International Journal of Database Theory and Application
Vol.7, No.5 (2014), pp.85-94
http://dx.doi.org/10.14257/ijdta.2014.7.5.06

Imbalanced Data Classification Based on AdaBoost-SVM

Li Peng[1,2], Bi Ting-ting[2], Yu Xiao-yang[1]and Li Si-ben[2]

[1] *Higher Educational Key Laboratory for Measuring and Control Technology, Instrumentations of Heilongjiang Province, Harbin University of Science and Technology, 150080 Harbin, China*
[2] *School of Computer Science and Technology, Harbin University of Science and Technology, 150080 Harbin, China*
{pli, yuxiaoyang }@hrbust.edu.cn.

Abstract

The classification of imbalanced data is one of the most challenging problems in data mining and machine learning research. Imbalanced dataset is a form that exists in reality area, which describes truly and objectively the essential characters of something. There will appear paucity of data and flooded in the classification of imbalanced dataset. Beside problems such as loss of information and data splitting phenomenon will also appear when using the traditional machine learning methods. So how to solve the classification problem of imbalanced data will be challenging. In this paper, aiming at the above problems, a classification algorithm based on AdaBoost-SVM is proposed. In the experiments with four typical forms of imbalanced data sets in UCI were validated the effectiveness of this strategy.

Keywords: *imbalanced data; classification; Adaboost; SVM*

1. Introduction

Along with the rapid development of computer network technology and in-depth application, all kinds of organizations can be obtained mass data easily. So analysis and obtain the helpful implicit information among the mass data have been deduced by employing data mining technology. The so-called imbalance dataset is the out-of-balance distribution of different category, the ratio of imbalance is even order of magnitude [1]. Although the imbalanced data reflects the objective things in world, the fact people tend to be more concerned about the situation described by small category data. Therefore, research on classification problem of imbalanced dataset has important theoretical and practical value. Experts and scholars from home and abroad have evolved this research for many years, they concluded that the method based on sample become the most effective classifier model [2]. At present the existing models have well solved the classification of small-scale data, the more complete data and the category is distribution uniformity relatively. But the classification problem is an ongoing issue, the classification technique still face many challenges such as data aliasing, the classification of mass data *etc* [3].

There have been many experts and scholars work on problems of imbalanced dataset. In 2000, topic of Workshop on learning Imbalanced Data Sets is first proposed on Artificial Intelligence International Conference, this conference make clear the concept

ISSN: 2005-4270 IJDTA

of imbalanced dataset , discuss the evaluation methods and the relationship between imbalanced data and cost-sensitive learning [4]**.** American scholar Chawla made research on the number and type of samples, provide the theoretical support for make certain of selecting samples reasonable [5]. In the light of imbalanced data, scholar Stamatatos apply the method of text sample to the filed of identity recognition [6]. In 2010, scholar Mike made deep exploration on feature selection about imbalanced data in his paper [7]. Alert analyzed the relationship between the overall scale of imbalanced data set and parameter setting of sorting algorithm [8]. Then Rukshan proposed FSVM-CIL algorithm based on fuzzy support vector machine and apply this theory into imbalanced data set learning and the field of bioinformatics' classification [9].

The investigation shows the effect of address such problem based on traditional classification algorithm is not effect [10]. The problems such as big different among individuals of data, loss of label and imbalanced distribution of categories will result in low classification accuracy. At present generally adopting two strategies: one is resample, up sampling and down sampling, which can reduce the amount of big category or decrease misclassification cost of small category. Another one is to explore more suitable classifier model to improve the classification ability [11]. In this paper view of the above question, propose a kind algorithm based on AdaBoost–SVM at the algorithm level to improve the accuracy rate of classification.

