A NEW SUPPORT VECTOR MACHINE FOR THE CLASSIFICATION OF POSITIVE AND UNLABELED EXAMPLES

***Junyan Tan*[1], *Ling Zhen*[1], *Naiyang Deng*[1], *Chunhua Zhang*[2]**

[1] *College of Science, China Agricultural University, Beijing 100083, China*
[2] *Information School, Renmin University of China, Beijing 100872, China*
*tanjunyan0@126.com, zhangchunhua@ruc.edu.cn**

Keywords: Support vector machine, feature selection, p-norm, PU learning.

Abstract

In this paper, we propose a new version of support vector machine named biased p-norm support vector machine (BPSVM) involved in learning from positive and unlabeled examples. BPSVM treats the classification of positive and unlabeled examples as an imbalanced binary classification problem by giving different penalty parameters to positive and unlabeled examples. Compared with the previous works, BPSVM can not only improve the performance of classification but also select relevant features automatically. Furthermore, an effective algorithm for solving our new model is proposed. BPSVM can be used to solve large scale problem due to the effectiveness of the new algorithm. Numerical results show BPSVM is effective in both classification and features selection.

1 Introduction

The traditional classification task is to construct a classification function based on the labeled training set. Different from the traditional classification task, another special kind of problem, namely, learning from positive and unlabeled examples (PU learning), gains more and more attention [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11]. It deals with such a case that no labeled negative examples exist in the training set, that is to say, only a few labeled positive examples and lots of unlabeled examples are available, without any information about negative class.

Currently, there are three kinds of methods for solving PU learning problem: the first one only use the labeled positive examples, the second one is based on two-step strategy and the third one is based on one-step strategy. For the first one, one-class support vector machine is the exact example[12]. One-class SVM tries to learn the final classification hyper plane only using the positive examples without using any information of the unlabeled examples. The numerical results show that it performs poorer than the learning methods that take advantage of the unlabeled examples. Most of the popular methods for solving PU learning are based on two-step strategy [1, 4, 5, 6, 11]. Two-step methods are iteratively conducted as the following two steps. Step 1: Identifying the reliable negative or positive examples from the unlabeled set to enlarge the original training set. Step 2: Building a set of classifiers by iteratively applying a classification algorithm and then selecting a good classifier from the set. These two steps together is an iterative method of increasing the number of unlabeled examples that are classified as negative while maintaining the positive examples correctly classified. [6] points out that if the sample size is large enough, maximizing the number of unlabeled examples classified as negative while constraining the positive examples to be correctly classified will give a good classifier. The one-step methods convert the PU learning into an unbalance binary classification problem [3, 7], such as biased-SVM. Biased-SVM gives bigger weights to the positive examples and small weights to the unlabeled examples which are regarded as negative examples with noise. The numerical results on the public benchmark data sets show that the performance of biased-SVM is better than most of two-step methods.

Although there are a lot of methods that can solve PU learning problem well, none of them consider the feature selection in PU learning. The benefit of feature selection is twofold. Firstly, it is meaningful because it can identify the features that contribute most to classification. Secondly, it is helpful for solving the classification problem because it can not only reduce the dimension of input space and speed up the computation procedure, but also improve the classification accuracy. But in PU learning, only positive examples are given, we have no negative examples. How to do the feature selection in PU learning? Which features should we keep? This paper answers these questions by conducting a new version of support vector machine which can perform feature selection and classification in PU learning simultaneously. Precisely, given the training set

$$T = \{(x_1, y_1), \cdots, (x_l, y_l), x_{l+1}, \cdots, x_{l+m}\} \in (\mathcal{X} \times \mathcal{Y})^l, \tag{1}$$

where $x_j \in R^n (j = 1, \cdots, l+m), y_j = 1, (j = 1, \cdots, l)$.

*Corresponding author: Chunhua, Zhang

There are two things to aim at in this paper. First, to get a classifier; Second, to select the relevant features at the same time.

