The Ninth International Symposium on Operations Research and Its Applications (ISORA'10)
Chengdu-Jiuzhaigou, China, August 19–23, 2010

Feature Selection in Multi-instance Learning*

Chun-Hua Zhang[1] Jun-Yan Tan[2,†] Nai-Yang Deng[2,‡]

[1]Information School, Renmin University of China, Beijing, China, 100872
[2]College of Science, China Agricultural University, Beijing, China,100083

Abstract This paper focuses on the feature selection in multi-instance learning. A new version of support vector machine named p-MISVM is proposed. In the p-MISVM model, the problem needs to be solved is non-differentiable and non-convex. By using the constrained concave-convex procedure (CCCP), a linearization algorithm is presented that solves a succession of fast linear programs that converges to a local optimal solution. Furthermore, the lower bounds for the absolute value of nonzero components in every local optimal solution is established, which can eliminate zero components in any numerical solution. The numerical experiments show that the p-MISVM is effective in selecting relevant features, compared with the popular MICA.

Keywords Support vector machine; feature selection; p-norm; multi-instance learning

1 Introduction

Feature selection is very important in many applications of data mining. By restricting the input space to a small subset of input variables, it has obvious benefits in terms of data storage, computational requirements, and cost of future data collection. This paper focuses on feature section in multi-instance learning via a new version of support vector machine (SVM).

Multi-instance learning (MIL) is a growing field of research in data mining. In the MIL problem, the training set is composed of many bags, each involves in many instances. A bag is positively labeled if it contains at least one positive instance; otherwise it is labeled as a negative bag. The task is to find some decision function from the training set for correctly labeling unseen bag.

MIL problem was first introduced by Dietterich et al.[1] in drug activity prediction. So far, MIL has been applied to many fields such as image retrieval ([2]), face detection ([3]), scene classification, text categorization, etc and is often found to be superior than a conventional supervised learning approaches. ([4]) proposed a framework called Diverse Density algorithm. Since then various variants of standard single instance learning algorithms like Boost-ing ([3], [5]), SVM ([2], [6]), Logistic Regression ([7]), nearest neighbor ([8]) etc. have been modified to adapt to the MIL problem.

*This work is supported by the Key Project of the National Natural Science Foundation of China (No. 10631070), the National Natural Science Foundation of China (No.10971223)
†Corresponding author. E-mail: tanjunyan0@126.com
‡Corresponding author. E-mail: dengnaiyang@cau.edu.cn

Based on the standard SVM, some methods including MI, mi [9], etc. have been proposed for the MIL problem. There are few works on feature selection in MIL. In [10], the MICA algorithm is introduced, which employs 1-norm, rather than 2-norm in MI and mi. Because the 1-norm SVM formulation is known to lead to sparse solutions ([11], [12]), MICA can get few features when a linear classifier is used. Recently, an effective method, named p-norm SVM ($0 < p < 1$), is proposed on feature selection in the standard classification problems in [13], which motivates us to apply it to the MIL problem. This paper proposes p-norm multi-instance SVM (p-MISVM), which replaces the 2-norm penalty by the p-norm ($0 < p < 1$) penalty in the objective function of the primal problem in the MI. The p-MISVM conducts feature selection and classification simultaneously. However, there are two difficulties in solving p-MISVM model: (i). It is impossible to solve the primal problem via its dual problem and the primal problem itself is hard to be solved, because it is neither differentiable nor convex; (ii). Feature selection needs to find the nonzero components of the solution to the primal problem. However, usually algorithms can only provide an approximate solution where nonzero components in the solution can not be identified theoretically.

Firstly, for the difficulty (i), by using the constrained concave-convex procedure (CCCP) ([14], [15]), a linearization algorithm is presented that solves a succession of fast linear programs that converges to a local optimal solution to the primal problem. Furthermore, for the difficulty (ii), the lower bounds for the absolute value of nonzero entries in every local optimal solution is established, which can eliminate zero entries in any numerical solution. Lastly, the performance of p-MISVM is illustrated on the simulated datasets.

Now we describe our notation. 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 $0 < p < 1$, 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.

This paper is organized as follows. In section 2, the p-MISVM for feature section is introduced. In section 3, the CCCP is proposed to solve p-MISVM model. In section 4, the absolute lower bounds of the local optimal solution is established. In section 5, numerical experiments are given to demonstrate the effectiveness of our method. We conclude this paper in section 6.

