Journal of Theoretical and Applied Information Technology
30th June 2013. Vol. 52 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

LP-NORM SVM BASED ON REWEIGHED MINIMIZATION WITH POSITIVE DAMPED ITEM

[1]HUANG HAILONG, [1]LIU JIANWEI, [2]LIU ZEYU , [1]ZUO XIN

[1]Research Institute of Automation, China University of Petroleum, Beijing 102249, China

E-mail: liujw@cup.edu.cn,huanghailong3520@163.com

[2]Computer Science and Technology Institute Jilin University,Jilin,130012,China

E-mail:2275045480@qq.com

ABSTRACT

A new kind of reweighed Lp-norm SVM, which solves simultaneously two major problems pattern classification and feature selection, based on positive damped item is proposed. The proposed algorithms attempts to select important features among the originally given plausible features, while maintaining the minimum error rate. The resulting value of variable p is not only related to the classification error rate but also connected to the degree of importance of feature. A convergence proof of this reweighed procedure is included and an efficient stopping criterion is employed. Different sets of experiments are conducted on the classification and feature selection tasks, and results compared with L2-norm SVM, and L1-norm SVM show that, our Lp-norm SVM algorithm is superior to these algorithm on both artificial datasets and real-world problems of analyzing DNA microarray data.

Keywords: *Reweighed Minimization, P Norm, Feature Selection, Prediction Error Rate*

1. INTRODUCTION

Supported Vector Machine (SVM)[1], proposed by Vapnik, is based on the popular theory of statistical learning theory. SVM fused many key technologies in the field of machine learning, including the largest margin, convex quadratic programming, Mercer kernel theory for nonlinear mapping and slack variable, etc. It can deal with the problem of high dimension but small samples effectively. And as it uses L2-norm as regularization term, traditional SVM problem is called L2-norm SVM, let $(x_i, y_i)_{i=1}^m \in R^n \times \{1,-1\}$ is the training dataset, L2-norm SVM problem is

$$\begin{aligned} &min\ \frac{1}{2}\|w\|_2^2 \\ &\text{s.t.} y_i\left(\langle w,x_i\rangle + b\right) \geq 1, i = 1,2,...,m \end{aligned} \tag{1}$$

where, w and b are weight and the bias of the discrimination function. L2-norm SVM algorithm can deal with the problem of high dimension but small samples effectively, but can't get sparse solutions, a more interesting alternative is to solve the following problem

$$\begin{aligned} &min\ \|w\|_0 \\ &\text{s.t. } y_i\left(\langle w,x_i\rangle + b\right) \geq 1, i = 1,2,...,m \end{aligned} \tag{2}$$

In (2), $\|w\|_0 = \sum_{j=1}^{n} 1_{\{w_j \neq 0\}}(w_j)$ is the L0-norm of a vector, i.e., the number of its non-zero coefficients, $1_A(x)$ is indicator function or a characteristic function. Problem (2) can get sparse solutions and it is called L0-norm SVM [2]. Unfortunately, the only known method to solve the equations (2) exactly is combinatorial, thus NP-complex [3]. Some researchers use L1-norm to take place of L0-norm and get the approximate problem of (2)

$$\begin{aligned} &min\ \|w\|_1 \\ &\text{s.t.} y_i\left(\langle w,x_i\rangle + b\right) \geq 1, i = 1,2,...,m \end{aligned} \tag{3}$$

Problem (3) is often called L1-norm SVM [4], and it is a convex problem which can be solved effectively by linear programming [5]. L1-norm is widely used in the filed of compressive sensing.

As above discussed about problem (2), it is a nonconvex problem, which is difficult to solve effectively. However, recently, in area of signal processing, Karthik Mohan has successfully solved the problem of matrix rank minimization using iterative reweighed algorithm for Lp-norm, where the range of p is $0 \leq p \leq 2$ [10]. He creatively introduced a positive parameter r, which amplitude is gradually decreasing to zero, to the object function, and this can avoid the singular during the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

iterative calculations. As r is gradually damped in iteration procedure, we call this r as positive damped item. This has already been used in Lp-norm least-squares problem [11].