Recently, p-norm ($p \in [0,1]$) attracts great attention in the optimization framework, the idea that using p-norm can find sparse solution is considered in [13, 14, 15, 16, 17, 18, 19, 20, 21]. Correspondingly, [17, 18, 19, 20] propose p-norm ($0 < p < 1$) support vector machine, which replace the 2-norm penalty by the p-norm ($p \in (0,1)$) penalty in the objective function in the primal problem in the standard linear SVM. p-norm SVM performs well for classification and feature selection. In this paper, we propose a new version of p-norm SVM named biased p-norm SVM (BPSVM) which gives different weights to the positive and unlabeled examples. BPSVM is solved approximately by an iteratively reweighted 2-norm SVM alternating between estimating normal vector w and redefining the weights. The numerical experimental results show that BPSVM is more effective in classification than some popular methods such as Biased-SVM [22] and the extended biased SVM (EBSVM). Moreover, BPSVM can realize feature selection while the other methods can not.

Now we describe our notation. All vectors are column vectors unless transposed to a row vector by a super script $\top$. For a vector x in R^n, $[x]_i (i = 1, 2, \cdots, n)$ denotes the i-th component of x. $|x|$ denotes a vector in R^n of absolute value of the components of x. $\|x\|_p$ denotes that $(|[x]_1|^p + \cdots + |[x]_n|^p)^{\frac{1}{p}}$. Strictly speaking, $\|x\|_p$ is not a general norm when $0 \leq p < 1$, but we still follow this term p-norm, because the forms are same except that the values of p are different. $\|x\|_0$ is the number of nonzero components of x. For two vectors $x \in R^n$ and $y \in R^n$, $(x \cdot y)$ denotes the inner product of x and y; $x \otimes y$ denotes a vector in R^n whose ith element is just $x_i y_i$.

This paper is organized as follows. In section 2, we first introduce some previous works related to this paper, then our new method, the biased p-norm support vector machine for both feature selection and classification is proposed. Furthermore, an iterative algorithm for solving the optimization problem of biased p-norm support vector machine is also carried out. In section 3, numerical experiments are given to demonstrate the effectiveness of our method. We conclude this paper in section 4.

2 Methods

2.1 Related Works

In this section, we briefly introduce several previous works related to this paper.

2.1.1 2-norm SVM

Consider the standard classification problem first. Given the training set

$$T1 = \{(x_1, y_1), \cdots, (x_l, y_l)\} \in (\mathcal{X} \times \mathcal{Y})^l, \qquad (2)$$

where $x_j (j = 1, \cdots, l) \in R^n$, $y_j \in \{1, -1\}$.The standard 2-norm SVM seeks an optimal separating hyperplane that maximizes the margin between two classes and the optimal decision function $\mathrm{sgn}((w^* \cdot x) + b^*)$ decided by the following optimization problem [23, 24, 25, 26].

$$\min_{w,b,\xi} \quad \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\sum_{i=1}^{l}\xi_i \ , \qquad (3)$$

$$\text{s.t.} \quad y_i((\boldsymbol{w}\cdot\boldsymbol{x}_i)+b) \geq 1-\xi_i \ , \ i = 1, \cdots, l \ , \qquad (4)$$

$$\xi_i \geq 0 \ , \ i = 1, \cdots, l, \qquad (5)$$

where C is the penalty parameter which can balance the empirical risk and the confidence interval, ξ_i is the slack variables measuring the classification loss of examples.

2.1.2 The p-norm SVM

The p-norm ($0 < p < 1$) support vector machines are proposed by Tan for the feature selection in supervised binary classification [19, 20]. The p-norm SVM replaces the 2-norm penalty by the p-norm ($0 < p < 1$) penalty in the objective function in the primal problem in the standard 2-norm SVM:

$$\min_{w,b,\xi} \quad \|\boldsymbol{w}\|_p^p + C\sum_{i=1}^{l}\xi_i \ , \qquad (6)$$

$$\text{s.t.} \quad y_i((\boldsymbol{w}\cdot\boldsymbol{x}_i)+b) \geq 1-\xi_i \ , \ i = 1, \cdots, l \ , \qquad (7)$$

$$\xi_i \geq 0 \ , \ i = 1, \cdots, l. \qquad (8)$$

The numerical experimental results show that the p-norm SVM is more effective in feature selection than some popular methods such as 1 -norm SVM and 0-norm SVM.