2 p-norm multi-instance support vector machine

For feature selection, p-MISVM is an embedded method in which training data are given to a learning machine, which returns a predictor and a subset of features on which it performs predictions. In fact, feature selection is performed in the process of learning.

Consider the multi-instance classification problem with the training set T is given by

$$\{(\mathscr{X}_1, y_1), \cdots, (\mathscr{X}_l, y_l)\}, \tag{1}$$

where $\mathscr{X}_i = \{x_{i1}, \cdots, x_{il_i}\}, x_{ij} \in R^n (i = 1, \cdots, l, j = 1, \cdots, l_i), y_i \in \{-1, 1\}$. Here, when $y_i = 1$, $\mathscr{X}_i$ is called as a positive bag and $(\mathscr{X}_i, y_i)$ implies that there exists at least one instance with positive label in $\mathscr{X}_i$; when $y_i = -1$, $\mathscr{X}_i$ is called as a negative bag and there exists no any instance with positive label in $\mathscr{X}_i$. The task is to find a function $g(x)$ such that

the label of any instance in R^n can be deduced by the decision function $f(x) = \mathrm{sgn}(g(x))$. For convenience, the training set (1) is represented as

$$\{(\mathscr{X}_1, y_1), \cdots, (\mathscr{X}_q, y_q), (x_{r+1}, y_{r+1}), \cdots, (x_{r+s}, y_{r+s})\}, \tag{2}$$

where $y_1 = \cdots = y_q = 1, y_{r+1} = \cdots = y_{r+s} = -1$, $(\mathscr{X}_i, 1)$ implies that there exists at least one instance with positive label and $(x_i, -1)$ implies that the label of the instance x_i is negative. All of instances in positive bags $\mathscr{X}_1, \cdots, \mathscr{X}_q$ are $x_1, \cdots, x_r$. $I(i)$ $(i = 1, \cdots, q)$ denotes the index set of instances in $\mathscr{X}_i$. The feature vector

$$g_i = ([x_1]_i, [x_2]_i, \cdots, [x_{r+s}]_i)^{\mathrm{T}}, (i = 1, \cdots, n) \tag{3}$$

denotes the values of i-th feature in all instances.

Suppose the decision function is given by $f(x) = \mathrm{sgn}((w \cdot x) + b)$, the p-MISVM solves the optimization problem:

$$\begin{array}{ll} \min\limits_{w,b,\xi} & \|w\|_p^p + C_1 \sum\limits_{i=1}^{q} \xi_i + C_2 \sum\limits_{i=r+1}^{r+s} \xi_i, \\ \text{s.t.} & \max\limits_{j \in I(i)} ((w \cdot x_j) + b) \geq 1 - \xi_i, i = 1, \cdots, q, \\ & (w \cdot x_i) + b \leq -1 + \xi_i, i = r+1, \cdots, r+s, \\ & \xi_i \geq 0, i = 1, \cdots, q, r+1, \cdots, r+s, \end{array} \tag{4}$$

where $C_1 (C_1 > 0)$, $C_2 (C_2 > 0)$ and $p (0 < p < 1)$ are parameters. Now we describe our new method such as following:

Algorithm 1. (p-MISVM)

(1) Given a training set (2); Select the parameters $C_1 (C_1 > 0), C_2 (C_2 > 0)$ and p $(0 < p < 1)$;

(2) Solve the optimization problem (4) and get its global solution (w^*, b^*, ξ^*);

(3) Select the feature set: $\{i | [w^*]_i \neq 0, (i = 1, \cdots, n)\}$;

(4) Construct the decision function $f(x) = \mathrm{sgn}(w^* \cdot x) + b^*)$.

Note that, in the Algorithm 1, there are two difficulties (i) and (ii) that have been addressed in Section 1, so the following sections will consider them respectively.

3 CCCP for the p-MISVM model

The constrained concave-convex procedure (CCCP) ([14], [15]) is an optimization tool for problems whose objective and constrained functions can be expressed as the differences of convex functions. Consider the following optimization problem:

$$\begin{array}{ll} \min_x & f_0(x) - g_0(x) \\ \text{s.t.} & f_i(x) - g_i(x) \leq c_i, i = 1, \cdots, m, \end{array} \tag{5}$$

where $f_i, g_i (i = 0, \cdots, m)$ are real-valued, convex and differentiable functions on R^n, and $c_i \in R$. Given an initial $x^{(0)}$, CCCP computes $x^{(t+1)}$ from $x^{(t)}$ by replacing $g_i(x)$ with its first-order Taylor expansion at $x^{(t)}$, and then setting $x^{(t+1)}$ to the solution of the following optimization problem:

$$\begin{array}{ll} \min_x & f_0(x) - [g_0(x^{(t)}) + \nabla g_0(x^{(t)})^{\top}(x - x^{(t)})] \\ \text{s.t.} & f_i(x) - [g_i(x^{(t)}) + \nabla g_i(x^{(t)})^{\top}(x - x^{(t)})] \leq c_i, i = 1, \cdots, m. \end{array} \tag{6}$$

