Research of the Improved Adaboost Algorithm Based on Unbalanced Data

Shang Fuhua

School of Computer and Information Technology, Northeast Petroleum University, Daqing, 163318 China

Summary
A large of Non-equilibrium data exist in the real world, because of the traditional classification methods based on assumptions of class balance and different categories misclassification the same costs as well as the evaluation criteria based on the accuracy of the overall sample classification , resulting in the classification of non-equilibrium data has not apply. Classification for unbalanced data, Adaboost algorithm and its adaptability in the classification of non-equilibrium data were analyzed in this paper first, followed by proposed an improved method for their classification in a non-equilibrium defects in the data, and finally proceed effectiveness analysis to improve methods through the experiment.

Key words:
Non-equilibrium data, AdaBoost algorithm, classification performance

1. Adaboost algorithm analysis

1.1 Adaboost algorithm principle

Boosting is an important element algorithm framework, the basic idea is the weak learning model as the basis for a number of iterations, the base classifiers get from each iteration will be iteration integrated, thus completing the improve of performance . Based on probably approximately correct (PAC) learning model, the conjecture of 1990 Schapire can learn is equivalent to the strong can learn [1], in order to provide a theoretical confirmed for the validity of Boosting.

AdaBoost is a typical representative of the Boosting algorithm, AdaBoost algorithm in an iterative process that can adaptively change the distribution of the sample by weight control, making the classification of the center of gravity easy to focus on the sample misclassification, while by aggregating the base classifier of each iteration obtained , learn weighted voting strategy, based on the importance of the base classifiers to calculate weighted , combined to get the final classifier, AdaBoost algorithm processes shown in Table 1.

As can be seen from Table 1, the importance of the base classifiers C_i depends on its misclassification ratio, once the step 7 predicate p is true, $\delta(p)$ equal to the 1,otherwise 0, according to the formula of α_i in step 10, if the current misclassification ratio approaching to 0, the base classifier weights will be large positive, AdaBoost

Table 1 Process of the AdaBoost Algorithm

Step 1:Initialize N number of the sample's weights， $w=\{w_j=1/N \mid j=1,2,\cdots,N\}$
Step2:Set T number of iterative rounds
Step3:for $i=1:T$
Step4:Sample to sample set S without replacement based on w, generate training sample set S_i
Step5:Based on the selected weak classifier , train basic classifier C_i on S_i
Step6:Use basic classifier C_i to classify S
Step7:Calculate misclassification ratio $\varepsilon_i=\frac{1}{N}\left[\sum_{j=1}^{N} w_j \delta(C_i(x_j)\neq y_j)\right]$
Step8:if ε_i <0.5 then Update sample weights to initial values， That is $w=\{wj=1/N \mid j=1,2,\cdots,N\}$, return Step4
Step9:end
Step10:Calculate importance parameter of C_i $\alpha_i=\frac{1}{2}\ln\frac{1-\varepsilon_i}{\varepsilon_i}$
Step11:Update weights Sample S $w_j^{(i+1)}=\frac{w_j^{(i)}}{z_j}\times\begin{cases}e^{\alpha i}, C_i(x_j)\neq y_j\\ e^{-\alpha i}, C_i(x_j)=y_j\end{cases}$, Z_j make $\sum_{j=1}^{N} w_j^{(i+1)}=1$
Step12:end
Step13:End up with strong classifier $C(x)=\arg\max_y \sum_{i=1}^{T}\alpha_i\delta(C_i(x)=y)$

Manuscript received April 5, 2015
Manuscript revised April 20, 2015

algorithm make the predict value of C_i based classifier weighted in α_i , instead of using the majority voting program, this mechanism makes AdaBoost algorithm to encourage higher accuracy based classification model, when one of these iterative rounds of misclassification situation worse than random conjecture, then sample weights will be reset to the initial value.

1.2 Advantages and defects of algorithm

As soon as Adaboost algorithm was put forward, it attracted a lot of attention from many researchers because of its good performance in the classification of learning, it have achieved good results in application in many fields, such as the handwritten signature certification [2], human action recognition [3], information retrieval [4] and iris detection and its expansion in the relevant hardware [5].

