.9	34
Breast Cancer Wisconsin	569	357	212	37.26	30
Echocardiogram	131	88	43	32.82	7
Statlog	1000	700	300	30	20

3.2 Evaluation indicators

Accuracy is commonly used as evaluation standard in classification problems, which reflects overall classification performance. However, it doesn't reflect the classification of imbalanced data well. We use seven common evaluation indexes to truly judge the quality of the classification results. Including Accuracy, Specificity, Precision, Recall, *F*-measure and *G*-mean. as in (4) to (9).

$$\text{Sensitivity} = \frac{TP}{TP+FN} \tag{4}$$

$$\text{Specificity} = \frac{TN}{TN+FP} \tag{5}$$

$$\text{Precision} = \frac{TP}{TP+FP} \tag{6}$$

$$\text{Recall} = \frac{TP}{TP+FN} \tag{7}$$

$$\text{F - measure} = \frac{2 \times \text{Re}\,call \times \text{Pr}\,ecision}{\text{Re}\,call + \text{Pr}\,ecision} \tag{8}$$

$$\text{G - mean} = \sqrt{\frac{TP}{TP+FN} \times \frac{TN}{TN+FP}} \tag{9}$$

Where FN is false negative. FP is false positive.TN is true negative.TP is true positive.

Since *F*-measure and *G*-mean comprehensively consider the accuracy of minority classes and majority classes, they have comprehensive comparison ability. In addition, if the classification of the minority samples is wrong, sometimes it will cause more serious losses in the classification of unbalanced data. Therefore, the accuracy of minority class is also an experimental key indicator in this paper.

3.3 Experimental design

In order to prove the superiority of method proposed in this paper, we compared our method and previous classification methods, and observed the influence degree of each method on different degree of data imbalance. In this study, the minority class samples in each data set were randomly undersampling to make sample proportion difference become larger. Minority class samples were undersampling to be 10%, 25% and 50% of the original minority samples. The data set was randomly divided into 5 sub-sets. Four sub-sets were used as the training set and the rest one was used as test set in each experiment. The above steps were repeated. Finally, we get the classification accuracy of five test sets, and the average value was taken as the average accuracy of our method.

We set the nearest neighbor parameter K to 1. The kernel function of SVM classifier is RBF kernel, parameter σ uses the method in reference [3]. Cost function *C* choose exponential sequence to search whose scope set as $2^{-12},2^{-10},\Lambda,2^{12}$; the parameter *NumT* of random forest set as 100, and *NumF* set as the attribute number in data set. Adaboost uses decision tree as its basic classifier and its *NumT* is 100. The basic classifier of RSM is decision tree and its *NumT* is 100, *NumF* is half of the number of attributes in the data set. SMOTEBoost uses a decision tree classifier with *NumT* setting as 100 and a buffer ratio of 100%. RB Bagging uses a decision tree as its base classifier with *NumT* setting as 100.

3.4 Experimental results and analysis

3.4.1 Performance comparison

We use six datasets and 4 different proportions of minority classes. The accuracy is shown in table 2 and table 3. We can find that classification accuracy of SVM, RF, RB Bagging, Reference[7](abbreviated as ref [7] in the table)and our algorithm is better in each data set. With minority class sample ratio decreasing, the accuracy of classification has a small fluctuation. In most data sets, our algorithm's classification accuracy even increased with reducing the proportion of minority class samples. It means that our algorithms can improve the classification results of unbalanced data effectively. Our algorithm performs better when the proportion of minority class samples is small, which indicates that our algorithm is more suitable for classifying data sets in cases where have a small proportion of little samples.

Table.2 performance comparison of correct rate

	Pima Indians Diabetes				**Blood Transfusion Service Center**				**Ionosphere**			
	100%	50%	25%	10%	100%	50%	25%	10%	100%	50%	25%	10%
Knn	0.68	0.73	0.83	0.94	0.70	0.68	0.79	0.89	0.87	0.89	0.92	0.96
Tree	0.69	0.74	0.83	0.91	0.73	0.82	0.87	0.96	0.89	0.97	0.96	0.95
SVM	0.74	0.77	0.77	0.84	0.73	0.71	0.69	0.74	0.95	0.95	0.96	0.98
RF	0.76	0.81	0.88	0.94	0.75	0.84	0.91	0.96	0.93	0.97	0.97	0.97
Adaboost	0.74	0.81	0.88	0.94	0.74	0.82	0.89	0.96	0.90	0.92	0.94	0.94
RSM	0.73	0.80	0.88	0.94	0.78	0.86	0.93	0.97	0.93	0.98	0.97	0.97
SMOTE Boost	0.71	0.79	0.85	0.93	0.73	0.80	0.85	0.94	0.91	0.80	0.95	0.96
RBB	0.75	0.76	0.72	0.74	0.39	0.26	0.16	0.21	0.91	0.92	0.75	0.54
ref [7]	0.74	0.73	0.72	0.75	0.63	0.59	0.64	0.63	0.93	0.95	0.90	0.66
Our	0.74	0.73	0.71	0.76	0.64	0.61	0.63	0.68	0.93	0.95	0.95	0.92

