A New Class Weighted Algorithm for PSVM

Xiaorong Zhu, Qian Zhang* and Sai Qiao
Department of Information Science and Technology, Taishan University, Taian, China

*Corresponding author email: qianzhang@tsu.edu.cn

Abstract. Focusing on the problem of slow training speed existing in the SVM in intrusion detection, this paper improves the faster PSVM (Proximal SVM) method and proposes a new kind of class - weighted PSVM algorithm. This kind of algorithm not only has a fast classification speed, but also improves the classification accuracy.

Keywords: Class weighted; PSVM; SVM.

1. Introduction
For practical problems such as fault diagnosis and intrusion detection, people often pay more attention to the classification accuracy of classes with a small number of training samples. However, due to inappropriate evaluation criteria, lack of training data, and large differences in the distribution of training samples of each category, the classification error rate of classes with small number of training samples is higher than that of classes with large number of training samples when using the traditional SVM method, which undoubtedly limits the scope of SVM.

The process of intrusion detection needs to train a large amount of data, which involves the problem of training speed. In the problem of intrusion detection, how to improve the training speed of intrusion detection is very important. The standard SVM algorithm usually needs to solve a large quadratic programming problem, while the PSVM\(^{[1]}\) algorithm only needs to solve a set of equations, so it is simple and fast. The above PSVM has the same penalty for positive and negative sample errors. In practice, the distribution of training samples is often unbalanced, and the same penalty for deviation will affect the classification results\(^{[2]}\). Literature [2] proposes a weighted support vector machine (SVM) algorithm for the standard C-SVM, but fails to provide an exact method for determining class weight and sample weight.

This paper analyzed PSVM algorithm in detail and the effect on the performance of the classifier classification caused by sample quantity, gives a weight selection strategy according to the characteristics of PSVM, which introduces a new diagonal matrix to the optimization problem, and improves the original method of linear and nonlinear PSVM method by using the method of weighting treatment to obtain the weighted processing solution. PSVM, as a fast algorithm, with the method of weighted processing, not only has fast classification speed, but also improves the classification accuracy.

2. The Influence of Sample Size on Classification
Unbalanced sample sets have always been a hot topic in machine learning. The imbalance of sample sets means that the number of samples of some classes in the sample set is higher than that of other classes. In the unbalanced sample set, due to the difference in the number of samples among different categories, the class with a small number of samples provides less classification information to the classifier, which makes the classification performance of the classifier reduce.
2.1. The Influence of Unbalanced Sample Set on Generalization Performance

Classification accuracy is the most commonly used evaluation criterion for classification problems. When this standard is used, categories with small sample size have less impact on classification accuracy than those with large sample size. Weiss’ experimental research \[3\] shows that the evaluation standard of classification accuracy leads to the classifier’s poor classification performance for the categories with a small number of samples, and the deviation rate of the categories with a small number of samples is 2 to 3 times or more than that of the categories with a large number of samples. At this point, the classifier tends to predict a sample as a category with a large number of samples. For extreme unbalanced data sets, the learning results may not have classification rules for categories with a small sample size.

Supposing \( N_{BSV_+} \) and \( N_{BSV_-} \) represent the number of boundary support vectors in the positive class and the negative class respectively, \( N_{SV_+} \) and \( N_{SV_-} \) represent the number of all support vectors in the positive class and the negative class respectively, and \( l_+ \) and \( l_- \) represent the number of samples of the positive class and the negative class respectively, according to the constraints of the dual problem, we can get

\[
\sum_{i=1}^{l_+} y_i \alpha_i = \sum_{y_i=+1} y_i \alpha_i + \sum_{y_i=-1} y_i \alpha_i \quad (1)
\]

Due to \( \sum_{i=1}^{l} y_i \alpha_i = 0 \) and \( \sum_{y_i=+1} \alpha_i = \sum_{y_i=-1} \alpha_i \quad (2)\)

and \( 0 \leq \alpha_i \leq C \), then

\[
N_{BSV_+} \cdot C \leq \sum_{y_i=+1} \alpha_i , N_{SV_-} \cdot C \geq \sum_{y_i=-1} \alpha_i \quad (3)
\]

So that \( N_{BSV_+} \cdot C \leq \sum_{y_i=+1} \alpha_i \leq N_{SV_+} \cdot C \quad (4)\)

In the same way with

\[
N_{BSV_-} \cdot C \leq \sum_{y_i=-1} \alpha_i \leq N_{SV_-} \cdot C \quad (5)
\]

Furtherly

\[
\frac{N_{BSV_+}}{l_+} \leq \frac{\sum_{y_i=+1} \alpha_i}{C \cdot l_+} \leq \frac{N_{SV_+}}{l_+} \quad (6)
\]

\[
\frac{N_{BSV_-}}{l_-} \leq \frac{\sum_{y_i=-1} \alpha_i}{C \cdot l_-} \leq \frac{N_{SV_-}}{l_-} \quad (7)
\]