Here, $\nabla g(\bar{x})$ is the gradient of the function g at $\bar{x}$. For non-smooth functions,the gradient should be replaced by the subgradient. It can be shown that CCCP converges to a local minimum solution of (5) in [15].

Consider the problem (4), we firstly introduce the variable $v = ([v]_1, \cdots, [v]_n)^\top$ to eliminate the absolute value from the objective function, which leads to the following equivalent problem:

$$\min_{w,b,\xi} \quad \|v\|_p^p + C_1 \sum_{i=1}^{q} \xi_i + C_2 \sum_{i=r+1}^{r+s} \xi_i, \tag{7}$$

$$\text{s.t.} \quad \max_{j \in I(i)} ((w \cdot x_j) + b) \geq 1 - \xi_i, i = 1, \cdots, q, \tag{8}$$

$$(w \cdot x_i) + b \leq -1 + \xi_i, i = r+1, \cdots, r+s, \tag{9}$$

$$\xi_i \geq 0, i = 1, \cdots, q, r+1, \cdots, r+s, \tag{10}$$

$$-v \leq w \leq v \tag{11}$$

where $\|v\|_p^p = [v]_1^p + \cdots + [v]_n^p$, due to the last constraint (11). Furthermore, we note that the objective function and the constraint functions in the problem (7)-(11) can be regarded as the differences of two convex functions. Hence, we can solve the problem (7)-(11) with CCCP. Note that $\max_{j \in I(i)} (w \cdot x_j)$ in (8) is convex, but a non-smooth function of w.To use the CCCP, we have to replace the gradient by the subgradients. It is easy to obtain that for $i = 1, \cdots, q$, the subgradients $\partial \max_{j \in I(i)} (w \cdot x_j) = \{\sum_{j \in I(i)} \beta_j x_j | \beta_j \in R, \beta_j \geq 0\}$, where $\beta_j = \begin{cases} = 0, & \text{if} (w \cdot x_j) \neq \max\limits_{j \in I(i)} (w \cdot x_j), \\ \geq 0, & \text{otherwise} \end{cases}$ with $\sum_{j \in I(i)} \beta_j = 1$. At the k-th iteration, denote the current w, b, ξ, v estimate and the corresponding β_j by $w^{(k)}, b^{(k)}, \xi^{(k)}, v^{(k)}$ and $\beta_j^{(k)}$, respectively. In the experiments, we initialized $\beta_j^{(0)} = \frac{1}{|I(i)|}$, for $j \in I(i)$. For convenience, we pick the subgradient with: $\beta_j^{(k)} = \begin{cases} 1, & \text{if } j = \text{argmax}_{k \in I(i)} (w \cdot x_k), \\ 0, & \text{otherwise.} \end{cases}$ then the optimization problem is:

$$\begin{aligned} \min_{w,b,\xi,v} \quad & p(v^{(k)})^{p-1} v + C_1 \sum_{i=1}^{q} \xi_i + C_2 \sum_{i=r+1}^{r+s} \xi_i, \\ \text{s.t.} \quad & (w \cdot x_i^{(k)}) + b \geq 1 - \xi_i, i = 1, \cdots, q, \\ & (w \cdot x_i) + b \leq -1 + \xi_i, i = r+1, \cdots, r+s, \\ & \xi_i \geq 0, i = 1, \cdots, q, r+1, \cdots, r+s, \\ & -v \leq w \leq v. \end{aligned} \tag{12}$$

which is a standard linear programming, then the following algorithm is established:

Algorithm 2.