2. Methodology

2.1. Imbalanced Data Classification

There may be two kinds of imbalances in a data set. One is between-class imbalanced, in which case some classes have much more examples than others, the other is within-class imbalanced, in which case some subsets of the same class. In imbalanced data sets, we call the classes having more examples the majority class and the ones having fewer examples the minority class. To describe the samples' data form with the help of decision information tables. A decision table $S =< U, R, V, f >$, here, U is a set of objects, $R = C \cup D$ is a set of attribute, subset C and D are condition attribute set and decision attribute set, $D \neq \varnothing$. $f : U \times R \to V$ is information function, this appoint the attribute value of every object x ,which $U = \{x_1, x_2 \cdots, x_n\}$ describe the set of samples, when n is relatively bigger show that the large scale of sample is. The set of condition attribute $C = \{c_1, c_2 \cdots, c_m\}$ describe intrinsic attributes, the value of m is relatively bigger show that the date set is high-dimensional. $D = \{d_1, d_2 \cdots, d_l\}$ are decision attribute set, if samples number of categories have huge difference we call these imbalanced data.

Due to the special distribution of different category samples, the traditional methods could have been solved these problems appears to be inadequate, even the classification results by some methods are not be accepted. So we need improve the existing classification strategy in order to solve the imbalance data classification problem. We have done this in two ways: processing data set and improving algorithm [12].

1. Processing Data Set

A common way is using resample technique [13]. This method changes data set distribution in space not involving existing classification algorithms. The main purpose is to reduce the imbalance ratio. The basic rule of sampling is avoiding loss of information with the maximum as the same time reducing imbalanced rates. This is a paradox question that to decrease the number of negative samples and want to remain the amount of information.

On the strategy of sampling, it can be divides into simple randomization sampling and heuristic sampling. Randomization sampling is not using the characteristics of elements in dataset, just add or delete some samples randomly. While heuristic sampling make the best of relevant information and guide the re-sampling of dataset. On the technology of sampling, it can be divides into up-sampling and down-sapling. Up-sampling is increasing the number of positive samples that change distribution of dataset. On the contrary down-sampling is deleting the number of negative samples.

2. Improving Algorithm

By far the most important means of improving algorithm is cost-sensitive learning and ensemble learning [14]. The mis-classification signification of majority class and minority class means different because of the characteristic of imbalanced dataset distribution. While cost-sensitive learning is a kind strategy of establishing different mis-classification cost aims at different categories data elements. And ensemble learning overcome classification bottleneck by setting up multiple classifiers and integrating internal mechanism, so as to solve the classification problem of imbalanced data set.

2.2. Classification Algorithm

A. AdaBoost

There are many kinds of ensemble learning algorithms. AdaBoost is one of them adopted by experts and scholars at home and abroad generally. Learning ability of a single classifier may unsatisfactory. So consider to get a stronger classifier by weighting multiple weak classifiers.

For two-category problem, if input n training samples: $\{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$, which x_i is training sample, $y_i \in \{-1, +1\}$ represent positive and example respectively, among them l is the number of positive example and m is number of negative , $n = l + m$.

The specific steps are as follows:

1) Initializing weight of every samples $w_l(i)$, $i \in D(i)$;

2) For every sample, $t = 1, 2, \cdots, T$ (T is number of weak classifiers)

①Initializing weight to a probability distribution

$$w_{t,i} = \frac{w_{t,i}}{\sum_{j=1}^{n} w_{t,j}} \tag{1}$$

②Train the weak classifier h_j for every feature f , calculate weight error rate of every weak classifiers:

$$\varepsilon_j = \sum_{i=1}^{n} w_t(x_i) \left| h_j(x_i) \neq y_i \right| \tag{2}$$

③Select the best weak classifier h_t (the minimum error rate) : ε_t

④Adjust weight in accordance with this classifier:

$$w_{t+1,i} = w_{t,i}\beta_t^{1-\varepsilon_t} \tag{3}$$

Among them $\varepsilon_i = 0$ represent is classified correctly, in contrast $\varepsilon_i = 1$ is classified falsely:

$$\beta_t = \frac{\varepsilon_t}{1-\varepsilon_t} \tag{4}$$

3) At last the stronger classifier is:

$$h(x) = \begin{cases} +1 & \sum_{t=1}^{r} \alpha_t h_t(x) \geq \frac{1}{2}\sum_{t=1}^{r} X_i, \quad \alpha_t = \log\frac{1}{\beta_t} \\ -1 & otherwise \end{cases} \tag{5}$$

