An Improved AdaBoost-SVM Model Based on Sample Weights and Sampling Equilibrium

Hongchen Guo[1, a],Junbang Ma[2,b] and Zhiqiang Li[2,c]

[1]Network Information Technology Centre, Beijing Institute of Technology, Beijing, 100081, China

[2] School of Software, Beijing Institute of Technology, Beijing 100081, China

[a]guohongchen@bit.edu.cn, [b]2803650957@qq.com, [c]lizq@bit.edu.cn (Corresponding Author)

Keywords: AdaBoost,SVM, Sample Weights, Sampling Equilibrium

Abstract. The existing model which combines AdaBoost and SVM has poor performance when dealing with the imbalance dataset in multi-label classification. To deal with this problem, we proposed a new model SAB-WSVM. In our model, we modified AdaBoost original sampling methods in order to make it more balanced and more informative. Also we combined the sample weights of SVM and weights of AdaBoost to make SVM pay more attention to the samples which are difficult to be classified.In the experiment, we test it with two datasets. The results show that our model has better performance in the unbalanced multi-label datasets.

Introduction

Support vector machines (SVM) [1] are supervised learning models for classification. And AdaBoost [2] is a classical ensemble method. We can use AdaBoost to combine with SVM so that we can obtain better classification effect.Most of the existing binding models are applied to single-label classification. When they are applied directly to multi-label classification, their performance is not very good, especially for unbalanced multi-label data sets. Therefore, in this paper, we propose a new modelSAB-WSVM. In our model, in order to deal with the imbalanced problem, we modify the sampling methods of AdaBoost. We combine random sampling withsequential sampling and add the distance from the samples to the hyperplane as the new ranking criteria. In this way, we can make the training sets more balanced and more informative. Also, we use the sample weights in AdaBoost iterative process as the sample weights in SVM. So that, we can make SVM classify the samples correctly as far as possible. The results prove that our model is valid.

Related Works

AdaBoost is an ensemble method, which combines multiple weak classifiers to form a final strong classifier. But SVM is a strong classifier. So it is not suitable to directly use SVM as the base classifier for AdaBoost.So how to weaken the SVM and in the meantime improve the efficiency of the final classifier has become the focus of previous studies.In the aspect of weakening SVM, there are two kinds of strategies, one is to weaken the SVM by adjusting the parameters of SVM kernel function, and the other is to reduce the number of training samples in order to weaken the SVM.

The most common strategy ofadjusting the parameters of SVM kernel function is to use RBFSVM, which has C andσ two parameters. Once we roughly determine the value of C parameter, the performance of RBFSVM largely depends on the value of theσparameter. So, we can use the σparameter to weaken SVM. For example, Hu Jinhai[3] proposed an AdaBoost-SVM model. It weakens the SVM by adjusting the σparameters in each iteration, and uses the weakSVM as the base classifier. Then it combines these weak classifiers to form its final strong classifier.On the basis of this AdaBoost-SVMmodel, Chang Tiantian [4] proposed a LDAB-SVM, which obtains the more diversified SVM as the base classifier to improve the efficiency of the final classifier.

As to reduce the number of training samples, the main idea is to extract some samples from the original data set to form the new training set.On the one hand, because of the less training samples,

we can weaken the classifier. On the other hand, we can accelerate the training speed. For example, Pavlov [5] improved the training speed of SVM by reducing the number of samples and combining SVM with AdaBoost. Elkin García [6] used the less training samples to weaken SVM to combine with AdaBoost.L Diao [7] used active learning to filter the samples and improved the final efficiency of classification.Efraín [8] proposed a model RAB-SLPSVM which uses a new sampling method to make the training set more informative.

Most of these models are applied to single-label classification. They can't do well in the unbalanced multi-label datasets. So next, we will introduce our model and what we do to deal with this problem.

Multi-label classification model SAB-WSVM

Our model SAB-WSVM is based on the AdaBoost-SVM modelwhich is proposed by Hu Jinhai. First, we use the Binary Relevance [9] (BR) algorithm to transform the multi-label problem into a set of binary classification problems. Then we use our model toclassify. We find that the performance of the AdaBoost-SVM model is not very good for the unbalanced data set. So in our model, in order to deal with this problem, we change the sampling methodof AdaBoost and use the sample weighted SVM as the base classifier for AdaBoost.