Adaboost algorithm has good performance because of the algorithm itself exist unique advantages. First, the algorithm can improve the distribution of the sample by the mistake ratio of based classifiers adaptively to focus on hard sample. Second, the initial weak classifier selected threshold is low, only required its performance better than the random guess, to enhance usability of the algorithm, while weakening its calculation complexity. Thirdly, Adaboost has fast convergence. Fourthly, generalization of Adaboost will not stop optimizing or reverse with more based classified occurs, the algorithm constantly adjust their spacing to achieve the best generalization performance by adding based classifier. Finally, Adaboost changed the priority of Boosting algorithm, its applications easier to promote in real-world.

With further research , many scholars in the process of discussion Adaboost algorithm found that the algorithm is not nearly perfect , its shortcomings still exist : first, applied to non- equilibrium problem , Adaboost treat the misclassification of rare class and the misclassification of majority class equal, update the weight based on the overall error rate of the base classifiers , resulting in algorithms reduce concern to the error class of rare sample; second , robustness to noise sample is poor , if the noise present in the sample ,which is entitled to re- update mechanism will be focused on the training of noise , resulting in training convergence premature, resulting in classifier occurs the phenomenon of "over learning"; third ,iteration termination condition currently only rely on the initial set of iterations , In Adaboost algorithm ,although the generalization performance of classifiers can continue optimization in an iterative process, the contribution of final base classifiers generated is small, set too high number of iterations not only may happen "over learning" , but also will increase the calculated pressure.

2. Improved Adaboost algorithm based on unbalanced data

2.1 The classification of Adaboost algorithm in the non-equilibrium samples

The non- equilibrium problems in practical application is prevalence , under normal circumstances, the correctly identify of rare class has a higher value , such as fault identification, equipment failures are rare category , this is classified existed fault as in normal operation, may produce huge losses in production operations.

For the classification of non- equilibrium sample has the following characteristics: when two non-equilibrium sample occurs misclassification , the cost of both misclassification is quite different in most cases; the rare instance of a class exists less by itself , the formation of a classification model tends to specialize , more susceptible to the effects of noise; because of the sample itself is a kind off balance, the overall error rate of sample measure the merits of the final classification model is biased , it is easy to ignore misclassification of rare class.

General classification model with minimum overall error rate for the structure target , so the tendency for most classes of samples correctly classified in the category of non- equilibrium , leading to the detect of the traditional classification model to non-equilibrium instances is failure . Adaboost algorithm is not, although the importance of each base classifier is evaluated by the overall error , it focuses attention on the misuse of samples , which also includes the majority of misclassification sample and the rare of misclassification sample , making Adaboost have some adaptive to the classification of non- equilibrium instances [6] . However, this adaptation is only relative to adapt , Adaboost same deficiencies exist in the non- equilibrium categories : first, the equivalent of two non- treated sample misclassification balanced category , the second , the noise impact on the adjustment of the sample or difficulty sample distribution large noise samples in the continuous process of misclassification , the weight will continue to increase , so that the distribution of distortion of the sample occurs , resulting in a noise pattern of intensive training classifiers phenomenon . Below Adaboost algorithm combines these two shortcomings improvement.

2.2 Some existing improved methods

To the defect of the Adaboost algorithm in Unbalanced Data Classification, some existing improved Adaboost Algorithm mainly focused on two directions. Firstly, the improvement of the updating strategies of the sample weights. Secondly, the improvement in the process of training combining with other algorithms.

To the improvement of the updating strategies of the

sample weights, it focuses on adjusting the center of gravity of samples, focusing on the concerns of the rare class of samples. In 2002, Asymmetric AdaBoost algorithms are proposed based on this idea. When the samples of rare class and the samples of major class are both detected incorrectly, the strategies of the sample weights updating given by Asymmetric AdaBoost Algorithm tends to the misclassified samples [7] of rare class. However, the algorithm ignores the control of the noise samples.In 2008, Guo Qiaojin etc proposed the method of the noise samples suppression .They take four different strategies of the weights updating for the sample which exceeds the set threshold samples and achieved good results on noise suppression [8]. But it is lack of class discrimination for misclassified sample of non-noise .In 2011, Li Wenhui etc changed the emphasis distorted weight distribution by two types of samples which are based on the classification error, combining with the growth of controlled weights of misclassified frequency and reduced the rate of false alarm [9]. For the second direction of improvement , there are many improved methods have been gradually raised , such as Adaboost SVM algorithms proposed in 2008 and SVM algorithms replace the original algorithm and using SVM classification method to replace the original algorithm and to train for weak classification have improved the classification performance of Adaboost algorithm [10] .In 2011 , Chen Jintan etc, improved the method of Naive Bayesian and make it as weak classifier and achieved good results in the classification performance of rare class .However , the disadvantage is that the computational complexity of the algorithm is slightly increase [11] .