B. Support Vector Machines for Classification

Support vector machine (SVM) was originally developed by Boser and Vapnik, which is based on the Vapnik-Chervonekis (VC) theory and structural risk minimization (SRM) principle, by trying to find the trade-off between minimizing the training set error and maximizing the margin, in order to achieve the best generalization ability and remains resistant to over fitting. In addition, one major advantage of SVM is the using of convex quadratic programming, which provides only global minima hence avoid being trapped in local minima. In this section we will be concentrated on the basic SVM concepts for typical binary-classification problems. $L = \{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$ is training set, here $x_i \in R^d$, $y_i \in \{1, -1\}, i = 1, 2, \cdots, n$, y_i is class label of x_i, n is the number of training sample, exists a hyperplane can be described as follow:

$$w * x + b = 0. \tag{6}$$

Among them w is the normal vector of hyperplane, and b is the offset of hyperplane. The minimum objective function is：

$$\min \phi(w) = \frac{1}{2}\| w \|^2 + C(\sum_{i=1}^{n} \xi_i). \tag{7}$$

Meets the condition:

$$y_i(w \cdot x_i + b) \geq 1 - \xi_i; i = 1, 2, \cdots, n. \tag{8}$$

Here, $1/2 \| w \|^2$ represent the complexity of structure; $C(\sum_{i=1}^{n} \xi_i)$ is empirical risk; ξ_i is slack variable; H is a constant which is punishment factor of samples wrongly classified. For the situation of linear inseparable the main idea of SVM is used to map the feature vector to the high dimensional feature space and constructs an optimal hyperplane in the feature space. To get the change of ϕ, x in space of R^n mapped into H :

$$x \to \phi(x) = (\phi_1(x), \phi_2(x), \cdots, \phi_i(x))^T. \tag{9}$$

Eventually can decide optimization classification function:

$$f(x) = \operatorname{sgn}(w \cdot \phi(x) + b) = \operatorname{sgn}\left(\sum_{i=1}^{n} a_i y_i \phi(x_i) \cdot \phi(x) + b\right) \tag{10}$$

C. AdaBoost-SVM

The basic idea of boosting algorithm is to produce several weak classifiers which are slightly better than random guessing, then incorporate them into an estimate with high accuracy. The advantages of Adaboost are as following: (1) Its training process focuses on the data difficult to be classified (2) weighted voting is adopted during the weak classifiers integration instead of average voting mechanism. (3)Features are selected contingent on the features.

In this paper, we use SVM as the base classifier in the framework of Adaboost. There are some reasons: (1) SVM can deal with problems, such as small size of samples, nonlinearity or high dimensions. (2) There exists mature and convenient software package of SVM, such an LIBSVM.Support $X(j)=\{x_i^{(j)}\},(i=1,2,\cdots,N,\ j=1,2\cdots,n)$ is the features bank for one sample, N is the total dimension of features, $y(i)\in\{1,-1\}(j=1,2,\cdots,n)$ is the label of each samples, T is the appointed feature number which is set beforehand, and n is the total number of training samples. Then the Adaboost-SVM process can be described as following:

1) Initialize weight for each sample:

$$D_i = 1/n(i=1,2,\cdots,n) \tag{11}$$

2) For each feature $x_i(i=1,2,\cdots,N)$, train a SVM classifier $h(x_i)\in\{-1,+1\}$ which is restricted to using a single feature. Then N SVM models are got.

3) For $t=1,\cdots,T$:

For each $h(x_i)$, calculate the weighted error:

$$\varepsilon_t^{(i)} = \sum_{j=1}^{n} D_i \,|\, h(X_j) - y(j) \,|\, \tag{12}$$

Choose a feature with the lowest weighted error rate ε_j and save its corresponding SVM model. Calculate the selected weak classifier's weight

$$a_t = \frac{1}{2}\ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right) \tag{13}$$

Update sample weights according to a_t :

$$D_{t+1}(i) = \frac{D_t(i)}{Z_t}\times F(E_t) = \frac{D_t(i)}{Z_t}\times\begin{cases} e^{-\alpha_t} & if \quad h_t(x_i)=y_i \\ e^{\alpha_t} & if \quad h_t(x_i)\neq y_i \end{cases} \tag{14}$$

And the normalized parameters:

$$\sum_{i=1}^{n} D_t(i) = 1 \tag{15}$$

4) Use strong classifier H integrated by SVM weak classifier to training set.

