The Ninth International Symposium on Operations Research and Its Applications (ISORA'10)
Chengdu-Jiuzhaigou, China, August 19–23, 2010
Copyright © 2010 ORSC & APORC, pp. 433–442

Multiple Instance Twin Support Vector Machines*

Yuan-Hai Shao[1] Zhi-Xia Yang[2] Xiao-Bo Wang[3]
Nai-Yang Deng[1,†]

[1]College of science, China Agricultural University, Beijing 100083, China
[2]College of Mathematics and System Science, Xinjiang University, Urumuchi 830046, China
[3]Department of Automation, Tsinghua University, Beijing 100084, China

Abstract Considering the multiple instance learning(MIL) in classification problem, a novel multiple instance twin support vector machines(MI-TWSVM) method is proposed. For linear classification, unlike other maximum margin SVM-based MIL methods, the proposed approach leads to two non-parallel hyperplanes. The non-linear classification via kernels is also studied. Preliminary experimental results on public datasets indicate that our MIL method is competitive with the previous MIL methods.

Keywords Multiple instance learning; Classification; Support vector machine; Twin support vector machine

1 Introduction

Multiple instance learning (MIL) was first introduced by Dietterich et al.[2] in the context of drug activity prediction. The task is to predict whether a drug molecule can bind to the targets (enzymes or cell-surface receptors). Each drug molecule (called a bag) can have multiple low-energy shapes or conformations (instances). The molecule is considered to be useful as a drug if one of its conformations can bind to the targets. However, biochemical data can only tell the binding capability of a molecule, but not a particular conformation. Thus, while each training pattern has a known label in the standard supervised learning, only the bags (but not the individual instances) have known labels in MIL. In other words, MIL only provides weak label information of the training data.

Following the seminal work of Dietterich et al.[2], a number of MIL methods emerged. Examples include the Diverse Density (DD)[5], EM-DD [12], MI-NN [8], mi-SVM [1], MI-SVM [1], MICA [4] and SVM-CC [11]. Besides classification, progress has also been made on MIL with real-valued outputs [8]. Moreover, many applications are now considered as MIL problems. Examples include content based image retrieval [8, 5], where each

[*] 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, No.10801112) and the China Agricultural University Innovation Fund for Graduate Student.

[†] Corresponding author. E-mail: dengnaiyang@cau.edu.cn

image is a bag and each local image patch an instance, and text categorization prediction [10, 1].

In this paper, we focus on SVM-based MIL methods. In this field, e.g. in [4, 1, 11], the usual approach is use the maximum “margin" principle and the “witness" instances in positive bags are obtained by selecting the farthest from the boundary constructed by SVMs. Inspired by twin support vector machine(TWSVM) [3], we propose a new MIL method called multiple instance twin support vector machines(MI-TWSVM), which is an extension of the TWSVM. For linear classification, our MI-TWSVM aims at generating a positive hyperplane and a negative hyperplane, such that the former one is close to at least one instance(“witness" instance) in every positive bag and is far from all instances belonging to negative bags, and the latter one is close to all instances belonging to negative bags and is far from the “witness" instances in positive bags. In other word, the “witness" instance in a positive bag by selecting the farthest from the boundary constructed by TWSVM. Moreover, instead of having QP problems, the MI-TWSVM optimization problem are bilevel programming problems (BLPPs). We using a simple optimization heuristic for solving these problems.

2 Primal Model

Multiple instance classification generalizes the standard classification by making significantly weaker assumptions on the labeling information. In multiple instance classification, instances are grouped into bags and a label is attached to each bag instead of to each instance. More formally, we consider the problem of classify l positive bags and m negative bags in n-dimensional real space R^n. Suppose that the positive bags are represented by $B_1,\cdots,B_l$ with the label of $+1$ and the negative bags by $B_{l+1},\cdots,B_{l+m}$ with the label of -1, where $B_i=\{x_{i1},x_{i2},\cdots,x_{in_i}\}$ is the ith bag containing instances $x_{ij}\in R^n, j=1,...,n_i, i=1,...,l,l+1,...,l+m.$

2.1 Linear Model

Different from the idea of maximizing the “margin" between two disjoint half planes of the positive bags and negative bags, we aim at generating two non-parallel hyperplanes(a positive hyperplane and a negative hyperplane)

$$f_1(x)=w_1^\top x+b_1=0 \quad and \quad f_2(x)=w_2^\top x+b_2=0, \tag{1}$$

such that the positive hyperplane is close to the “positive" instances in positive bags and is far from the all instances in negative bags, and the negative hyperplane is close to the all negative instances and is far from the “positive" instances in positive bags.