Improved Sampling Method. According to [10], There are three commonly used strategies when using emphasis criterion in sampling: Trimming, Unique Uniform Sampling (UUS) and WeightedSampling (WS).In AdaBoost,the sampling is the first one. Itextracts the sampleswhich have larger weights.This makes the algorithm pay more attention to the samples which aredifficult to be classified and improvesthe final classification effect.But for the unbalanced datasets, because the number of positive samples is less, using this sampling method will lead classification hyperplane to tilt and can't reach very good effect of classification.

In order to overcome this shortcoming of AdaBoost sampling method in multi-label classification, we modify the sampling method in RAB-SLPSVM [8] algorithm and use this new method in our model. L is the number of the samples to be selected. In each round of the training process, the samples are sorted according to their weights and the distance from the samples to the classification hyperplane, where the sample weightsare the primary ranking criteria and the distance is the secondary. Then we divide the samples into L groups and randomly select a sample from each group to form the final training data set. In the SVM, the sample which is closer to the hyperplane is more difficult to be classified. By using this, in each round of iterativeprocess, training set both has the sampleswhich have bigger weights and closer distance and the samples which have smaller weights and longer distance. It can insure that the training set has sufficient information anddiversity between the samples in the set. By using this sampling method, we canmake the hyperplanemore balanced and effectively improve the performance of our model. As for the calculating the distance, we can simply use the value of $\|w^T x_i+b\|$to approximate the distance from the samples to the hyperplane.

Modifying the sample weighted SVM. In order to improve the performance of our model, we use the sample weighted SVM as the base classifier. In the sample weighted SVM, the optimization problem is Eq. 1.

$$\min \frac{1}{2}\|w\|^2+C\sum_{i=1}^{n} W_i\xi_i \tag{1}$$

$W_i$is the weights of the sampleswhich areselected as the slack variable. For the sample weighted SVM, the weights affect the likelihood that the samples will be selected as a slack variable. The bigger the weightof a sample, the less it will be selected as the slack variable, the more it should be classified correctly. In this case, we can combine the sample weights in AdaBoost with the sample weights in SVM. In AdaBoost iterative process, the samples with larger weights are often difficult to be classified. Therefore, we can use these weights in SVM so that SVM can pay more attention to these samples to classify them correctly. In this way, we can combine the SVM and AdaBoost more

closely and improve the final classification results.

In practical application, we find that the weights of AdaBoost can't be used directly in SVM.Because the sample weights of AdaBoost is generally initially set to 1 / N, where N is the number of samples. Each sample will have a small weightifthe number of samples is large. The weightmaybe is only 0.1% or 0.01%. If we directly use the weights in SVM, the slack variable will be very small andwill have no effect to the classification. Therefore, wemodify the sample weighted SVM.Its new optimization problem isEq. 2

$$\min \frac{1}{2}\|w\|^2+C\sum_{i=1}^{n} W_i Q_i \xi_i \quad (2)$$

$Q_i$is the weighting coefficient. Through this amplification factor, on the one hand, the sample weights of AdaBoost can be appropriately enlarged so that the weights of AdaBoost can be used in the sample weights of SVM. And on the other hand, the weights of different samples can be adjusted by this coefficient. As to determine the $Q_i$coefficient, it can be roughly setto the number of samples N.

Iterative process of our model. With this two method, we proposed our new model SAB-WSVM. The iterative process of it is as below.

Algorithm 1 SAB-WSVM

1. Input: a set of training samples with labels$\{(x_1,y_1),\ldots,(x_m,y_m)\}$,the initial kernel parameter σ_{ini},the minimal kernel parameter σ_{min},the step of kernel parameterσ_{step}.
2. Initialization: the weights of training samples $W_i=\frac{1}{m}$, m is the total number of training samples
3. While($\sigma>\sigma_{min}$)
   a) Use the new sampling method to extract the samples.
   b) Train a sample weighted classifier h_i on the new training sets.
   c) Calculate the training error of h_i :$\varepsilon_i=\sum_{i=1}^{n} W_i,\ y_i\neq h_i(x_i)$.
   d) If $\varepsilon_i > 0.5$, decrease σ_{ini}byσ_{step}, go to (a); else go to (d).
   e) Set the weights of component classifier h_i :$\alpha_i=\frac{1}{2}\ln(\frac{1-\varepsilon_i}{\varepsilon_i})$
   f) Update the weights of training samples :$W_{i+1}=\frac{W_i\exp(-\alpha_i y_i h_i(x_i))}{Z_i}$, where Z_i is a normalization constant and $\sum_i^n W_{i+1}=1$
   g) Update the distance from samples to hyperplane for each sample
4. Output: $H(x)=\text{sign}(\sum_{i=1}^{T}\alpha_i h_i(x_i))$