$$H(x) = sign[\sum_{i=1}^{n} a_t h_t(x)] \tag{16}$$

3. The Results and analysis of experiment

3.1. Evaluation Index of Experiment

In order to evaluate the effectiveness of our method filtering spammers, we utilize ROC(receiver operating characteristic curve) and AUC （receiver operating characteristic curve） to evaluate the effectiveness of our method.

$$logit(p) = \log \frac{p}{1-p} \tag{17}$$

$$logit^{-1} = \frac{e^{x}}{1+e^{x}} \tag{18}$$

$$fpr = \frac{|FP| + 0.5}{|FP| + |TN| + 1} \tag{19}$$

$$fnr = \frac{|FN| + 0.5}{|FN| + |TP| + 1} \tag{20}$$

$$AUC = \int_0^1 \frac{TP}{P} d\frac{FP}{N} = \frac{1}{P \cdot N} \int_0^N TP\, dFP \tag{21}$$

Here, *FP* (False Positive) represents the number of relevant pairs misjudged by workers; *TN* (True Negative) represents the number of irrelevant pairs judged by workers; *FN* (False Negative) represents the number of irrelevant pairs misjudged by workers; *TP* (True Positive) represents the number of relevant pairs judged by workers. So the number of virtual positive $P = TP + FN$, while the number of negative $N = TN + FP$.

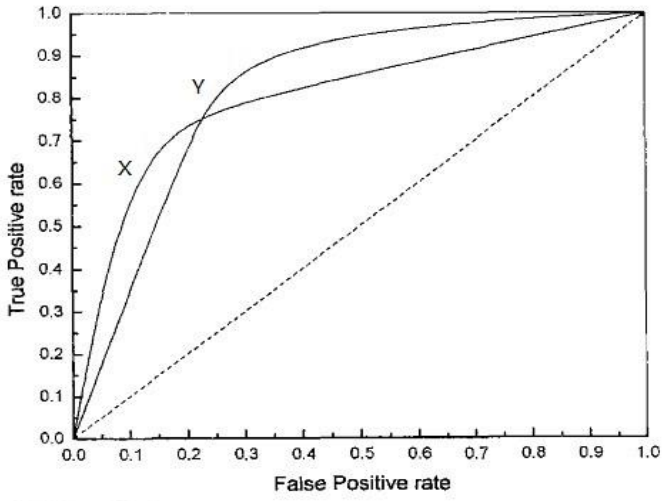

Figure 1. Two ROC Curves X and Y

3.2. Experimental Results and Analysis

The two interior factors of imbalanced data set are category unbalanced distribution and information deficiency. Imbalance ratio is majority class to minority class, which indicate degree of imbalance. And information deficiency means the information content of minority class. In order to verify the performance of our method, we select four group datasets of UCI public data platform (The essential information of four data sets are listed in Table 1) which representing four situations of imbalanced data respectively. It reflects characteristic from all aspects of imbalanced data by using these data sets. So can proving the effectiveness and feasibility of this method.

Table 1. The Basic Information of Four UCI Datasets

Dataset	The number of negative examples	The number of positive examples	Imbalance ratio	Data description
Shuttle	57829	171	338:1	High imbalance ratio and the big amount of information
Abalone	4145	32	130:1	High imbalance ratio and small amount of information
Yeast	1433	51	28:1	Low imbalance ratio and small amount of information
Churn	4293	707	6:1	Low imbalance ratio and big amount of information

Do comparative experiments on these four kind datasets using our method (compared with SVM), the ROC curve as follows:

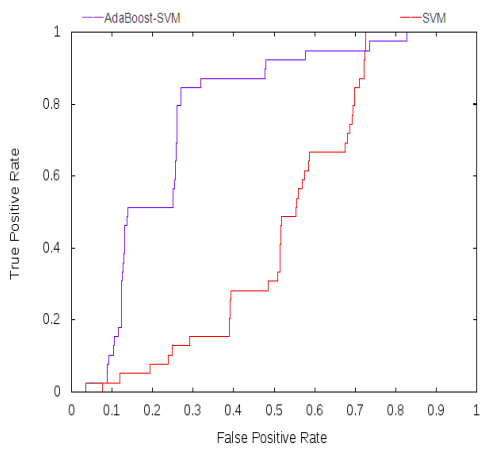

Figure 2. Comparative Experiment of Shuttle Dataset

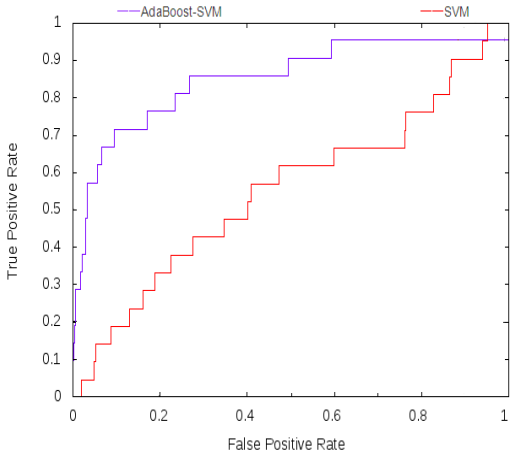

Figure 3. Comparative Experiment of Yeast Dataset

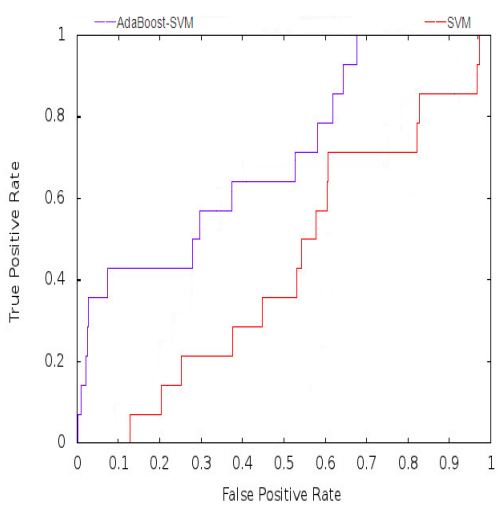

Figure 4. Comparative Experiment of Churn Dataset

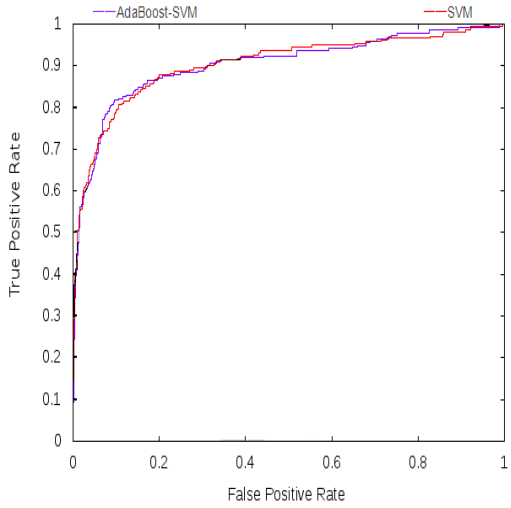

Figure 5. Comparative Experiment of Abalone Dataset

The experimental results have been reflected the advantage of our method compared to traditional SVM classification algorithm intuitively. As a result of ROC curve can't give quantifiable expression the level of classification performance promoting. In order to comparing in a quantitative level further, we calculated AUC values (in table 2) of these four datasets.

Table 2. The AUC Value of 2 Methods in Different Data Sets

Data Set	SVM	AdaBoost-SVM
Shuttle	0.5091	0.7670
Abalone	0.5159	0.7004
Yeast	0.5540	0.8487
Churn	0.8745	0.9053

The ROC curve and AUC values explain the advantage of AdaBoost: it can improve the accuracy of classification greatly in the case of large imbalance ratio, don't need to know the weaker classifier's floor level of correct rate in advance and apply it to real-life problems easily. We also can conclude the advantage of AdaBoost-SVM: the main classifier adjusting kernel function parameter values of every component classifier according to training samples, that to make precision and otherness of component classifiers achieve a certain balance. Base classifier performance is better than the single one combined by ensemble learning method. The classification experiment in standard dataset show AdaBoost-SVM is effective, meanwhile this algorithm applies equally in imbalanced dataset.That's a complete system of feasible and well classification performance by using cost sensitive support vector machine with penalty factor and ensemble learning method.