Table.3 performance comparison of correct rate

	Breast Cancer Wisconsin				**Echocardiogram**				**Statlog**			
	100%	50%	25%	10%	100%	50%	25%	10%	100%	50%	25%	10%
Knn	0.92	0.94	0.96	0.98	0.52	0.59	0.75	0.75	0.60	0.72	0.83	0.92
Tree	0.92	0.94	0.96	0.97	0.70	0.67	0.81	0.81	0.69	0.75	0.84	0.92
SVM	0.96	0.95	0.96	0.97	0.73	0.67	0.75	0.75	0.68	0.75	0.78	0.76
RF	0.96	0.96	0.98	0.98	0.70	0.75	0.87	0.87	0.77	0.84	0.90	0.96
Adaboost	0.91	0.93	0.96	0.97	0.69	0.73	0.84	0.84	0.75	0.83	0.90	0.96
RSM	0.96	0.96	0.98	0.97	0.66	0.73	0.85	0.85	0.76	0.83	0.90	0.96
SMOTEBoost	0.93	0.94	0.9	0.98	0.67	0.71	0.85	0.85	0.72	0.82	0.89	0.94

			5									
RBB	0.94	0.90	0.93	0.87	0.65	0.60	0.56	0.56	0.69	0.67	0.67	0.60
ref [7]	0.96	0.95	0.96	0.94	0.65	0.64	0.67	0.67	0.71	0.69	0.69	0.67
Our	0.96	0.94	0.95	0.93	0.65	0.69	0.73	0.73	0.71	0.70	0.71	0.69

Table.4 Performance comparison of F-measure

	Pima Indians Diabetes				**Blood Transfusion Service Center**				**Ionosphere**			
	100%	50%	25%	10%	100%	50%	25%	10%	100%	50%	25%	10%
Knn	0.55	0.37	0.24	0.23	0.39	0.22	0.16	0.15	0.78	0.68	0.56	0.23
Tree	0.56	0.43	0.28	0.14	0.37	0.21	0.08	0.19	0.85	0.92	0.82	0.53
SVM	0.67	0.56	0.31	0.13	0.47	0.31	0.15	0.03	0.93	0.89	0.83	0.79
RF	0.64	0.44	0.28	0.00	0.36	0.17	0.10	0.05	0.91	0.93	0.88	0.60
Adaboost	0.62	0.45	0.30	0.00	0.38	0.17	0.14	0.05	0.86	0.82	0.71	0.41
RSM	0.58	0.37	0.23	0.00	0.31	0.12	0.12	0.00	0.91	0.94	0.87	0.56
SMOTEBoost	0.59	0.48	0.32	0.06	0.40	0.18	0.10	0.03	0.87	0.77	0.80	0.63
RBB	0.68	0.57	0.37	0.24	0.30	0.18	0.09	0.05	0.87	0.83	0.44	0.09
ref [7]	0.65	0.54	0.38	0.23	0.45	0.26	0.19	0.06	0.90	0.89	0.69	0.29
Our	0.68	0.55	0.40	0.25	0.45	0.28	0.21	0.08	0.90	0.88	0.82	0.49

Table.5 Performance comparison of F-measure

	Breast Cancer Wisconsin				**Echocardiogram**				**Statlog**			
	100%	50%	25%	10%	100%	50%	25%	10%	100%	50%	25%	10%
Knn	0.89	0.86	0.82	0.77	0.34	0.03	0.07	0.07	0.36	0.19	0.18	0.07
Tree	0.89	0.87	0.84	0.72	0.50	0.20	0.17	0.17	0.49	0.32	0.26	0.08
SVM	0.94	0.90	0.87	0.77	0.64	0.32	0.23	0.23	0.56	0.26	0.21	0.12
RF	0.95	0.92	0.91	0.75	0.46	0.06	0.00	0.00	0.53	0.31	0.07	0.00
Adaboost	0.88	0.85	0.83	0.74	0.49	0.05	0.13	0.13	0.50	0.23	0.07	0.00
RSM	0.95	0.91	0.92	0.73	0.44	0.16	0.00	0.00	0.49	0.26	0.03	0.00
SMOTEBoost	0.91	0.87	0.82	0.77	0.51	0.04	0.20	0.20	0.50	0.29	0.09	0.04
RBB	0.93	0.82	0.76	0.44	0.56	0.36	0.26	0.26	0.59	0.43	0.29	0.12
ref [7]	0.93	0.89	0.87	0.73	0.55	0.39	0.29	0.29	0.60	0.45	0.31	0.13
Our	0.95	0.89	0.82	0.63	0.56	0.40	0.39	0.39	0.60	0.47	0.33	0.14

F-measure performance comparison of 6 data sets is shown in Table 4 and table 5. SVM, RBB and our algorithm have better classification effect. With the decrease of the proportion of minority classes samples, our algorithm performance is more stable.