(1) Given a training set (2); Select the parameters $C_1 (C_1 > 0)$, $C_2 (C_2 > 0)$ and $p (0 < p < 1)$;

(2) Let $k = 0$, select $x_i^{(k)} = \frac{1}{|I(i)|} \sum_{j \in I(i)} x_j, i = 1, \cdots, q$ and $v^{(k)} = 0$;

(3) Solve the following optimization problem (12), and get its solution $(w^{(k+1)}, b^{(k+1)}, \xi^{(k+1)}, v^{(k+1)})$;

(4) Compute $g(x_j) = (w^{(k+1)} \cdot x_j) + b^{(k+1)}$, for $j \in I(i)$ and $i = 1, \cdots, q$, select $x_i^{(k+1)} = \text{argmax}_{j \in I(i)} g(x_j)$;

(5) If $\|x_i^{(k)} - x_i^{(k+1)}\| = 0$, for $i = 1, \cdots, q$, then let $w^* = w^{(k+1)}, b^* = b^{(k+1)}, \xi^* = \xi^{(k+1)}$ and stop; otherwise, let $k = k+1$, go to step (3).

4 The lower bounds of nonzero entries of local optimal solution to the problem (4)

Using the same strategy in [16], we get the following theorem 1, which can be used to identify nonzero components in the local optimal solutions to the problem (4), even though the Algorithm 2 can only find the approximate local optimal solution.

Theorem 1 For any local optimal solution (w^*, b^*, ξ^*) to the problem (4), if $|[w^*]_i| \leq (\frac{p}{\sqrt{C_1^2 q + C_2^2 s}\|g_i\|})^{\frac{1}{1-p}}$, then $[w^*]_i = 0$, $i = 1, 2, \cdots, n$, where g_i is defined in (3).

Proof: Suppose $\|w^*\|_0 = k$. Without loss of generality, let $w^* = ([w^*]_1, [w^*]_2, \cdots, [w^*]_k, 0, 0 \cdots 0)^T$ and $z^* = ([w^*]_1, [w^*]_2, \cdots, [w^*]_k)^T$. For the new instance $\widetilde{x_i} = ([x_i]_1, [x_i]_2, \cdots, [x_i]_k)^\top$, we consider the following optimization problem

$$\begin{array}{ll} \min\limits_{z,b,\xi} & \|z\|_p^p + C_1 \sum_{i=1}^q \xi_i + C_2 \sum_{i=r+1}^{r+s} \xi_i \ , \\ \text{s.t.} & \max\limits_{j \in I(i)} ((z \cdot \widetilde{x}_j) + b) \geq 1 - \xi_i, i = 1, \cdots, q, \\ & (z \cdot \widetilde{x}_i) + b \leq -1 + \xi_i, i = r+1, \cdots, r+s, \\ & \xi_i \geq 0, i = 1, \cdots, q, r+1, \cdots, r+s. \end{array} \tag{13}$$

It has been pointed out by [10] that the constraint $\max_{j \in I(i)}((z \cdot \widetilde{x}_j) + b) \geq 1 - \xi_i$, is equivalent to the fact that there exist convex combination coefficients $v_j^i \geq 0, \sum_{j \in I(i)} v_j^i = 1$, such that $(w \cdot \sum_{j \in I(i)} v_j^i \widetilde{x}_j) + b \geq 1 - \xi_i$. Then, the above problem (13) is equivalent to:

$$\begin{array}{ll} \min\limits_{z,b,\xi} & \|z\|_p^p + C_1 \sum_{i=1}^q \xi_i + C_2 \sum_{i=r+1}^{r+s} \xi_i \ , \\ \text{s.t.} & (z \cdot \sum_{j \in I(i)} v_j^i \widetilde{x}_j) + b \geq 1 - \xi_i, i = 1, \cdots, q, \\ & (z \cdot \widetilde{x}_i) + b \leq -1 + \xi_i, i = r+1, \cdots, r+s, \\ & \xi_i \geq 0, i = 1, \cdots, q, r+1, \cdots, r+s, \\ & v_j^i \geq 0, j \in I(i), i = 1, \cdots, q, \\ & \sum_{j \in I(i)} v_j^i = 1, i = 1, \cdots, q. \end{array} \tag{14}$$

The Lagrange function of (14) is:

$$L(z, b, \xi, \alpha, \zeta, \lambda, \mu) = \|z\|_p^p + C_1 \sum_{i=1}^{q} \xi_i + C_2 \sum_{i=r+1}^{r+s} \xi_i - \sum_{i=1}^{q} \alpha_i ((z \cdot \sum_{j \in I(i)} v_j^i \widetilde{x}_j) + b - 1 + \xi_i) +$$
$$\sum_{i=r+1}^{r+s} \alpha_i ((z \cdot \widetilde{x}_i) + b + 1 - \xi_i) - \sum_{i=1}^{q} \zeta_i \xi_i - \sum_{i=r+1}^{r+s} \zeta_i \xi_i - \sum_{i=1}^{q} \sum_{j \in I(i)} \lambda_j^i v_j^i - \mu (\sum_{j \in I(i)} v_j^i - 1).$$