4. Conclusion

In this paper, we analysis the performance of SVM based on imbalanced data classification problem and introduce AdaBoost algorithm. Constructing nonlinear SVM

classifier and focusing on the fallible samples by layered combination and iterative weights. And verify the algorithm's advantage of AdaBoost-SVM on four UCI datasets.

We found improved SVM solving classification problem of imbalanced dataset is better. But the best thing is design a special kernel function that can solve the problem completely, which also is the difficulty and the innovation point. So we will focus on special kernel function of imbalanced data in the future study and work.

Acknowledgements

This paper is partially supported by Foundation for University Key Teacher of Heilongjiang Province (1252G023). National Natural Science Foundation of China (61103149), Technological Innovation Foundation for Youth Scholars of Harbin (2012RFQXG093) and Province Natural Science Foundation of Heilongjiang (QC2013C060).

References

[1] C. Seiffert, T. M. Khoshgoftaar and J. V. Hulse, *et al.*, "RUSBoost: A hybrid approach to alleviating class imbalance", [J], Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on, vol. 40, no. 1, **(2010)**, pp. 185-197.

[2] S. Maldonado and J. López, "Imbalanced data classification using second-order cone programming support vector machines", [J], Pattern Recognition", **(2013)**.

[3] Z. Zhao, P. Zhong and Y. Zhao, "Learning SVM with weighted maximum margin criterion for classification of imbalanced data", [J], Mathematical and Computer Modelling", vol. 54, no. 3, **(2011)**, pp. 1093-1099.

[4] J. H. Fu and S. L. Lee Certainty-enhanced active learning for improving imbalanced data classification[C]//Data Mining Workshops (ICDMW), (**2011**) IEEE 11th International Conference on. IEEE, **(2011)**, pp. 405-412.

[5] N. V. Chawla, K. W. Bowyer and L. O. Hall, *et al.*, "SMOTE: synthetic minority over-sampling technique", [J], arXiv preprint arXiv: 1106.1813, **(2011).**

[6] M. Wasikowski and X. Chen, "Combating the small sample class imbalance problem using feature selection", [J], Knowledge and Data Engineering, IEEE Transactions on, vol. 22, no. 10, **(2010)**, pp. 1388-1400.

[7] A. Orriols-Puig, E. Bernadó-Mansilla and D. E. Goldberg, *et al.*, "Facetwise analysis of XCS for problems with class imbalances", [J], Evolutionary Computation, IEEE Transactions on, vol. 13, no. 5, **(2009)**, pp. 1093-1119.

[8] J. P. Hwang, S. Park and E. Kim, "A new weighted approach to imbalanced data classification problem via support vector machine with quadratic cost function", [J], Expert Systems with Applications, vol. 38, no. 7, **(2011)**, pp. 8580-8585.

[9] A. Statnikov, C. F. Aliferis and I. Tsamardinos, *et al.*, "A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis", [J], Bioinformatics, vol. 21, no. 5, **(2005)**, pp. 631-643.

[10] H. Yu, J. Ni and J. Zhao, "ACOSampling: An ant colony optimization-based undersampling method for classifyi ng imbalanced DNA microarray data", [J], Neurocomputing, **(2013)**, pp. 309-318.

[11] G. Ditzler, R. Polikar and N. Chawla, "An incremental learning algorithm for non-stationary environments and class imbalance", [C], //Pattern Recognition (ICPR), 2010 20th International Conference on IEEE", **(2010),** pp. 2997-3000.

[12] D. Bitouk, R. Verma and A. Nenkova "Class-level spectral features for emotion recognition", [J], Speech Communication, vol. 52, no.7, **(2010)**, pp. 613-625.

[13] H. Yu, J. Ni and J. Zhao, "ACOSampling: An ant colony optimization-based undersampling method for classifying imbalanced DNA microarray data", [J], Neurocomputing, vol. 101, **(2013)**, pp. 309-318.

[14] R. Prasad, D. Arpit and S. Wu, *et al.*, "Ridge Regression based classifiers for large scale class imbalanced datasets", [C], //Proceedings of the 2013 IEEE Workshop on Applications of Computer Vision (WACV). IEEE Computer Society, **(2013),** pp. 267-274.