Comparison of G-mean performance of 6 data sets is shown in table 6 and table 7. They indicate that in each data set, SVM, U_RSM, RBB, References[7] and our algorithm all have better effects. In the case of decreasing the proportion of minority samples, the performance of this algorithm is more stable.

Table.6 Performance comparison of G- mean

	Pima Indians Diabetes				**Blood Transfusion Service Center**				**Ionosphere**			
	100%	50%	25%	10%	100%	50%	25%	10%	100%	50%	25%	10%
Knn	0.64	0.55	0.46	0.38	0.56	0.49	0.48	0.45	0.80	0.73	0.64	0.26
Tree	0.66	0.61	0.50	0.30	0.54	0.40	0.22	0.27	0.88	0.94	0.89	0.69
SVM	0.74	0.74	0.59	0.45	0.65	0.60	0.52	0.21	0.94	0.92	0.90	0.84
RF	0.72	0.58	0.44	0.00	0.51	0.34	0.20	0.10	0.92	0.95	0.91	0.62
Adaboost	0.70	0.59	0.45	0.00	0.54	0.35	0.32	0.10	0.88	0.89	0.79	0.61
RSM	0.66	0.51	0.34	0.00	0.45	0.26	0.20	0.00	0.92	0.95	0.92	0.59
SMOTEBoost	0.68	0.63	0.52	0.11	0.56	0.37	0.28	0.10	0.90	0.74	0.87	0.80
RBB	0.75	0.76	0.70	0.77	0.43	0.35	0.27	0.36	0.90	0.92	0.75	0.49
ref [7]	0.73	0.73	0.71	0.66	0.63	0.56	0.57	0.44	0.92	0.94	0.90	0.73
Our	0.75	0.75	0.75	0.78	0.64	0.58	0.65	0.55	0.92	0.94	0.93	0.83

Table.7 Performance comparison of G- mean

	Breast Cancer Wisconsin				**Echocardiogram**				**Statlog**			
	100%	50%	25%	10%	100%	50%	25%	10%	100%	50%	25%	10%
Knn	0.91	0.90	0.88	0.86	0.47	0.08	0.13	0.13	0.51	0.39	0.41	0.16
Tree	0.92	0.92	0.91	0.84	0.60	0.32	0.27	0.27	0.62	0.53	0.50	0.19
SVM	0.95	0.94	0.94	0.91	0.73	0.48	0.43	0.43	0.68	0.46	0.49	0.46
RF	0.96	0.94	0.93	0.84	0.57	0.09	0.00	0.00	0.63	0.45	0.15	0.00
Adaboost	0.90	0.90	0.92	0.84	0.59	0.09	0.14	0.14	0.61	0.37	0.10	0.00
RSM	0.96	0.94	0.94	0.81	0.55	0.27	0.00	0.00	0.60	0.40	0.05	0.00
SMOTEBoost	0.92	0.91	0.94	0.86	0.62	0.09	0.27	0.27	0.62	0.45	0.17	0.11
RBB	0.95	0.92	0.92	0.86	0.65	0.58	0.56	0.56	0.70	0.68	0.67	0.61
ref [7]	0.95	0.93	0.95	0.86	0.62	0.48	0.38	0.38	0.00	0.00	0.00	0.00
Our	0.96	0.94	0.95	0.94	0.65	0.57	0.75	0.75	0.71	0.72	0.71	0.65

The above comparison method is to combine 7 different evaluation indicators to judge the quality of classification results. However, this way can’t be used for quantitative comparison. We choose G-mean as a representative from seven evaluation indicators for comparison to find out the highest value in each data set under different proportion. If the gap between two algorithms was not greater than 0.02, the values are considered the same. The statistical results of the highest number of G-mean are shown in figure 2.

From Figure 2, when the original proportion of minority class samples is classified, SVM algorithm has the best results of 6 data sets. Our algorithm's results of 4 data sets are the second best. However, when the proportion of minority class samples decreases, results of our algorithm are significantly better than others. When the proportion equals to 50%, 25% and 10%, our results are the best, it means that our algorithm can improve classification results of unbalanced data sets effectively. The smaller proportion of minority class samples is, the better classification results we have.