From (6) and (7): if \( l_+ < l_- \), then the upper bound of the boundary support vector proportion in the positive class is greater than the upper bound of the boundary support vector proportion in the negative class, similarly, vice versa when \( l_- < l_+ \). Therefore, if the number of samples in the two categories is different, the misclassified proportion of samples in the small sample category is greater than that in the large sample category. This deviation shows that the standard SVM is not suitable for class unbalanced classification. In addition, in a number of practical problems, the importance between samples is not the same. Ignoring the importance of samples can lead to important samples being
misclassified. There are few new attacks in network intrusion detection. If they are distributed near the boundary or in the normal sample area, we will not detect these new attacks by using SVM classification method without considering class balance to classify them. If these attacks are very dangerous to the system, then price of misclassification can be very costly.

2.2. The Solution to the Unbalanced Problem of Sample Set
As for the imbalance of sample set, there are few researches\cite{4} on class distribution differences and most of them consider the imbalance of training sample number. Among them, reconstruction of data set and improvement of algorithm are two main directions.

The purpose of data set reconstruction is to obtain a relatively balanced data set, which can be divided into two types: one is to reduce the number of samples of classes with large sample size, and the other is to increase the number of samples of classes with small sample size. Simple data set reconstruction can cause a lot of problems, so a heuristic approach and its improvements are generated\cite{5}. In the aspect of algorithm improvement, two regularization parameters are mainly used to control two kinds of error penalty respectively for standard C-SVM and V-SVM problems. By studying two regularization parameters C-SVM and V-SVM, hong-gunn Chew et al. found that when the ratio of two regularization parameters is equal to the ratio of the number of samples of two types, a SVM with similar classification error rate will be obtained.

3. Improved Class - weighted PSVM Method

3.1. PSVM Method

The standard SVM method solves the following optimization problems

\[
\min \frac{1}{2} \|w'w + ve'\xi\|_2^2
\]
\[
s.t. \quad D(Aw - e\gamma) + \xi \geq e
\]
\[
\xi \geq 0
\]

The PSVM method is further improved on the standard SVM method. Firstly, the measure of the degree of right and wrong in the objective function is changed from the norm 1 to the norm 2 to get rid of the nonnegative constraint on \(\xi\); Secondly, in order to make the maximum distance between the two boundary hyperplane, considering the two factors \(w, \gamma\), this modified optimization problem not only does classification as well as classic (8), and get the hyperplane as (8), but also has other advantages such as strong convex the objective function. The most important improvement of PSVM is the third point. By changing the inequality constraint in the constraint condition to the equality constraint, it can quickly find a solution to approximate the classification hyperplane of the original problem, which also leads to the change of support vector mentioned below. With the above improvements, PSVM can solve the following optimization problems

\[
\min_{(w, \gamma, \xi)} \frac{1}{2} \|\xi\|^2 + \frac{1}{2} (w'w + \gamma^2)
\]
\[
s.t. \quad D(Aw - e\gamma) + \xi = e
\]

After these three improvements, the PSVM method removes the non-negative constraint on the relaxation variables, thus simplifying the constraint conditions. More importantly, the problem is changed to the equality constraint problem, which fundamentally changes the nature of the original optimization problem. As described in the following text, the exact solution of the problem can be obtained quickly by using the training data.

Figure 1 and figure 2 graphically illustrate the standard SVM and PSVM methods.
3.2. Improved PSVM Approach

When there is a large difference in two kinds of samples, the classification effect of PSVM will be affected by the number of samples. For the party with a small sample size, the classification result is not ideal, so the factor of sample size should be considered in the penalty of misclassification. Supposing that the total number of samples is $m$, $m_1, m_2$ respectively is the number of samples of positive class and negative class. Define a diagonal matrix $N$ of $m \times m$ size as the weight matrix and use this matrix to adjust the penalty for misclassification.

$$N = \begin{cases} 
\frac{1}{m_1}, & \text{if } y_i = 1 \\
\frac{1}{m_2}, & \text{if } y_i = -1
\end{cases}$$  \hspace{1cm} (10)

The following two cases, linear and nonlinear, are discussed.