It is easy to know that (z^*, b^*, ξ^*, v^*) is a local optimal solution of (14), according to the KKT condition, we have

$$p|z^*|^{p-1} sign(z^*) = \sum_{i=1}^{q} \alpha_i \sum_{j \in I(i)} v_j^i \widetilde{x}_i - \sum_{i=r+1}^{r+s} \alpha_i \widetilde{x}_i, \tag{15}$$

$$0 \leq \alpha_i \leq C_1, i = 1, \cdots, q, \tag{16}$$

$$0 \leq \alpha_i \leq C_2, i = r+1, \cdots, r+s. \tag{17}$$

According to (15)-(17) and Cauchy-Schwarz inequality, we have

$$\begin{aligned} p|[z^*]_i|^{p-1} &= [|\sum_{i=1}^{q} \alpha_i \sum_{j \in I(i)} v_j^i \widetilde{x}_i - \sum_{i=r+1}^{r+s} \alpha_i \widetilde{x}_i|]_i & (18) \\ &\leq \|(\alpha_1, \cdots, \alpha_q, -\alpha_{r+1}, \cdots, -\alpha_{r+s})\| \\ &\quad \|([\sum_{i \in I(1)} v_j^1 \widetilde{x}_j]_i, \cdots, [\sum_{i \in I(p)} v_j^q \widetilde{x}_j]_i, [\tilde{x}_{r+1}]_i, \cdots, [\tilde{x}_{r+s}]_i)\| & (19) \\ &\leq \sqrt{C_1^2 q + C_2^2 s} \|g_i\| & (20) \end{aligned}$$

which means $|[z^*]_i| \geq (\frac{p}{\sqrt{C_1^2 q + C_2^2 s}\|g_i\|})^{\frac{1}{1-p}}$, then the conclusion is obtained. □

According to Theorem 1, we can identify the nonzero components of the local optimal solution to (4). Based on the Algorithm 2 and the Theorem 1, the new algorithm 3 is established as follows:

Algorithm 3.

(1) Given a training set (2); Select the parameters $C_1(C_1 > 0), C_2(C_2 > 0)$ and p $(0 < p < 1)$;

(3) Using the Algorithm 2 to get the local optimal solution (w^*, b^*, ξ^*) to the problem (4);

(4) Compute $L_i = (\frac{p}{\sqrt{C_1^2 q + C_2^2 s}\|g_i\|})^{\frac{1}{1-p}}$, for $i = 1, \cdots, n$; Select the feature set: $\{i | [w^*]_i > L_i, (i = 1, \cdots, n)\}$;

(5) Construct the decision function $f(x) = \text{sgn}((\tilde{w}^* \cdot \tilde{x}) + b^*)$, where the components of $\tilde{w}^*$ are nonzero components of w^* and the components of $\tilde{x}$ are also corresponding to nonzero components of w^*.

Note that, in the following section, our experiments are conducted according to the algorithm 3.

5 Numerical experiments

In this section, some experiments on four simulated datasets are conducted, by comparing p-MISVM with MICA. The four simulation datasets (I, II, III, IV) are generated by the following steps:

• According to two different distributions, independently generate n_0 positive and negative feature vectors $g_i^+ \in R^{l_+}$, $g_i^- \in R^{l_-}, i = 1, 2, \cdots, n_0$ where l_+ and l_- are respectively the number of the positive and negative points;

Table 1: Four simulate datasets

Data	features	relevant features	Distribution of g^+	Distribution of g^-
I	20	3	$N(1,0.5)$	$N(-1,0.5)$
II	20	3	$U(-0.5,1)$	$U(-1,0)$
III	100	5	$N(1,0.5)$	$N(-1,0.5)$
IV	100	5	$U(-0.5,1)$	$U(-1,0)$

Table 2: Results on the four Simulated datasets

Dataset	Methods	No. of selected features	Percent of relevant features(%)	Average accuracy(%)	Parameters
I	p-MISVM	**2.96**	**88.1**	**99.88**	p=0.5, $C=2.8$
	MICA	5.05	59.4	99.86	C=2
II	p-MISVM	**2.91**	**92.7**	**99.98**	$p=0.5, C=1$
	MICA	3.36	89	99.88	C=0.7
III	p-MISVM	**5.06**	**89**	**99.98**	$p=0.6, C=0.57$
	MICA	6.97	71	99.98	C=2
IV	p-MISVM	**3.51**	**83.7**	99.36	$p=0.6, C=0.7$
	MICA	3.83	75.9	**99.46**	C=0.7

• According to other distributions, independently generate some stochastic vectors that are irrelevant to the class;

• The positive bags contains three points that are stochastic generated in the rectangular region, the radius of this rectangular region is 2 and the center is the positive points generated by the first step.