2.1.3 Biased 2-norm SVM

Now, we focus on the PU learning problems with the training set (1). Biased 2-norm SVM (BSVM) is a one-step method by converting the classification of positive and unlabeled examples to an imbalanced binary classification, supposing that the unlabeled examples in (1) are negative examples, i.e. the labels of $x_{l+1}, \cdots, x_{l+m}$ in (1) are supposed to be -1. The classifier is conducted by giving appropriate weights to the positive examples error and negative examples error respectively:

$$\min_{w,b,\xi} \quad \frac{1}{2}\|\boldsymbol{w}\|^2 + C_1\sum_{i=1}^{l}\xi_i + C_2\sum_{i=l+1}^{l+m}\xi_i \ , \qquad (9)$$

$$\text{s.t.} \quad y_i((\boldsymbol{w}\cdot\boldsymbol{x}_i)+b) \geq 1-\xi_i \ , i = 1, \cdots, l+m \qquad (10)$$

$$\xi_i \geq 0 \ , \ i = 1, \cdots, l+m \ , \qquad (11)$$

where C_1 and $C_2$are the penalty factors of misclassification for positive and unlabeled example sets respectively. Usually, C_1 is larger than C_2. $\xi_i, i = 1, \cdots, l+m$ are slack variables.

2.2 New methods

In this section, our new methods are introduced in detail including the optimization problem of biased p-norm SVM, the solving algorithm for the optimization problem and the final classification algorithm for the classification of positive and unlabeled examples.

2.2.1 Optimization Problem

We now present the biased p-norm support vector machine (BPSVM) formulation of the problem. BPSVM is an embedded feature selection method in which training data are given to a learning machine. BPSVM returns a predictor and a subset of features on which it performs predictions. In fact, feature selection is performed in the process of learning.

For the PU learning problem, [7] indicates that if the sample size is large enough, minimizing the number of unlabeled examples classified as positive while constraining the positive examples to be correctly classified will give a good classifier. Following this idea and the motivation of feature selection, we propose the following biased p-norm SVM (no error for positive examples but only for unlabeled examples).

$$\min_{w,\,b,\,\xi} \quad F(w,b,\xi) = \|\boldsymbol{w}\|_p^p + C\sum_{i=l+1}^{l+m}\xi_i\ , \quad (12)$$

$$\text{s.t.} \quad y_i((\boldsymbol{w}\cdot\boldsymbol{x}_i)+b) \geq 1, i = 1,\cdots,l, \quad (13)$$

$$y_i((\boldsymbol{w}\cdot\boldsymbol{x}_i)+b) \geq 1-\xi_i\ ,\ i = l+1,\cdots,l+m, \quad (14)$$

$$\xi_i \geq 0\ ,\ i = l+1,\cdots,l+m\ , \quad (15)$$

where $0 < p < 1$ and $C > 0$ are parameters. While, the positive examples may contain some errors in practice. Thus, we allow error in the positive examples and propose the following soft margin version of BPSVM which uses two parameters C_1 and C_2 to weight positive errors and negative errors differently.

$$\min_{w,\,b,\,\xi} \ F_p(w,b,\xi) = \|\boldsymbol{w}\|_p^p + C_1\sum_{i=1}^{l}\xi_i + C_2\sum_{i=l+1}^{l+m}\xi_i\ , \quad (16)$$

$$\text{s.t. } y_i((\boldsymbol{w}\cdot\boldsymbol{x}_i)+b) \geq 1-\xi_i\ , i = 1,2,\cdots,l+m\ , \quad (17)$$

$$\xi_i \geq 0\ ,\ i = l+1,\cdots,l+m\ . \quad (18)$$

C_1 and C_2 can be adjusted to achieve our objective. Intuitively, we give a big value for C_1 and a small value for C_2 because the unlabeled set, which is assumed to be negative, also contains positive data.

We now give the geometric interpretation of the BPSVM. The first term $\|\boldsymbol{w}\|_p^p (0 < p < 1)$ in the objective function of problem (16-18) is the regularizer that can control the sparsity of the final classification hyperplane. The second and the third term of the objective function of problem (16-18) minimize the sum of error variables, which attempts to over-fit the training examples.

Note that, its very difficult to find the global solution of problem (16-18) because its objective function is neither convex nor differentiable. This will be considered seriously in the following section.

2.2.2 Algorithm for Solving Problem (16-18)

Although the objective function of problem (16-18) is composed of a concave term $\|\boldsymbol{w}\|_p^p$ $(0 < p < 1)$ and a convex term $C_1\sum_{i=1}^{l}\xi_i + C_2\sum_{i=l+1}^{l+m}\xi_i$, which can be regarded as a difference between two convex functions, it can't be solved by CCCP because the concave term is not differentiable. By the idea of CCCP , we propose a new algorithm which is an iterative process.