3.2.1. Linear PSVM. After introducing the weight matrix $N$, the original PSVM problem can be further modified as

$$\min_{(w, \gamma, e)} \frac{1}{2} w^T N \xi + \frac{1}{2} (w^T w + \gamma^2)$$  \hspace{1cm} (11)

s.t. $D(Aw - e \gamma) + \xi = e$

Its Lagrange function is

$$L(w, \gamma, \xi, u) = \frac{1}{2} w^T N \xi + \frac{1}{2} (w^T w + \gamma^2)$$  \hspace{1cm} (12); From KKT condition

$$-u'(D(Aw - e \gamma) + \xi - e)$$

$$w = A' Du; \quad \gamma = -e' Du; \quad N \xi = u \sqrt{v}$$  \hspace{1cm} (13); $D(Aw - e \gamma) + \xi - e = 0$ (14); Substitute (3.13) into (3.14), the expression $u$ can be available. $u = (NDEE' D + \frac{I}{v})^{-1} Ne$, $E = [A - e]$ (15); For $u$, by using Sherman-Morrison-Woodbury formula[18], we can get $u = (vN - vNDE(\frac{I}{v} + E' NE)^{-1} E' D N)e$ (16). And then

![Figure 1. Classical SVM](image1)

![Figure 2. PSVM method](image2)
we can simplify the relationship between \( w, \gamma \) and \( u(w, \gamma) = (I_v + E'NE)^{-1} E'Nd e \) (17).

The final classification hyperplane is \( x'w - \gamma = 0 \). In this way, it can be obtained \( w, \gamma \) simply and quickly through (17), and the classification accuracy is improved by taking into account the difference between the number of samples of the positive class and the number of samples of the negative class, so that the classification effect is not affected by the number of samples.

### 3.2.2. Nonlinear PSVM.

In order to obtain nonlinear PSVM, the \( w = A'Du \) in (3.13) is substituted into the constraint function in 3.11 and the nonlinear kernel \( K(A, A') \) was used to replace the linear kernel \( AA' \), and the objective function is correspondingly modified into a minimization problem, then

\[
\min_{(u, \gamma, \xi)} \frac{1}{2} \xi'N\xi + \frac{1}{2}(u'u + \gamma^2)
\]

\[
\text{s.t.} \quad D(K(A, A')Du - e\gamma) + \xi = e
\]

Suppose \( \beta \) is its Lagrange multiplier, which is similar to the linear PSVM, then

\[
u = DK(A, A')D\beta, \quad \gamma = -e'D\beta, \quad N\xi = \frac{\beta}{\nu}
\]

Finally obtaining \( \beta = (NDE'ED + \frac{I}{\nu})^{-1}Ne \), and \( E = [K(A, A') - e] \) (20).

In the linear case, the classification hyperplane is \( x'w - \gamma = x'A'Du - r = 0 \), and in the nonlinear case, the nonlinear kernel function \( K(x', A') \) replaces \( x'A' \), then the final classification \( K(x', A')Du - \gamma = 0 \) hyperplane can be obtained.

In the linear case, \( A \) is a \( m \times n \) matrix and \( n \equiv m \), so we can use the SMW formula to find the inverse. In the nonlinear case, the kernel matrix \( K(A, A') \) is a \( m \) order square, so it is meaningless to apply SMW formula.

### 3.3. Other Selection Methods of Weight

There are many strategies for weight selection. Literature [6] proposes another method for weight selection. According to the knowledge of multivariate analysis, the main eigenvalues of the covariance matrix express the dispersion degree of each uncorrelated direction of the data, while the traces of the covariance matrix express the overall dispersion degree of the data. Therefore, the trace of different covariance matrix can provide important information for data classification of the same distribution.

Assuming that the covariance matrices of positive and negative samples in the kernel space are \( \Sigma^+ \) and \( \Sigma^- \) respectively, by kernel principal component analysis, we can obtain the first \( p \) main eigenvalues of \( \Sigma^+ \), which are \( \lambda_1, \lambda_2, \ldots, \lambda_p \) and the first \( p \) main eigenvalues of \( \Sigma^- \), which are \( \mu_1, \mu_2, \ldots, \mu_p \).

The trace of two kinds of covariance matrix, \( Tr(\Sigma^+) \approx \sum_{i=1}^{p} \lambda_i \) and \( Tr(\Sigma^-) \approx \sum_{i=1}^{p} \mu_i \) is used to represent the dispersion degree of two kinds of data. Let \( C^+ \) and \( C^- \) respectively represent the penalty for the positive and negative samples, then if \( C^+ = \lambda C \), \( C^- = (1-\lambda)C \), and

\[
\lambda = \frac{Tr(\Sigma^+)}{Tr(\Sigma^+)/N^+ + Tr(\Sigma^-)/N^-}
\]

\( C \) is a normal number, represents the penalty.

The disadvantage of this method is that it is limited to the classification problem of the same type of distribution, and in the case of large amount of data, which requires incremental learning, the SVM
algorithm needs more training time due to the calculation of kernel function matrix, covariance matrix, eigenvalue, etc.