2.3 Weight updating based on consideration and misclassification error

For the analysis of the non-equilibrium data characteristics and Adaboost algorithm exists deficiencies in the non-equilibrium data classification application in Section 2.1, and combines the analysis of existing improved method in Section 2.2, this section will introduce cost-sensitive thinking of Adaboost algorithm is improved based on the sample misclassification costs and misclassification times to enhance the classification performance of algorithm to unbalanced data. Because of the consideration is different when different types of sample is misclassified, so this section will classified misclassification of training samples into three categories for discussion, in order to facilitate discussion, these three types of samples are called positive misclassified samples , negative misclassification of samples and noise samples . The positive samples are that rare category of misclassification sample mistaken for sample categories, the negative misclassification sample are that sample categories are misinterpreted as a rare kind of sample, noise sample is that hard classified sample or special samples were misclassified.

Adaboost algorithm after each round calculated , all samples will be re- adjusted according to the distribution of sample weights , this updated strategy is Adaboost make the training sample to maintain the core of self- adaptive , a new round of sample weights which is the original Adaboost algorithm is calculated according to the formula , such as formulas (1) , (2) and (3) .

$$w_j^{(i+1)} = \frac{w_j^{(i)}}{z_j} \times \begin{cases} e^{\alpha_i}, C_i(x_j) \neq y_j \\ e^{-\alpha_i}, C_i(x_j) = y_j \end{cases} \quad (1)$$

$$\alpha_i = \frac{1}{2} \ln \frac{1-\varepsilon_i}{\varepsilon_i} \quad (2)$$

$$\varepsilon_i = \frac{1}{N} \left[\sum_{j=1}^{N} w_j \delta(C_i(x_j) \neq y_j) \right] \quad (3)$$

$w_j^{(i)}$ is the weights for the sample j which is i round , $w_j^{(i+1)}$ is the weights for the next round of the sample .From the Sample weights calculated formula can be found that when the sample occurs misclassification, the weights will exponentially increase with the importance of the base classifier ,if the training results of the sample is correct, its weight decreases exponentially . This change in approach means three samples of positive misclassified samples , the negative samples and noise misclassification have the same misclassification costs , and with the increase times of training , the noise will get higher attention than the positive misclassification sample and the negative misclassification samples, the constantly expanding of cumulative noise weights will make the classifier produce excess trained , which is the root causes of poor noise robustness of Adaboost . In fact ,first, the positive misclassification sample should be subject to greater attention than negative misclassification samples ,and therefore in the process of adjustment sample distribution, it should be considered that the right of the positive misclassified samples have a large increase , while the weights of the negative misclassification sample should be increased smaller , and the second , the noise misclassification samples should be taken seriously , so once they are identified as noise misclassified samples , which should be a greater degree of weight reduction , making the classification is no longer concerned about the correctness of classification . Based on the above analysis, the value of new weights calculated by the following formula (4) and (5) to optimize:

$$w_j^{(i+1)} = \begin{cases} \frac{w_j^{(i)}}{z_j} \times \begin{cases} e^{\lambda \alpha_i} & , \quad C_i(x_j) \neq y_j \\ e^{-\alpha_i} & , \quad C_i(x_j) = y_j \end{cases} \\ \min(w^{(i)}) \quad , \quad C_i(x_j) \neq y_j \quad and \quad k \geq \mu \end{cases} \quad (4)$$

$$\lambda = \begin{cases} (1+c)/k \ , \ y_j = 1 \ \ and \ \ 0 < k < \mu \\ (1-c)/k \ , \ y_j = 0 \ \ and \ \ 0 < k < \mu \\ 1+c \ \ , \ y_j = 1 \ \ and \ \ k = 0 \\ 1-c \ \ , \ y_j = 0 \ \ and \ \ k = 0 \end{cases}, \ c \in [0,1] \quad (5)$$