Roughly speaking, the traditional classification models usually take p value as 2, 1 or 0. However, some classification models which value of p not restricted to $\{0,1,2\}$ have also been proposed, and it has already been applied, for example, facing reorganization[9], where $0 < p < 1$, and solving the linear underdetermined system[5], where, $0 < p \le 1$.

We argue that the above three different models can get a unified form, we use Lp-norm to replace the L2-norm, L0- norm and L1-norm in (1),(2) and (3), furthermore, we expand the selected range from 0 to 2, i.e., $0 \le p \le 2$,we proposed to solve below problem

$$\begin{aligned} &min\ \frac{1}{2}\|w\|_p^p \\ &\text{s.t. } y_i\left(\langle w,x_i\rangle + b\right) \ge 1, i = 1,2,...,m \end{aligned} \tag{4}$$

Unlike previous L2 norm, L0 norm and L1 norm regularization SVM problem, in problem (4), the value of p is not determined in advance, but as unsolved variable to be determined by data.

The paper is organized as follows. In section 2, the proposed reweighed Lp-norm SVM based on positive damped item is discussed, and the proof of convergence of the proposed algorithm also can be found in the section. In section 3, different sets of experiments are conducted on the classification and feature selection tasks and some analysis and comparisons with others the state of the arts method on synthetical datasets and the real datasets are given. And section 4 concludes and presents some perspectives.

2. REWEIGHTED LP-NORM SVM BASED ON POSITIVE DAMPED ITEM

2.1 Reweighed Lp-norm SVM based on Positive Damped Item

In this section, we proposed the algorithm for Lp-norm SVM using iterative reweighted minimization methods with positive damped item, where $0 \le p \le 2$. We take relaxation method and introduce positive damped item into Lp-norm of problem (4) and obtain a relaxation of (4)

$$\begin{aligned} &min\ \frac{1}{2}\sum_{j=1}^{n}\left(|w_j| + r\right)^p + \frac{C}{2}\sum_i^m \xi_i^2 \\ &s.t.\ y_i\left(\langle w,x_i\rangle + b\right) \ge 1 - \xi_i \\ &\xi_i \ge 0, i = 1,2,...,m \end{aligned} \tag{5}$$

Now, we use Lagrange multipliers to find the optimal point of problem (5). We introduce Lagrange multiplier α and β into (5), then it can be derived that the associated dual is

$$\begin{aligned} L(\alpha,\beta) &= \frac{1}{2}\sum_{j=1}^{n}\left(|w_j| + r\right)^p + \frac{C}{2}\sum_i^m \xi_i^2 - \sum_i^m \beta_i\xi_i \\ &+ \sum_i^m \alpha_i\left[1 - \xi_i - y_i\left(\langle w,x_i\rangle + b\right)\right] \\ &= \frac{1}{2}\sum_{j=1}^{n}\left(|w_j| + r\right)^p + \frac{C}{2}\sum_i^m \xi_i^2 + \sum_i^m \alpha_i - \sum_i^m \alpha_i\xi_i \\ &- \sum_i^m \alpha_i y_i x_i^T w + \sum_i^m \alpha_i y_i b - \sum_i^m \beta_i\xi_i \\ &\approx \frac{1}{2}\sum_{j=1}^{n}\left(|w_j| + r\right)^{p-2}\left(|w_j| + r\right)^2 + \frac{C}{2}\sum_i^m \xi_i^2 + \sum_i^m \alpha_i \\ &- \sum_i^m \alpha_i\xi_i - \sum_i^m \alpha_i y_i x_i^T w + \sum_i^m \alpha_i y_i b - \sum_i^m \beta_i\xi \\ &\approx \frac{1}{2}\sum_{j=1}^{n}\left(|w_j| + r\right)^{p-2} w_j^2 + \frac{C}{2}\sum_i^m \xi_i^2 + \sum_i^m \alpha_i - \sum_i^m \alpha_i\xi_i \\ &- \sum_i^m \alpha_i y_i x_i^T w + \sum_i^m \alpha_i y_i b - \sum_i^m \beta_i\xi \end{aligned}$$