Experiments

Dataset. In order to show the performance of our model in the unbalanced datasets, we chose the scene data set and BITWebCorp. The scene data is provided by libsvm, which is divided into six categories, including 1211 training samples and 1196 test samples. TheBITWebCorpis aChineselanguagecorporacollectedbyus. We crawled the text contents according to the web log information in our school and used ICTCLASforsegmentation. The data set isdivided into 10 categories, including2956trainingsamplesand 1267 test samples.These datasets are imbalanced. The proportion of positive and negative samplesin some categories of the scene dataset can be 1: 5, and in some categories of BITWebCorp theproportion can be 1:50. So it is very suitable to use these datasets to test the performance of our model in the unbalanced multi-label data sets.

Results and Discussion. In the experiment, we mainly comparedour model with the SVM algorithm and the AdaBoost-SVM model. There are two parameters to be determined. One is the C parameter of RBFSVM and the other is the number of training samples in AdaBoost iteration process.Through some experiments, we set the C parameter to 1.0 and sampling ratio to 30%. In terms of evaluation criteria, we used the macro-accuracy, macro-precision, macro-recall, macro-f1 and Hamming losses.

The experimental results of scene dataset and BITWebCorpare shown in Figure 1.In the five evaluation indicators, most indicators of SAB-WSVM are better than the other two algorithms, expect theprecision.In SVM and AdaBoost-SVM,due to the imbalance of the positive and negative samples, the class hyperplane is inclined to the positive samples. So that a large number of positive samples are misclassified into negative classes and the classifiers perform poorly on these four metrics. And our model makes the class hyperplane more balance,so that we can have better performance of our classifier. As to macro-precision, in unbalanced datasets, most of the negative samples are classified correctly due to the inclined hyperplane. So the macro-precision of SVM and AdaBoost-SVM is particularly high. In our model, we make hyperplanemore balanced.It inevitably leads to an increase in the number of negative-class misclassified samples and leads to a decrease in macro-precision.