λ is Weight control coefficient, λ is related to the three variable of c , k , μ . c is regulatory factor of misclassification costs of sample, k is the cumulative times of misclassification error of sample , μ is cumulative upper limit of misclassification error of sample .Update value k of samples, After each iteration of the algorithm .If samples were misclassified, the value of k plus 1 .When the cumulative times of misclassification error of sample k exceeds the value μ, the sample was identified misclassified noise samples

Formula (4) and (5) shows that the updating strategy of weights of misclassification sample of three types is discussed respectively. Updating value of weights of positive misclassified samples can greatly increase, if it is based on cost coefficient c and the cumulative times of misclassification error k and it controls the weight increase by the cumulative times of misclassification error k .Before the sample is identified as noise samples, minimize its impact. Because of the participation of c , the weight growth extent of negative misclassified sample reduce. Compared with the positive misclassification sample, the weights is increase small, and also control the weight through the misclassification cumulative k ; If the sample is determined as noise misclassification sample, the sample will receive a minimum weight of all the current sample weights rounds, this update strategy means that which is identified as the noise will be reduced the degree of being taken seriously to the minimum. The updating strategy of weights of sample based on cost and misclassification error makes improved Adaboost algorithm more focused on the misclassification of rare class sample and reduces the sensitivity of Adaboost algorithm to noise samples and improve the classification performance of disequilibrium dataset and does not affect the computational complexity of the Adaboost algorithm.

3. Experimental Consideration

3.1 Experimental environment and experimental data

In order to select SVM ,the experiments in this section use Matlab (R2010b) development environment and weak classification method , in which SVM algorithm uses LIBSVM package developed by LinZhiren of National Taiwan University, all the parameters of SVM during the experiment set to default values.

Experimental data are derived from the standard data set, the selected three data sets all have non-equilibrium nature, as to the CMC sample coming from UCI, which would need to be processed before the experiments are carried on, the original of CMC sample should be classified into multi-classification of samples, the sample classification number is 3,at first ,its unbalanced treatment should be carried on, the category label 2 is classified as a rare sample, making the other two combined, which is classified as the majority of the sample data category, the parameters about the experimental data would be shown in Table 2.

Table 2 describes the experimental data

Name	Property	Categories	Scale	Classification	Category
cmc	9	3	1473	333:1140	2:other
haberman	3	2	306	81:225	2:1
ionosphere	34	2	351	126:225	b:g

3.2 Experiments and results analysis

1. Experiment 1

Experiment 1 will focus on rare species in classification accuracy to measure the improved Adaboost algorithm to enhance the performance and using the recall rate, accuracy and measure F_1 to measure the performance of classification. The recall rate is the ratio of the number of samples of rare class of correct classification and the number of samples of all rare class .The higher the recall rate index is, the higher the proportion of correct identification of is; The accuracy is the ratio of the number of samples of rare class of correct classification and the number of samples of rare class judged by classification model .The higher the accuracy is , the lower the error rate of a classification model to predict the result of a rare class is The measure F_1 is the harmonic mean of the two indicators of recall rate and accuracy. The higher the value of F_1 is, it means that the overall accuracy of the classification model and the recall rate are both higher.

Firstly, the experiments evaluate on the classification performance of the selected weak classifier SVM and singly use the selected method of weak classifiers to classify the sample set .The overall misclassification error rate and the recall rate was evaluated, the results are shown in Table 3.

Table 3 The results of initial weak classifiers SVM

Sample set name	Accuracy	Recall rate
cmc	72.52%	3.76%
haberman	75.58%	4.76%
ionosphere	91.89%	82.61%

As we can see form Table 3, the training accuracy of the above data set is higher than a random guess, which meets the requirements of Adaboost algorithm to weak classifiers, so the method of weak classifiers which selects Adaboost is appropriate.