The above can also be written in the matrix form

$$\begin{aligned} L &= \frac{1}{2}w^T \Lambda w - \alpha^T \Delta x w + \frac{C}{2}\xi^T\xi - \beta^T\xi - \alpha^T\xi \\ &+ \sum_i^m \alpha_i + \sum_i^m \alpha_i y_i b \end{aligned} \tag{6}$$

where,

$$\Lambda = \begin{bmatrix} \left(|w_1| + r\right)^{p-2} & & \\ & \left(|w_2| + r\right)^{p-2} & \\ & & \ddots \\ & & \left(|w_n| + r\right)^{p-2} \end{bmatrix}$$

$$\Delta = \begin{bmatrix} y_1 & & \\ & y_2 & \\ & & \ddots \\ & & y_m \end{bmatrix}$$

and which we minimize w.r.t α, β and ξ_i. Setting the respective derivatives to zero, we get

Journal of Theoretical and Applied Information Technology
30th June 2013. Vol. 52 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

$$\begin{cases} \dfrac{\partial L}{\partial w} = w^T \Lambda - \alpha^T \Delta x = 0 \quad \Rightarrow w^T = \alpha^T \Delta x \Lambda^{-1} \\ \dfrac{\partial L}{\partial b} = \sum_i^m \alpha_i y_i = 0 \\ \dfrac{\partial L}{\partial \xi} = C\xi^T - \alpha^T - \beta^T = 0 \Rightarrow \xi = \dfrac{1}{C}(\alpha + \beta) \end{cases} \quad (7)$$

By substituting (7) into (6), we obtain the Lagrange (Wolfe) dual objective function

$$\min_{\alpha,\beta} \frac{1}{2}\alpha^T \Delta x \Lambda^{-1} x \Delta \Delta + \frac{1}{C}\left(\alpha^T\alpha + \beta^T\beta + \alpha^T\beta + \beta^T\alpha\right) - \sum_i^m \alpha_i \quad (8)$$

Note that the two unsolved variables are α and β in (8), and Λ can be calculated by w and r , therefore before we solve (8), w and r should be known in advance. The role of r is to avoid Λ becoming a singular matrix when some weights of w are close to zero, also $\{r_k\}$ should be non-increasing sequence, r_{k+1} should be smaller than r_k , which make sure that the optimal solution of final problem approaches asymptotically to the ones of original problem, for example, we may select $r = e^{-\lambda k}$ or $r = \lambda / k$ [6], where λ is a constant parameter and k is the number of iteration.

The algorithm we proposed is shown as below:

Algorithm1: Reweighed Lp-norm SVM

Input: $(x_i, y_i)_{i=1}^m$, $w^{(0)}$, C , p , absolute error bound ε , maximum number of iteration N

Output: model vector w

1. for p=0, 0.2,0.4,0.6,⋯,2 (outer iteration)
2. for k from 1 to N (inner iteration)
3. using the initial $w^{(0)}$ to calculate Λ^k and Δ
4. solve problem (8), and denotes its solution as α^{k+1} and β^{k+1};
5. check whether the criterion is satisfied, if so, break, otherwise, Step to 6;
6. calculate w^{k+1}, Λ^{k+1} through (7)

2.2 Convergence

In this section, we give proof of the convergence of our algorithm.

Theorem1. Suppose that $r > 0, 0 \le p \le 2$, given the problem represented by (5), $\{\alpha^k\}$ be the sequence generated by reweighed Lp-norm SVM algorithm, the sequence $\{\alpha^k\}$ is bounded.

Proof. As (5) shows that, the Lagrange multiplier β is related with constrains $\xi_i \ge 0$, and $\beta \ge 0$ .From (8), we see that $\beta = 0$. So the problem (8) can be transformed into (9)

$$\min f(\alpha, w, r, p) = \frac{1}{2}\alpha^T \Delta x \Lambda^{-1} x \Delta \Delta + \frac{1}{C}\alpha^T\alpha - \sum_i^m \alpha_i \quad (9)$$

Then, by algorithm1, we can derive the following two iteration formula

$$\alpha^{k+1} = \arg\min_{\alpha} f(\alpha, w^k, r^k, p) \quad (10)$$

$$w^{k+1} = [\alpha^k \Delta x g(w^k, r^k, p)]^T \quad (11)$$

where,

$$g(w^k, r^k, p) = \begin{bmatrix} \left(\left|w_1^k\right| + r^k\right)^{p-2} & & & \\ & \left(\left|w_2^k\right| + r^k\right)^{p-2} & & \\ & & \ddots & \\ & & & \left(\left|w_n^k\right| + r^k\right)^{p-2} \end{bmatrix}^{-1}$$

then we have

$$\begin{aligned} f(\alpha^{k+1}, w^{k+1}, r^{k+1}, p) &\le f(\alpha^{k+1}, w^k, r^{k+1}, p) \\ &\le f(\alpha^{k+1}, w^k, r^k, p) \\ &\le f(\alpha^k, w^k, r^k, p) \end{aligned} \quad (12)$$