At the k-th iteration, denote the current (w,b,ξ) estimate by $(w^{(k)},b^{(k)},\xi^{(k)})$, respectively, and then setting $(w^{(k+1)},b^{(k+1)},\xi^{(k+1)})$ as the solution to the following weighted biased SVM:

$$\min_{w,\,b,\,\xi} \ \frac{1}{2}\|\beta^{(k+1)}\otimes w\|^2 + C_1\sum_{i=1}^{l}\xi_i + C_2\sum_{i=l+1}^{l+m}\xi_i\ , \quad (19)$$

$$\text{s.t. } y_i((\boldsymbol{w}\cdot\boldsymbol{x}_i)+b) \geq 1-\xi_i\ ,\ i = 1,\cdots,l+m\ , \quad (20)$$

$$\xi_i \geq 0\ ,\ i = 1,\cdots,l+m\ , \quad (21)$$

where $\beta^{(k+1)} = (\beta_1^{(k+1)},\cdots,\beta_n^{(k+1)})^\top$ is the weight vector and satisfies that:

$$\nabla F_p(w^{(k)},b^{(k)},\xi^{(k)}) = \nabla F_2(w^{(k)},b^{(k)},\xi^{(k)}), \quad (22)$$

$F_2(w^{(k)},b^{(k)},\xi^{(k)})$ is the objective function of problem (19-21). It is easy to have $\beta_i^{(k+1)} = p|[w^{(k)}]_i + \varepsilon|^{p-2}$, where $\varepsilon > 0$ is to guarantee that $\beta_i^{(k+1)}$ is well defined, $i = 1,2,\cdots,n$.

Based on the above idea, we propose the following solving algorithm for the problem (16)-(18).

Algorithm1: Solving the problem (16)-(18)

(1) Given $C_1 > 0, C_2 > 0$ and $p(p \in (0,1))$, start with a random $\beta^{(0)}$ and let $k = 1$;

(2) Solve the weighted optimization problem (19)-(21)and get the solution $(w^{(k)},b^{(k)},\xi^{(k)})$;

(3) Terminate on convergence or where k attains a specified maximum number of iteration K_{max}. Otherwise, set $k = k+1$ and update the weights for each $i = 1,2,\cdots,n$,

$$\beta_i^{(k+1)} = p|[w^{(k)}]_i + \varepsilon|^{p-2}, i = 1,\cdots,n$$

and go to step 2.

Note that, the optimization problem solved in step (2) can be converted to a standard 2-norm biased SVM, which assures that it can be solved by the well known software easily. Thus, we can apply Algorithm 1 to solve large-scale problems.

2.2.3 Biased p-norm Support Vector Classification (BPSVC)

Because p-norm can induce the sparse solution, there will be many components which are as closed as zero, we can eliminate these components and realize the

feature selection. The new algorithm for classification and feature selection in PU learning is established as follows:

Algorithm 2:BPSVC

(1) Given the parameters $C_1 > 0, C_2 > 0$, $p(0 < p < 1)$ and a very small number $\varepsilon > 0$; using the set given by (1), construct the optimization problem (16)-(18);

(2) Using the Algorithm 1 to get the local optimal solution (w^*, b^*, ξ^*) to (16)-(18);

(3) Select the feature index set: $F' = \{i|[w^*]_i| > \varepsilon, i = 1, \cdots, n\}$;

(4) Construct the decision function $f(x) = \text{sgn}((\tilde{w}^* \cdot \tilde{x}) + b^*)$, where $\tilde{w}^*$ are composed by the components in the F' of w^* and the components of $\tilde{x}$ are also corresponding to components in the feature set F' of w^*.

In the following section, our experiments are conducted according to the Algorithm 2.

3 Results

The numerical experiment and results on several real data sets are carried out in this section.

3.1 Experiment Setup

Datasets In this section, some experiments on Reuters corpus [28] and Phospho.ELM (version 1009) are conducted. For Reuters corpus the top ten popular categories are used. Each category is employed as the positive class, and the rest as the negative class. This gives us 10 data sets. Phosphorylation data contains a collection of experimentally verified serine (S), threonine (T), and tyrosine (Y) specific phosphorylation sites in eukaryotic proteins. The entries provide the information about the phosphorylated proteins and the exact positions of the known phosphorylated residues, which are catalyzed by a given kinase. We consider four common kinase: CDK1, CDK2, CK2 and CDK. Therefore, we have four data sets. The statistics of all data sets are shown in Table 1.