• The negative bags is just the negative points generated by the first step.

The description of the four data sets is listed in Table 1.

According to Algorithm 3, 100 experiments are conducted for every dataset. Note that, there are three parameters C_1, C_2 and p in Algorithm 3. Usually we set $C_1 = C_2 = C$ in our experiments, and the best value of these parameters is chosen by ten-fold cross validation. Our experimental results are illustrated in Table 2, where the best results are given by the bold form. Obviously, p-MISVM performs the best among two methods. In Table 2, the data in 4th column shows the percentage of the number of the right features over the number of the selected features, which means the bigger the value the better the result. The average accuracy is computed by averaging the test accuracy among 100 experiments. It is easy to see that p-MISVM selects the least features with the high accuracy, compared with the MICA.

6 Conclusion

Feature selection is very important in many applications of data mining. This paper introduces a new version of SVM named p-MISVM to feature selection and multi-instance

classification. By using the CCCP method, a linearization algorithm is proposed to get the approximate local optimal solution to p-MISVM. And the lower bounds for the absolute value of nonzero components in every local optimal solution is established, which can eliminate zero components in any numerical solution. The numerical experiments show that the p-norm support vector machines is effective in selecting relevant features, compared with the popular MICA.

References

[1] Dietterich, T. G., Lathrop, R. H., Lozano Perez, T. Solving the multiple-instance problem with axis-parallel rectangles. Artificial Intelligence, 89, 31-71, 1997.

[2] Andrews, S., Tsochantaridis, I., Hofmann, T. Support vector machines for multiple instance learning. In Advances in neural information processing systems 15, 2002.

[3] Viola, P., Platt, J., Zhang, C. Multiple instance boosting for object detection. In Advances in neural information processing systems 18, 1417-1424, 2006.

[4] Maron, O., Lozano-Perez, T. A framework for multiple-instance learning. In Advances in neural information processing systems 10, 570-576, 1998.

[5] Xin, X., Frank, E. Logistic regression and boosting for labeled bags of instances. Proc 8th Pacific-Asia Conference on Knowledge Discovery and Data Mining, 272-281, 2004.

[6] Fung, G., Dundar, M., Krishnapuram, B., Rao, R. B. Multiple instance learning for computer aided diagnosis. In Advances in neural infor- mation processing systems 19, 425-432, 2007.

[7] Settles, B., Craven, M., Ray, S. Multiple instance active learning. In Advances in neural information processing systems 20, 2008.

[8] Wang, J., Zucker, J. Solving the multiple- instance problem: A lazy learning approach. In Proceedings of the 17th international conference on machine learning, 1119-1125, 2000.

[9] Andrews S., Tsochantaridis I., Hofmann T. Support vector machines for multiple-instance learning. Advances in Neural Information Processing Systems. 15 Cambridge, MA: MIT Press, 2003.

[10] Mangasarian O.L., Wild E.W. Multiple instance classification via successive linear programming. Journal of Optimization Theory and Application. 137(1): , 555-568,2008.

[11] Bradley P.S., Mangasarian O.L. Feature selection via concave minimization and support vector machines. In Proc. 13th ICML, 82-90, 1998.

[12] Zhu J., Rosset S., Hastie T. Tibshirani R. 1-norm svms, Advances in Neural Information Processing Systems 16, 2003.

[13] Tan, J.Y., Zhang, C.H., Deng, N.Y. Cancer gene identification via p-norm support vector machine, ISB2010, to be accepted.

[14] Yuille A., Rangarajan A. The concave-convex procedure. Neural Computation. 15:, 915-936, 2003.

[15] Smola A.J, Vishwanathan S.V.N., Hofmann T. Kernel methods for missing variables. Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics. Barbados. 2005.

[16] Chen X.J.,Xu F.M., and Ye Y.Y. Lower bound theory of nonzero entries in solutions of l_2-l_p minimization. Technical report, Department of Applied Mathematics, the Hong Kong Polytechnic University, 2009.