Inequality (12) shows that the iterative process of algorithm1 produces a convergent sequence $\{\alpha^k\}$, as long as $r_{\min} = \lim_{k\to\infty} r^k > 0$, due to we suppose $r > 0$, $\{r_k\}$ is non-increasing sequence, this condition is obviously satisfied, therefore we arrive that the sequence $\{\alpha^k\}$ of iteration of Lp-norm SVM is bounded.

Theorem 2. The cluster point of the sequence $\{\alpha^k\}$ is a stationary point of (8).

Proof. Suppose that the statement is not true. Let $\tilde{\alpha}$ be a cluster point of $\{\alpha^k\}$, but not a stationary point. By the definition of cluster point, there exists a subsequence $\{\alpha^{n_i}\}$ of $\{\alpha^k\}$ converging to $\tilde{\alpha}$. Furthermore, subsequence $\{\alpha^{n_i+1}\}$ is also convergent and we denote its limit

Journal of Theoretical and Applied Information Technology
30th June 2013. Vol. 52 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

by $\hat{\alpha}$. Then, α^{n_i+1} is the minimum of the following problem

$$min\ f(\alpha, w^{n_i}, r^{n_i}, p) \quad (13)$$

Suppose that problem (8) has a solution set S, and then α^{n_i+1} satisfies $\alpha^{n_i+1} w^{n_i} \in S$.

Taking limits to α^{n_i+1}, we see that

$$\tilde{a}\tilde{w} \in S \quad (14)$$

Equation (14) shows that $\hat{\alpha}$ is the solution of below problem

$$min\ f(\alpha, \tilde{w}, r_{min}, p) \quad (15)$$

By assumption, $\tilde{\alpha}$ is not a stationary point of (8), which implies that

$$f(\hat{\alpha}, \hat{w}, r_{min}, p) < f(\tilde{\alpha}, \tilde{w}, r_{min}, p) \quad (16)$$

However, as Theory 1 shows that $f(\alpha^k, w^k, r^k, p)$ is convergent, thus

$$\begin{aligned} &lim\ f(\alpha^i, w^i, r^i, p) = lim\ f(\alpha^{n_i}, w^{n_i}, r^{n_i}, p) \\ &= lim\ f(\tilde{a}, \tilde{w}, r_{min}, p) = lim\ f(\alpha^{n_i+1}, w^{n_i+1}, r^{n_i+1}, p) \\ &= lim\ f(\hat{a}, \hat{w}, r_{min}, p) \end{aligned} \quad (17)$$

which contradicts with (16). Hence, the cluster point of the sequence $\{\alpha^k\}$ is also a stationary point of (8).

Algorithm1 is based on the technique of reweighed iterations and the idea of positive damped item. We first set an initial value for w and r , and solve the problem (8). In the next iteration we use the last w and a new r to solve problem (8). Theorem 1 proves that the sequence $\{\alpha^k\}$ of iteration is bounded and Theorem 2 proves that the cluster point of the sequence $\{\alpha^k\}$ is a stationary point of (8). Therefore, we can get the stable solution of an approximate problem of (8).

3. EXPERIMENTS

In this section, we give the results of our experiments both on artificial datasets and real datasets.

3.1. Experiments on Artificial Datasets

The method to generate artificial datasets follows the following principles:

1) Each sample should have the same size of feature dimensions and noise dimensions;
2) The labels are determined by feature dimensions, and have nothing to do with noise dimensions;
3) All samples are independent;
4) Labels are assigned with equal probabilities.