Preprocessing For each dataset, 30% of the documents are randomly selected as test documents. The remaining (70%) are used to create training sets as follows: δ percent of the documents from the positive class is first selected as the positive set P. The rest of the positive documents and negative documents are used as unlabeled set U. For all data sets, we range δ from 30%,70% (0.3,0.7)

In the experiment, we would compare our method with 2-norm biased SVM (BSVM) and the extended 2-norm biased SVM (EBSVM). The parameters in the optimization problem are optimized on training sets. The range of C_1 and C_2 is chosen from $\{2^{-5}, 2^{-4}, \cdots, 2^5\}$, p in biased p-norm SVM (BPSVM) is chosen from 0.1 to 0.9

Evaluation Criteria F-score on the positive class is used to evaluate the performance of the classifiers on the test sets. F-score takes into account of both recall and precision and is defined as:

$$F = \frac{2pr}{p + r},$$

where $r = TP/(TP + FN)$, $p = TP/(TP + FP)$, TP and FP represent the number of true positive and false positive examples respectively. FN is the number of false negative examples.

F-score cannot be calculated on the validation set during the training process because there are no negative examples. An approximate method [6] is used to evaluate the performance by

$$F = \frac{r_P^2}{Prob(f(x) > 0)},$$

where x is an input vector, $Prob(f(x) > 0)$ is the probability of this input example x classified as positive, r_P is the recall for positive set P in the validation set. The approximate F is used to select the optimal parameters during the training process.

3.2 Experimental Results

Results on Reuters Collection

The classification performance is shown in Table 2-Table 4. We can see that BPSVM performs better. For each data set with different δ, most of the mean of average F scores of BPSVM is higher than the other two methods. Table 3 shows the average F scores on each data set with all δ. It is clearly that BPSVM is the best, because it performs best on seven data sets. Table 4 shows the average number of selected features, the most is 779 and the least is 145. While, the other two methods use all the featurese and can't select any feature. All in all, BPSVM is the best one, because it conducts feature selection and classification simultaneously. Furthermore, the accuracy of BPSVM is better than the other two methods in most cases and it is comparable with the other two methods when the accuracy of BPSVM is lower than the other two methods.

Results on Phosphorylation Data Sets

Table 5 shows the average F scores on 4 data sets. All methods performs worse on the four data sets, all of the average F scores are lower than 0.5. This results indicate that the problem of sites prediction is not suitable to be seeing as a PU learning problem. Since BPSVC can solve the imbalanced binary classification problem, we also conduct the numerical experiments on the pure data sets by comparing ensemble SVM (ESVM), BSVM and BPSVC. To compare their performance, the F-score in binary classification is used and is defined as $F = \frac{2SnSp}{Sn+Sp}$, where $Sn = \frac{TP}{TP+FN}$,$Sp = \frac{TN}{TN+FP}$. The results are shown in Table 6. We can see that for the imbalanced binary classification problem, BPSVC performs best. BPSVC achieves the highest F scores and can select relevant features. While, SVM and BSVM can not select relevant features..

4 Conclusion

In this paper, we propose the biased p-norm support vector machine (BPSVM), which is an entirely new feature selection method for PU learning problem. BPSVM gives larger weights to positive examples and smaller weights to unlabeled examples. The effective algorithm for solving BPSVM is also conducted. Numerical results show that BPSVM performs better compared with other PU learning methods. In addition, BPSVM can be used in unbalanced binary classification problem and it performs a little better than other methods.

Acknowledgments

This paper is supported by National Natural Science Foundation of China (No. 11271361, No. 11201480).

References

[1] Fung, G.P.C., Yu, J.X., Lu, H., Yu P.S., Text Classification without Negative Examples Revisit. IEEE Transactions on Knowledge and Data Engineering, 2006,18(1), pp.6–20.

[2] Joachims, T., Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In: Proceedings of the 10th European Conference on Machine Learning, Chemnitz, Germany ,1998, pp. 137–142.

[3] Lee, W.S., Liu, B.: Learning with Positive and Unlabeled Examples Using Weighted Logistic Regression. In: Proceedings of the 20th International Conference on Machine Learning, Washington DC, United States,2003, pp. 448–455.