Because classification decision method inside SVM does not have disequilibrium, the effect of classification for rare samples is poor, and the reaction of the recall rate indicators is poor. The recall rate is only 60.87% in the contrast layer of sandstone sample set, but it is less than 5% in the sample sets of CMC and Hagerman.

In the experiment , SVM act as weak classifiers and the number of iterations T is set to 20 , the misclassification cumulative threshold μ is set to $T/4$, the cost coefficient c is set to 0.2、0.5 and 0.8 respectively. Use the original Adaboost algorithm and improved Adaboost algorithm respectively to do the assorted experiment. The parameter values of two algorithms is set the same. Use the three indicators of the recall rate, accuracy, and measure F_1 for evaluation. The classification results are shown in the following table 4.

Table 4 Comparison of the original algorithm

Name	Algorithm	Cost	Recall rate	Accuracy	Measure
cmc	Adaboost		36.84%	52.69%	43.36%
	improved algorithm	0.2	48.87%	53.72%	51.18%
		0.5	70.68%	43.12%	53.56%
		0.8	76.69%	40.80%	53.26%
haber man	Adaboost		42.86%	40.91%	41.86%
	improved algorithm	0.2	61.90%	54.17%	57.78%
		0.5	66.67%	56.00%	60.87%
		0.8	80.95%	54.84%	65.38%
ionos phere	Adaboost		91.30%	93.33%	92.13%
	improved algorithm	0.2	91.30%	97.67%	94.38%
		0.5	93.48%	91.49%	92.47%
		0.8	95.56%	91.67%	93.62%
	improved algorithm	0.2	73.91%	56.67%	64.15%
		0.5	82.61%	55.88%	66.67%
		0.8	91.30%	45.65%	60.87%

As we can see from the experimental data in Table 4 , whether the evaluation criteria set provided by UCI or the generated sample data of sandstone contrast layer in this paper, using an improved method of updating the weight , in case of three kinds of different coefficients consideration , four data recall rate of concentrated rare class has improved a certain level , when the consideration factor is set on 0.8 , CMC recall rate target the enhance rate to more than doubled , F1 measure results than the original Adaboost algorithm is also improved. From the data in the table, we can also see that the extent of enhance of rare species identification is related to cost coefficient at the same time . From the analysis of the data, the recall rate of the four dataset showed an increasing relationship with cost coefficient .When the recall rates were highest value, the cost coefficient is 0.8 .According to the analysis of the chapter 2.3, because the setpoint of cost coefficient is larger, positive misclassified samples get greater attention. Known that the higher the the setpoint of cost coefficient is, the larger the recall rate is at the same condition. From the analysis of the data in the table, when the recall rate increases ,the accuracy of rare class classification slightly decreases at the same time .That is because we concern the accuracy of rare class classification while sacrificing the accuracy of the major class classification ,but the accuracy of the floating range is acceptable from the data. As a result, the cost coefficient can be set according to the actual situation of the data and the setpoint of cost coefficient can not be too high, because too high setpoint may lead to the decline of classification accuracy. Form the analysis of F_1 which is the harmonic mean of the two indicators of recall rate and accuracy, the value F_1 of rare class of the improved algorithm is all higher than the original algorithm' and the value F_1 has not necessary relationship of ascending or descending .

For the analysis of these three indicators, the improved Adaboost algorithm proposed in this section deal with classification problem of unbalanced data, especially for the identification of rare class and played a good effect.

2. Experiment 2

Compared improved Adaboost algorithm in this article, and other improvements Adaboost algorithm in experiment 2, and it mainly compared an improved method proposed in the literature [8] with Asymmetric AdaBoost algorithm by the use of the recall rate, accuracy, and measure F_1 three indicators for evaluation.

Asymmetric AdaBoost algorithm and the results of improved methods in the literature [8] adopt the experimental results in the literature [8] , There are four improved algorithm in the literature [8], this section adopt the improved method of D, In this article, the improved algorithm of experimental parameters setting by the weak classifier and experiment 1 use the same parameters .the number of iterations is 10 times, misclassification cumulative threshold μ is set to $T/3$,cost coefficient c is set to 0.5. Comparative results are shown in Table 5.