According to the above four principles, we give the steps to generate dataset:

1) Generate label set *y*. Using standard normal distribution whose mean is 0 and variance is 1 to generate randomly m (m is the number of samples) numbers. When the number is greater than 0 then the label of the corresponding samples is 1, when it is less than 0, the label of the corresponding samples is -1;
2) Generate samples sets *X*. The feature vectors are divided into two groups, the first *fn*/2 (fn denotes feature number) features follow the distribution $X(i)=yN(i,1)$, and the other *fn*/2 features follow the distribution $X(i)=N(0,1)$;
3) Generate noise components of samples sets. The noise components obey the distribution $N(0,10)$.

In order to express clearly our artificial dataset, we define the following notions.

S: artificial dataset
M: number of samples
N: dimension of a sample
NF: Number of Features
NN: Number of Noises
DF: Distribution of Features
DN: Distribution of Noises

For designing experiment, we generate two types of artificial datasets. In the first type of datasets, we set *N*, *NF*, *NN as constants*, and change value of *M*, the details shown in table1. In the second type of datasets, we set *M* and *N as constants*, change simultaneously value of *NF* and *NN*, shown in table2. *DF* obeys normal distribution, while *DN* is white noise.

Figures should be labeled with "Figure" and tables with "Table" and should be numbered sequentially, for example, Figure 1, Figure 2 and so on (refer to table 1 and figure 1). The figure numbers and titles should be placed below the figures, and the table numbers and titles should be placed on top of the tables. The title should be placed in the middle of the page between the left and right margins. Tables, illustrations and the corresponding text should be placed on the same page as far as possible if too large they can be placed in singly column format after text. Otherwise they may be placed on the immediate following page. If its size should be smaller than the type area they can be placed after references in singly column format and referenced in text

Table 1: The First Type Artificial Datasets.

NO.	*M*	*N*	*NF*	*NN*
1	150	100	50	50
2	100	100	50	50
3	50	100	50	50
4	25	100	50	50

Journal of Theoretical and Applied Information Technology
30th June 2013. Vol. 52 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 2: The Second Type Artificial Datasets.

NO.	*M*	*N*	*NF*	*NN*
1	50	100	50	50
2	50	100	40	60
3	50	100	30	70
4	50	100	20	80
5	50	100	10	90

The experiment results of two artificial datasets are shown in Fig.1 and Fig.2.

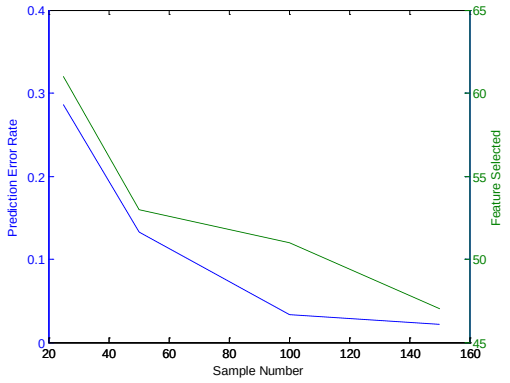

Figure 1: Results Of The First Type Artificial Datasets

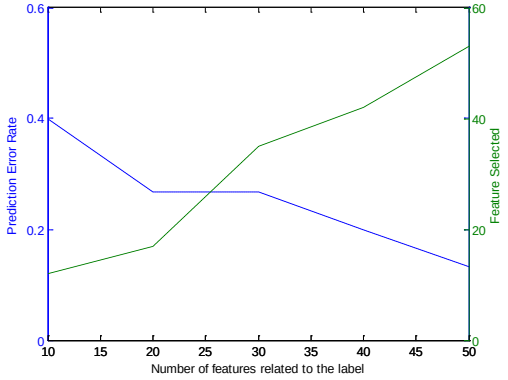

Figure 2: Results Of The Second Type Artificial Datasets

We can draw the following conclusions from the two experiments: on first type artificial datasets, when *N*, *NF* and *NN* is taken fixed value , prediction error rate decreases with the change of *M*, furthermore, the number of selected features in the four datasets are 47,51,53 and 61, and approximately the same size as corresponding one of *NF*. For second type artificial datasets, the value of *M and N* is predefined as 50 and100 respectively; the value of *NF and NN* is changed between 10-50 and 90-50 respectively. It is well known that with the increase of the number of samples, classification accuracy will rise, and Figure 2 verified this point of view. We also see that with the increase in the value of *NF*, the number of selected features increase correspondingly.