[4] Li, X., Liu, B.: Learning to Classify Text Using Positive and Unlabeled Data. In: Proceedings of the 18th International Joint Conference on Artificial Intelligence, Acapulco, Mexico, 2003, pp. 587–594.

[5] Li, X., Liu, B., Ng, S.: Negative Training Data can be Harmful to Text Classification. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, Massachusetts, United States, 2010, pp. 218–228.

[6] Liu, B., Lee, W.S., Yu, P.S., Li, X.: Partially Supervised Classification of Text Documents. In: Proceedings of the 19th International Conference on Machine Learning, Sydney, Australia, 2002, pp. 387–394.

[7] Liu, B., Dai, Y., Li, X., Lee, W.S., Yu, P.S.: Building Text Classifiers Using Positive and Unlabeled Examples. In: Proceedings of the 3rd IEEE International Conference on Data Mining, Florida, United States, 2003, pp. 179–188.

[8] Nigam, K., McCallum, A.K., Thrun, S.: Learning to Classify Text from Labeled and Unlabeled Documents. In: Proceedings of the 15th National Conference on Artificial Intelligence, AAAI Press, United States, 1998, pp. 792–799.

[9] Nigam, K., McCallum, A.K., Thrun, S., Mitchell, T.: Text Classification from Labeled and Unlabeled Documents Using EM. Mach. Learn. 2000, 39, pp.103–134.

[10] Sebastiani, F.: Machine Learning in Automated Text Categorization. ACM Computer Surveys. 2002, 34, pp.1–47.

[11] . Yu, H., Han, J., Chang, K.C.C.: PEBL: Positive Example-Based learning for web page classification using SVM. In: Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, United States, 2002, pp. 239–248.

[12] Manevitz, L., Yousef, M.: One-class SVMs for document classification. J. Mach. Learn. Res. 2001, 2, pp. 139–154.

[13] Chen XJ, Xu FM, Ye YY Lower bound theory of nonzero entries in solutions of l_2-l_p minimization. http://www.polyu.edu.hk/ama/staff/xjchen/ cxy-final.pdf, 2009.

[14] Bruckstein AM, Donoho DL, Elad M From sparse sulutions of systems of equations to sparse modeling of signals and images. SIAM Reviewer, 2009, 51,pp. 34-81.

[15] Fan J, Li R Varible selection via nonconcave penalized likelihood and its oracle properties. J Amer Statis Assoc. 2001,96,pp. 1348-1360.

[16] Xu Z, Zhang H, Wang Y, Chang X $L_{\frac{1}{2}}$ regularizer. Science in China Series F-InfSci, 2009, 52, pp. 1-9.

[17] Chen WJ, Tian YJ, L_p-norm proximal support vector machine and its applications. Procedia Computer Science, ICCS 2010, 1(1), pp. 2417-2423.

[18] Tian YJ , Yu J, Chen WJ l_p-norm support vector machine with CCCP. In Proc. the 7th FSKD, 2010, pp.1560-1564.

[19] Tan JY, Zhang CH, Deng NY , Cancer related gene identification via p-norm support vector machine. The 4th International Conference on Computational Systems Biology, 2010, pp.101-108.

[20] Jun-Yan Tan, Zhi-Qiang Zhang, Ling Zhen, Chun-Hua Zhang*, Nai-Yang Deng*, Adaptive feature selection via a new version of support vector machine. Neural Computing and Applications, doi: 10.1007/s00521-012-1018-y, 2012.

[21] Chun-Hua Zhang, Yuan-Hai Shao, Jun-Yan Tan*, Nai-Yang Deng, A mixed-norm linear support vector machine. Neural Computing and Applications, doi: 10.1007/s00521-012-1166-0, 2012.

[22] Ke T., Tan J.Y. Building High-performance Classifiers Using Positive and Unlabeled Examples for Text Classification

[23] V. N.Vapnik. Statistical learning theory[M]. New York:Wiley, 1998.

[24] V. N.Vapnik. The Nature of Statistical Learning Theory. Springer, New York, second edition, 2000.

[25] B.Scholkopf, A.J.Smola. Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond, MIT Press, Cambridge,

MA, USA, 2001.
[26] N.Y. DengY.J. Tian, C.H. Zhang Support Vector Machines - Optimization Based Theory, Algorithms, and Extensions CRC Press (Taylor & Francis Group), 2013.
[27] A.L. Yuille., A.Rangarajan, The concave-convex procedure. Neural computation,2003,15(4) pp.915-936.
[28] http://www.research.att.com/ lewis/reuters21578.