From the experimental results, the SVM algorithm has a better classification effect on the original scale data set, but the classification effect is significantly reduced when the proportion of minority class sample becomes lower. SVM is not suitable for classifying unbalanced data. RB Bagging algorithm has a good classification effect in all six data sets, but its results are unstable. Although it has a high accuracy of minority classes in some data sets, when the accuracy of multiple classes quickly drops at the same time, overall classification effect is bad. Our algorithm has a good classification effect in various data sets and minority class proportions, and it even has a better classification effect when the proportion of minority class samples and majority class samples has a big difference between each other. In original proportions, the classification effect results in 4 data sets are the best; in case of 50% and 10% of minority class samples, 5 results are the best; in case of 25% of minority class samples, all results are the best. This indicates that our algorithm is very suitable for classifying unbalanced data. Moreover, our algorithm is well suitable for unbalanced data classification algorithm.

3.4.2 Importance of variables

Our algorithm calculates the variable importance of each subset. After summing up the variables' importance of 5-fold, the data range is between 0 and 1 through the method of data normalization. Table 8 shows the variables' importance in Pima Indians Diabetes dataset whose value is between 0 and 1. the value is closer to 1, the more important of variable is. From Table 8, the most important variables are attribute 2 and attribute 6.

Table.8 importance of variables for Pima Indians Diabetes data sets

	attribute 1	**attribute 2**	**attribute 3**	**attribute 4**	**attribute 5**	**attribute 6**	**attribute 7**	**attribute 8**
100%	0.18	1.00	0.00	0.15	0.19	0.42	0.28	0.39
50%	0.19	1.00	0.12	0.00	0.08	0.48	0.03	0.13
25%	0.17	1.00	0.00	0.20	0.03	0.45	0.10	0.40
10%	0.06	1.00	0.00	0.10	0.14	0.35	0.04	0.07

4. Conclusion

There are unbalanced data problems in many different fields. This target receives much attentions in recent years related to relative issues. The past research mainly combined with a multiple classifier and many people focus on Bagging and Boosting classifier. Few people notice RSM classifier which has not only the advantages of multiple classifier, but also the ability of reducing classification dimension, indirectly promote the minority class sample effect. We use RSM combined with binomial distribution sampling to propose a binomial random subspace classification algorithm for unbalanced data. This method integrates the concepts of RSM and resampling. Experiments show that our algorithm has a higher, more balanced and more stable classification accuracy when classifying unbalanced data, and it also has the ability of model interpretation.

References

[1] WANG Canwei, YU Zhilou, ZHANG Huaxiang, New algorithm of AdaBoost for unbalanced datasets, Computer Engineering and Applications. 2011, (28)169-172,175.

[2] B Yuan, X Ma. Sampling + reweighting: Boosting the performance of AdaBoost on imbalanced datasets// International Joint Conference on Neural Network. Australia: Brisbane 2012, 1-6.

[3] Li C H, Ho H H, Liu Y L, et al., An automatic method for selecting the parameter of the normalized kernel function to support vector machines. Journal of Information Science and Engineering 2012 (1),1-15.

[4] XY Liu, J Wu, ZH Zhou, Exploratory under-sampling for class-imbalance learning, IEEE Transactions on Systems Man & Cybernetics Part B Cybernetics A Publication of the IEEE Systems 2009(2), 539-50.

[5] W Klement, S Wilk, W Michalowski, et al., classifying severely imbalanced data, Germany: Springer Berlin Heidelberg,2011.

[6] TR Hoens, Q Qian, NV Chawla, et al., Building decision trees for the multi-class imbalance problem, 16th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Malaysia: Kuala Lumpur,2012,122-134

[7] Gu Yuping, Cheng Longsheng, Classification of unbalanced data based on MTS-AdaBoost, Application Research of Computers 2018(2), 346-353.

[8] GO Feng, HUANG Hai-yan, Imbalanced Data Classification Method Based on Neighborhood Hybrid Sampling and Dynamic Ensemble, Computer Science 2017 (8),225-229.

[9] S Hido, H Kashima, Y Takahashi, Roughly balanced bagging for imbalanced data, Statistical Analysis & Data Mining 2010 (2), 412-426.

[10] MYOUNG J K, DAO K K, HONG B K, Geometric mean based boosting algorithm with over-sampling to resolve data imbalance problem for bankruptcy prediction, Expert Systems with Applications 2015 (42),1074-1082.

[11] WANG B X, JAPKOWICZ N, Boosting support vector machines for imbalanced data sets, Knowledge and Information Systems 2010 (1),1-20.

[12] Michael J Procopio, Jane Mulligan, Greg Grudic, Coping with Imbalanced Training Data for Improved Terrain Prediction in Autonomous Outdoor Robot Navigation, 2010 IEEE International Conference on Robotics and Automation Anchorage Convention District ,USA: Anchorage Alaska, 2010, 518-525.