3.2. The experimental result on Real Datasets

In this section, we compare the proposed Lp-norm SVM algorithm with the standard L2-norm SVM and L1-norm SVM on four cancer datasets.

The four datasets that we select are Colon, Bladder, Melanoma and Lymphoma. In the Colon cancer problem [12], 62 samples contain 40 colon cancer samples and 22 normal ones, and the dimension is 2000. Bladder cancer datasets contain 57 samples, each sample has 2215 features [13], among 12 samples belonging to Grade 1, and 45 samples belonging to Grade 2 and 3. Melanoma dataset has 78 samples, of which the 35 samples were taken from the patients who contributed to the development of the tumor metastasis in 24 hours, and 43 samples had no contributions [13]. Lymphoma dataset [12] has 96 samples. 61 samples of this dataset are in classes “DLCL”, “FL” or “CLL” (malignant) and 35 samples are labeled normal. Details of each dataset can be seen in Table 3.

Table 3: Feature Of The Selected Cancer Datasets.

dataset	*N*	*M*	Label 1	Label -1
Colon	2000	62	22	40
Bladder	2215	57	12	45
Melanoma	3750	78	35	43
Lymphoma	4026	96	61	35

In our experiments, the order p of regularization is adjustable parameter, so we would like to find a suitable p for four cancer datasets. The results are shown in Fig.3 and Fig.4.

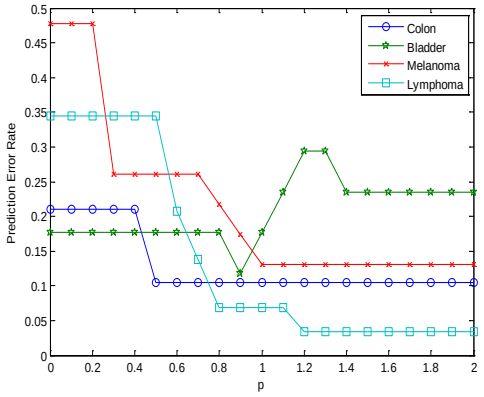

Figure 3: Relationship Between Prediction Error rate and p

Journal of Theoretical and Applied Information Technology
30th June 2013. Vol. 52 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

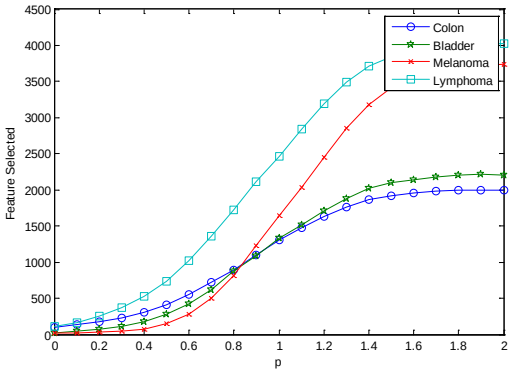

Figure 3: Relationship Between Feature selected and p

As we can see from Fig.3, the larger p is, the lower prediction error rate we get, and more features are selected. When p=2, the problem that we solve is equal to problem (1), which has better classification performance. When p = 0, the problem degenerate into problem (2), which can only select a part features among all the features. Therefore, the prediction error rate and number of selected feature are contradictory from each other. We try to maintain our prediction error rate low and select as few features as possible. Taking Colon dataset as an example, when p equals to 0.6, the prediction error rate is 10.35% and 275 out of 2000 features are selected. When p increases, the prediction error rate keeps the same level of p=0.6, but the number of selected features increases. so that, for Colon dataset, we set p=0.6, and use the selected 275 features to construct classification model. As for the other three datasets, we set p=1.0 for Bladder dataset, p=1.1 for Melanoma dataset, and p=1.3 for Lymphoma dataset, and 1029 out of 2215, 1643 out of 3750, and 3492 out of 4026 features are selected respectively. The results are shown in Table 4.

Table 4: Prediction Error Rate And Number of Selected Features In Our Experiments.

Dataset	Error	features	*N*
Colon	10.35%	275	2000
Bladder	11.76%	1092	2215
Melanoma	13.04%	1643	3750
Lymphoma	3.45%	3492	4026

Below, we compare our results with L2-norm SVM and L1-norm SVM. L2-norm SVM algorithm selects all features to train a classifier, so that it can not select features which are more related with the classifier performance. L1-norm SVM is an approximate algorithm to L_0-SVM, and it can select features by sparse model vector. The comparisons are shown in Table 5 and Table 6.