Table 1: Statistics of Data sets

Data set	l_+/l_-	No.of features	Data set	l_+/l_-	No.of features
R_1	799/18407	26214	R_8	990/17856	26214
R_2	973/17873	26214	R_9	996/17850	26214
R_3	985/17861	26214	R_{10}	994/17852	26214
R_4	982/17864	26214	CDK	93/2353	273
R_5	963/17883	26214	CDK1	144/5361	273
R_6	988/17858	26214	CDK2	67/1763	273
R_7	975/17871	26214	CK2	215/4710	273

The abbreviations $R_1, R_2, \cdots, R_{10}$ refer to ten data sets in Reuters corpus, l_+ is the number of positive examples, l_- is the number of unlabeled or negative examples

Table 2: Average F Scores on Reuters collection

Class	EBSVM	BSVM	BPSVM	EBSVM	BSVM	BPSVM
		0.7			0.3	
R_1	0.984	0.981	0.969	0.983	0.959	0.963
R_2	0.957	0.952	0.950	0.939	0.897	0.922
R_3	0.934	0.904	0.930	0.857	0.814	0.859
R_4	0.915	0.915	0.929	0.882	0.779	0.871
R_5	0.863	0.863	0.865	0.759	0.792	0.842
R_6	0.854	0.854	0.903	0.836	0.814	0.853
R_7	0.846	0.794	0.909	0.576	0.592	0.698
R_8	0.956	0.956	0.971	0.955	0.763	0.958
R_9	0.985	0.937	0.985	0.937	0.802	0.985
R_{10}	0.909	0.909	0.837	0.782	0.749	0.837
Mean	0.921	0.907	0.925	0.850	0.796	0.879

The abbreviations $R_1, R_2, \cdots, R_{10}$ refer to ten data sets in Reuters corpus, EBSVM is the extended biased SVM, BSVM is biased SVM and BPSVM is the biased p-norm SVM.

Table 3: Average over-all F Scores on Reuters collection

Class	EBSVM	BSVM	BPSVM	Class	EBSVM	BSVM	BPSVM
R_1	0.984	0.970	0.966	R_6	0.845	0.834	0.878
R_2	0.948	0.925	0.936	R_7	0.711	0.693	0.803
R_3	0.895	0.859	0.895	R_8	0.956	0.860	0.965
R_4	0.899	0.847	0.900	R_9	0.961	0.870	0.985
R_5	0.811	0.828	0.854	R_{10}	0.846	0.829	0.837

The abbreviations $R_1, R_2, \cdots, R_{10}$ refer to ten data sets in Reuters corpus, EBSVM is the extended biased SVM, BSVM is biased SVM and BPSVM is the biased p-norm SVM.

Table 4: Average Selected features by BPSVM on Reuters collection

Class	R_1	R_2	R_3	R_4	R_5	R_6	R_7	R_8	R_9	R_{10}
Features	561	442.2	392.2	267	779.7	552.5	644	289.5	305	145

The abbreviations $R_1, R_2, \cdots, R_{10}$ refer to ten data sets in Reuters corpus.

Table 5: Average F Scores on Phosphorylation data sets

Data	CDK		CDK1		CDK2		CK2	
	0.3	0.7	0.3	0.7	0.3	0.7	0.3	0.7
ESVM	0.450	0.461	0.381	0.381	0.351	0.391	0.392	0.395
BSVM	0.446	0.415	0.364	0.368	0.408	0.376	0.398	0.374
BPSVM	0.425	0.473	0.368	0.391	0.336	0.383	0.353	0.331

The abbreviation ESVM is the ensemble SVM, BSVM is biased SVM and BPSVM is the biased p-norm SVM.

Table 6: Performance on Pure data sets

Data	CDK		CDK1		CDK2		CK2	
	F	No.features	F	No.features	0F	No.features	F	No.features
ESVM	0.921	all	0.850	all	0.926	all	0.836	all
BSVM	0.934	all	0.950	all	0.941	all	0.842	all
BPSVM	0.941	34	0.954	147	0.948	18	0.872	185

The abbreviation ESVM is the ensemble SVM, BSVM is biased SVM and BPSVM is the biased p-norm SVM.