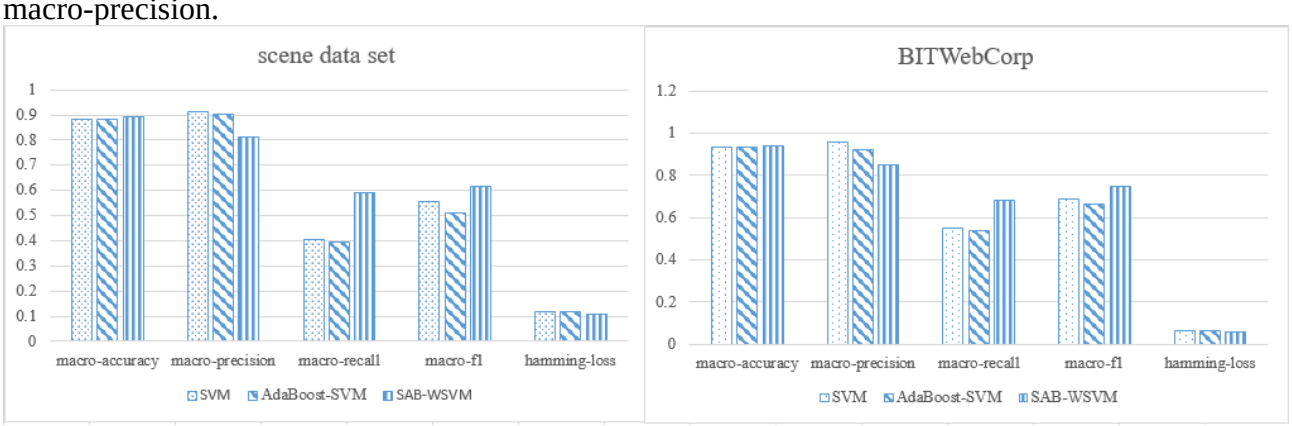

Fig. 1.experimental result of the data sets

In order to further analyze the impact of unbalanced data sets in the experiment, several representative categories were selected for analysis from the scene data set. The results of the scene data set experiment are shown in Table 1. For SVM and AdaBoost-SVM, we can see that the number of negative-class misclassified samples is often very small, even 0, and the number of positive-class misclassified samples is more than half of the total positive samples, even more than 90%.Their precision is often every high.The recall and F1 value are generally low. For the SAB-WSVM, we can see that the number of positive-class misclassified samples is reduced and the number of negative-class misclassified samples is inevitably increased. But Generally speaking,our model has better performance than SVM and AdaBoost-SVM in unbalanced data sets.

Table 1partial experiment results of scene data set

category	classifier	positivesample num	negativesamplenum	positive sampleerr num	negative sampleerr num	precision	recall	f1
0	SVM	200	996	116	14	0.86	0.42	0.56
	AdaBoost-SVM	200	996	127	6	0.92	0.36	0.52
	SAB-WSVM	200	996	84	30	0.79	*0.58*	*0.67*
2	SVM	200	996	104	0	1	0.48	0.65
	AdaBoost-SVM	200	996	86	1	0.99	0.57	0.72
	SAB-WSVM	200	996	73	14	0.90	*0.63*	*0.74*

3	SVM	237	959	113	4	0.97	0.52	0.68
	AdaBoost-SVM	237	959	208	0	0.96	0.62	0.74
	SAB-WSVM	237	959	90	7	0.95	0.62	0.75
4	SVM	256	940	176	25	0.76	0.31	0.44
	AdaBoost-SVM	256	940	248	1	0.88	0.03	0.06
	SAB-WSVM	256	940	140	65	0.64	0.45	0.53

Conclusion

In this paper, we propose a new combinatorial model SAB-WSAM, which is based on the AdaBoost-SVM model proposed by Hu Jinhai. We modified the sampling method of AdaBoost and combined the sample weighted SVM with AdaBoost to improve the performance of the combined model in the unbalanced datasets.

In the experiment, we use scene dataset and BITWebCorp to test our model.The experiment shows that, compared with SVM and AdaBoost-SVM, our model has been improved in most indicators for unbalanced multi-label data sets. In the future, we will test our model in other data sets in order to further improve the performance of our model in unbalanced multi-label data sets.

References

[1] Cortes, C.; Vapnik, V. (1995). "Support-vector networks". Machine Learning. 20 (3): 273–297.

[2] Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences 55 (1997) 119–139

[3] Hu J, Xie S, Yang F, et al. Ensemble of classification methods based on SVM and its application in diagnosis[J]. JOURNAL OF PROPULSION TECHNOLOGY-BEIJING-, 2007, 28(6): 669.

[4] Tiantian C, Hongwei L, Shuisheng Z. Large scale classification with local diversity AdaBoost SVM algorithm[J]. Journal of Systems engineering and electronics, 2009, 20(6): 1344-1350.

[5]Pavlov D, Mao J, Dom B. Scaling-up support vector machines using boosting algorithm[C]//Pattern Recognition, 2000. Proceedings. 15th International Conference on. IEEE, 2000, 2: 219-222.

[6] García E, Lozano F. Boosting Support Vector Machines[C]//MLDM Posters. 2007: 153-167.

[7] Diao L, Hu K, Lu Y, et al. A method to boost support vector machines[M]//Advances in Knowledge Discovery and Data Mining. Springer Berlin Heidelberg, 2002: 463-468.

[8] Mayhua-López E, Gómez-Verdejo V, Figueiras-Vidal A R. A new boosting design of Support Vector Machine classifiers[J]. Information Fusion, 2015, 25: 63-71.

[9]Boutell M R, Luo J, Shen X, et al. Learning multi-label scene classification[J]. Pattern recognition, 2004, 37(9): 1757-1771.

[10] Kalal Z, Matas J, Mikolajczyk K. Weighted Sampling for Large-Scale Boosting[C]//BMVC. 2008: 1-10.