Table 5: Comparisons Of Prediction Error Rate With Lp-norm SVM, L2-norm SVM And L1-norm SVM.

Dataset	Lp-norm SVM	L2-norm SVM	L1-norm SVM
Colon	10.35%	12.20%	16.39%
Bladder	11.76%	23.60%	26.41%
Melanoma	13.04%	14.71%	14.11%
Lymphoma	3.45%	7.39%	6.10%

Table 6: Comparisons Of Number Of Selected features With Lp-norm SVM, L2-norm SVM And L1-norm SVM.

Dataset	Lp-norm SVM	L2-norm SVM	L1-norm SVM
Colon	275	-	341
Bladder	1092	-	921
Melanoma	1643	-	1491
Lymphoma	3492	-	3112

Table 5 and Table 6 show that the prediction error rates of the proposed Lp-norm SVM algorithm on these four cancer datasets are competitive with L2-norm SVM and L1-norm SVM. And Lp-norm SVM can also select more related features with pathogenesis.

4. CONCLUSIONS

In this article, we first review some classical classification models, L2-norm SVM, L_0-SVM and L1-norm SVM. We unify the three classification models and proposed framework of Lp-norm SVM classifier, where $0 \le p \le 2$, based on reweighed minimization technique and a positive damped item, we also proposed effective solved numerical method for Lp-norm SVM classifier. We give proof of convergence of the proposed algorithm. Compared with L2-norm SVM and L1-norm SVM both on the classification and feature selection tasks, experimental results show that our Lp-norm SVM algorithm is superior on both artificial datasets and real problems of analyzing DNA microarray data.

REFRENCES:

[1] H Vladimir N.Vapnik. "Statistical Learning Theory". Wiley, New York, 1998.

[2] T. Blumensath, M. E. Davies. "Iterative hard thresholding for compressed sensing". *Applied and Computational Harmonic Analysis*, Vol.27, No.3, 2009, pp. 265-274.

[3] Edoardo Amaldi, Viggo Kann. "On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems". *Theoretical Computer Science*, Vol.209,No.1,1998, pp. 237-260.

[4] A. Beck and M. Teboulle. "A fast iterative shrinkage thresholding algorithm for linear inverse problems". *Journal on Imaging Sciences*, Vol. 2, No. 1, pp. 183-202.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[5] S. Boyd and L. Vandenberghe, "Convex Optimization", Cambridge University Press, 2004.

[6] S. Foucart and M. Lai. "Sparsest solutions of underdetermined linear systems via lp minimization for $0<p \leq 1$". *Applied and Computational Harmonic Analysis*, Vol.26, 2009, pp.395-407.

[7] Yiwei Chen and Chihjen Lin. "Combining SVMs with various feature selection strategies.Feature extraction", foundations and applications. Springer, 2006.

[8] Guyon, I., Weston, J., Barnhill, S., Vapnik, V. "Gene selection for cancer classification using support vector machines". *Machine Learning* Vol. 46, 2002, pp.389-422.

[9] Song Guo, Zhan Wang, Qiuqi Ruan. "Enhancing sparsity via lp($0<p<1$) minimization for robust face recognition", *Neurocomputing* Vol. 99, 2013, pp.592-602.

[10] Karthik Mohan, Maryam Fazel. "Iterative reweighted least squares for matrix rank minimization", *Communication, Control, and Computing* (Allerton), 2010 48th Annual Allerton Conference on, pp.653-661.

[11] X. Chen and W. Zhou. "Convergence of reweighted l1 minimization algorithms and unique solution of truncated lp minimization". Preprint, April 2010.

[12] Jason Weston, Andre Elisseeff, Bernhard Scholkopf. "Use of Zero-Norm with Linear Models and Kerkel Methods". *Journal of Machine Learning Research*, Vol. 3, 2003, pp.1439-1461.

[13] Franck Rapaport, Emmanuel Barillot, Jean-Philippe Vert. "Classification of arrayCGH data using fused SVM". *Bioninformatics*. Vol. 24,2008, pp.375-382.