4. Numerical Experimentation

4.1. Data Preprocessing and Experimental Results
The experiment was conducted on Matlab 7.1 on the Windows platform. The computer CPU was Intel(R)Core(TM)i3-7100 CPU@3.90GHz and the memory was 4G. The kernel function of the experiment adopts the common radial basis kernel function $K(x, y) = \exp(-\|x - y\|^2 / 2\sigma^2)$, and uses the grid search method to select the penalty parameters $\nu$ in PSVM problem and the parameters $\sigma$ in the kernel function. The parameter $n_1 \times n_2$ grid is obtained by taking a value $n_1, n_2$ for $\nu, \sigma$ respectively, from which the parameter combination with the strongest generalization ability is searched. Then divide the grid around the combination and search until a satisfactory solution is obtained. Formula (10) is used to construct the weight matrix $N$ of each training.

4.2. Analysis of Experimental Results
Table 1 lists the Detection Rate (DR), False Positive Rate (FP) and Training Time (TT) of intrusion Detection classifier when the training set takes different training parameters in the order of standard SVM, PSVM and class-weighted PSVM. RR refers to the reduction Rate, which is the ratio of the reduced data set to the original data set in the Reduced SVM method.

| PC       | NC       | SVM       | PSVM       | Balanced PSVM |
|----------|----------|-----------|------------|---------------|
|          |          | Linear    | Nonlinear  | Linear        | Nonlinear     | RR          |
| PC       | parameter | $\nu=100$ | (50,0.45)  | $\nu=100$    | (100,0.2)     | $\nu=100$   | (100,0.5) |
| 600      | DR(%)    | 94.23     | 95.20      | 93.12        | 94.43         | 94.72       | 96.73     |
| NC       | FP(%)    | 0.81      | 0.75       | 1.21         | 0.98          | 0.61        | 0.43      |
| 1200     | TT(s)    | 2.23      | 43.37      | 0.1150       | 24.89         | 0.1335      | 25.02     |
| PC       | parameter | $\nu=50$  | (100,0.2)  | $\nu=20$     | (60,0.1)      | $\nu=100$   | (100,0.2) |
| 600      | DR(%)    | 85.12     | 84.20      | 84.42        | 83.13         | 93.12       | 96.01     |
| NC       | FP(%)    | 2.13      | 1.83       | 3.41         | 1.94          | 1.72        | 0.96      |
| 2400     | TT(s)    | 18.74     | 82.33      | 0.1229       | 8.10          | 0.1411      | 7.94      |
| PC       | parameter | $\nu=20$  | -          | $\nu=10$     | (50,0.8)      | $\nu=100$   | (100,0.1) |
| 600      | DR(%)    | 78.05     | -          | 76.13        | 80.43         | 94.13       | 95.03     |
| NC       | FP(%)    | 3.10      | -          | 9.76         | 2.45          | 2.85        | 0.80      |
| 4000     | TT(s)    | 5.12      | -          | 0.1394       | 1.71          | 0.1512      | 1.83      |

5. Conclusion
As can be seen from table 1, compared with the standard SVM method, the training time of PSVM method with class weight is greatly reduced, which reflects the advantages of computing time. However, with the increase of the proportion of the number of training samples in the two categories, the classification effect of the standard method and PSVM method without the weighted method decreased rapidly, while that of the PSVM method with the weighted method was not affected much,
and all indexes were better than that of the standard method and PSVM.

Acknowledgements
Tai’an City science and technology development plan project of 2018 (2018GX0061), a reconstruction method of undersampled magnetic resonance image based on BM3D.

References
[1] G.Fung, O.L. Mangasarian. Proximal Support Vector Machine Classifiers[R] ,Data Mining Institute Technical Report 01-02,Februray 2001.Proceedings of KDD-2001,San Francisco,August 26-29,2001.Association for Computing Machinery, New York, 77-86
[2] Jia yin-shan, Jia chuan-ying. An algorithm of weighted support vector classifier [J]. Computer engineering, 2005, 31 (12): 23-25
[3] G.M.Weiss, F.Provost. Learning When Training Data Are Costly: The Effect of Class Distribution on Tree Induction[J], Journal of Artificial Intelligence Research, 2003, 19(3):315-354
[4] Y.L.Gracew. Support Vector Machines for Classification in Nonstandard Situations[J], Machine Learning, 2002, 46(2):191-202
[5] A.Estabrooks, J.O. Taeho and N. Japkowicz. A Multiple Resampling Method for Learning from Imbalanced Data Sets[J], Computational Intelligence , 2004, 20(1):18-36
[6] Wang jinyan, Feng Jianwu, Liu wanli. A method for correcting unbalanced support vector machines [J]. Computer applications, 2007, 27 (12): 2896-289