As we can see form Table 5, the F_1 measurement results of this improved Adaboost algorithm is higher than the Asymmetric AdaBoost algorithm as well as improved method D of the literature [8], reflecting the improved algorithm of this paper is useful for the overall level improvement of the recall rate and precision. For the recall rate and precision of these two individual indicators,

they Both have promotion in the CMC data set by the use of the improved algorithm in the paper , for Hagerman data sets , the improvement of the precision is more obvious, the recall rate fell slightly from the other two algorithms, experimental results of ionosphere data set is just on the contrast, the recall rate improve obviously than other two algorithms, the precision fell compared with the improvement method in literature [8] , but the recall rate is higher than Asymmetric AdaBoost algorithm. Overall, in this paper the improvement of Adaboost algorithm is effective and has certain advantages.

Table 5 Comparison of other improved algorithm

Name	Algorithm	Recall rate	Accuracy	Measure
cmc	AsymBoost	69.37%	35.16%	46.67%
	Improved methods	62.76%	36.73%	46.34%
	Paper	69.92%	48.19%	57.06%
haber man	AsymBoost	76.54%	34.64%	47.69%
	Improved methods	71.60%	36.71%	48.54%
	Paper	66.67%	53.58%	59.57%
ionos phere	AsymBoost	89.68%	79.58%	84.33%
	Improved methods	89.68%	94.96%	92.24%
	Paper	93.48%	91.49%	92.47%

4. Conclusion

In this paper, we analysis the Adaboost algorithm , describes the advantages and defects of algorithm ,because of Adaboost tend to concern misclassified samples in the classification process , making it have more ability to adapt to the non-equilibrium classification problems than traditional classifiers . Yet for the strategy that all misclassification samples without distinction look makes Adaboost algorithm have some problems in the application process, including the lack of interest in the rare class of non-equilibrium sample classification process, sensitive to noise and so on. For the above analysis, this paper proposes a value update method based on cost and misclassification, the Adaboost method can be improved by the important degree of these three misclassification samples which are positive misclassification, negative misclassification and noise misclassification of in classification, treated differently in the process of updating the weights so that the effect of the classifier can focus on the classification of rare class. Experimental results show that the improved Adaboost method has a good effect in improving the performance aspects of the classification of rare class.

References

[1] Schapire R E. The strength of weak learnability [J]. Machine learning, 1990, 5(2): 197-227.

[2] Hu J, Chen Y. Writer-independent off-line handwritten signature verification based on real adaboost[C]//Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC), 2011 2nd International Conference on. IEEE, 2011: 6095-6098.

[3] Yan X, Luo Y. Recognizing human actions using a new descriptor based on spatial–temporal interest points and weighted-output classifier [J]. Neurocomputing, 2012, 87: 51-61.

[4] Wu X, Ngo C W, Zhu Y M, et al. Boosting web video categorization with contextual information from social web World Wide Web, 2012, 15(2): 197-212.

[5] Wang Q, Zhang X, Li M, et al. Adaboost and multi-orientation 2D Gabor-based noisy iris recognition[J]. Pattern Recognition Letters, 2012, 33(8): 978-983.

[6] Landesa-Vázquez I, Alba-Castro J L. Shedding light on the asymmetric learning capability of AdaBoost [J]. Pattern Recognition Letters, 2012, 33(3): 247-255.

[7] Viola P, Jones M. Fast and robust classification using asymmetric adaboost and a detector cascade [J]. Advances in Neural Information Processing Systems, 2002, 2: 1311-1318.

[8] Gao Qiaojin, Li Libin, Li Ning.Novel modified AdaBoost algorithm for imbalanced data classification .Computer Engineering and Applications, 2008, 44(21):217- 221.

[9] Li Wenhui, Ni Hongyin. An Improved Adaboost Training Algorithm[J]. Journal of Jilin University(Science Edition), 2011,49(3):498-504.

[10] Li X, Wang L, Sung E. AdaBoost with SVM-based component classifiers [J]. Engineering Applications of Artificial Intelligence, 2008, 21(5): 785-795.

[11] Chen Jintan, Kang Hengzheng, Yang Yan, Zhou Weixiong. A Classification method for Class-unbalanced data[J]. Journal of Shandong University(Engineering Science),2011,41(1):96-101.