Only one instance called “witness" in every positive bag matters. The positive witness is usually the “most positive" one in the positive bag. In order to introduce the degree of an instance x belonging to the positive class, consider the distance between the instance x and the boundary consisting of the separating hyperplanes

$$\frac{w_1^\top x+b_1}{\|w_1\|}+\frac{w_2^\top x+b_2}{\|w_2\|}=0, \quad \frac{w_1^\top x+b_1}{\|w_1\|}-\frac{w_2^\top x+b_2}{\|w_2\|}=0, \tag{2}$$

generated by positive and negative hyperplanes (1). Obviously the distance between an instance x and the boundary is

$$D(x) = \min\ \{D_+(x), D_-(x)\}, \tag{3}$$

where $D_+(x)$ and $D_-(x)$ are respectively the distances between the instance x and the hyperplanes (2)

$$\begin{aligned} D_+(x) &= \frac{\left|\frac{w_1^\top x+b_1}{\|w_1\|} + \frac{w_2^\top x+b_2}{\|w_2\|}\right|}{\left|\left|\frac{w_1}{\|w_1\|} + \frac{w_2}{\|w_2\|}\right|\right|} = \frac{\left|\frac{w_1^\top x+b_1}{\|w_1\|} + \frac{w_2^\top x+b_2}{\|w_2\|}\right|}{\sqrt{2+\frac{2(w_1\cdot w_2)}{||w_1||||w_2||}}}, \\ D_-(x) &= \frac{\left|\frac{w_1^\top x+b_1}{\|w_1\|} - \frac{w_2^\top x+b_2}{\|w_2\|}\right|}{\left|\left|\frac{w_1}{\|w_1\|} - \frac{w_2}{\|w_2\|}\right|\right|} = \frac{\left|\frac{w_1^\top x+b_1}{\|w_1\|} - \frac{w_2^\top x+b_2}{\|w_2\|}\right|}{\sqrt{2-\frac{2(w_1\cdot w_2)}{||w_1||||w_2||}}}. \end{aligned} \tag{4}$$

Noticing that $\frac{|w_1^\top x+b_1|}{|w_2^\top x+b_2|} \cdot \frac{||w_2||}{||w_1||} < 1$ and $\frac{|w_1^\top x+b_1|}{|w_2^\top x+b_2|} \cdot \frac{||w_2||}{||w_1||} > 1$ imply that x is in the positive and negative region respectively. it is reasonable to define the degree of the instance x belonging to the positive class as

$$d(x) = \begin{cases} D(x) & \frac{|w_1^\top x+b_1|}{|w_2^\top x+b_2|} \cdot \frac{||w_2||}{||w_1||} < 1; \\ 0 & \frac{|w_1^\top x+b_1|}{|w_2^\top x+b_2|} \cdot \frac{||w_2||}{||w_1||} = 1; \\ -D(x) & \frac{|w_1^\top x+b_1|}{|w_2^\top x+b_2|} \cdot \frac{||w_2||}{||w_1||} > 1. \end{cases} \tag{5}$$

The bigger the value $d(x)$ is, the more positive the x is. For positive bag $B_i, i = 1,...l$, we introduce a selector which finds the positive “witness" in B_i by selecting the most positive instance $x_i = \arg\max_{x\in B_i} f(x)$. Once these “witness" instances are identified, the relative position of other instances in positive bags would become irrelevant. For negative bags, they are unfolded into instances. Thus, the positive and negative hyperplanes are obtained respectively by solving the following bilevel programming problem (BLPP):

$$\begin{aligned} \min_{w_1,w_2,b_1,b_2,\eta,\xi,x_i} \quad & \sum_{i=1}^{l}((w_1^\top x_i+b_1)^2+c_2\eta_i) + \sum_{i=l+1}^{l+m}\sum_{k=1}^{n_i}((w_2^\top x_{ik}+b_2)^2+c_1\xi_{ik}) && (6) \\ \text{s.t.} \quad & w_1^\top x_{ik}+b_1 \geq 1-\xi_{ik}, \xi_{ik} \geq 0, k = 1,...,n_i, i = l+1,...m, && (7) \\ & w_2^\top x_i + b_2 \geq 1-\eta_i, \eta_i \geq 0, i = 1,...l, && (8) \\ & x_i = \arg\max_{x_{ij}\in B_i}\ \{d(x_{ij})\}, i = 1,...l, && (9) \end{aligned}$$

where $c_1 > 0$ and $c_2 > 0$ are positive parameters, $d(x)$ is defined by (5).

Once the w_1, w_2, b_1, b_2 and the “witness" instances x_i $(i = 1,...,l)$ are found from the (6)–(9), the bag label of a particular bag B is deduced by

$$F(B) = \text{sgn}\ (1-\max_{x\in B}(\frac{|w_1^\top x+b_1|}{|w_2^\top x+b_2|}\cdot\frac{||w_2||}{||w_1||})). \tag{10}$$

2.2 Nonlinear Model

The above approach can also be extended to construct nonlinear multi-surface classifiers by considering the following kernel-generated surfaces instead of the hyperplanes

$$f_1(x) = u_1^\top K(x^\top, C^\top) + b_1 = 0 \quad \textit{and} \quad f_2(x) = u_2^\top K(x^\top, C^\top) + b_2 = 0, \tag{11}$$

where $C^\top$ is a matrix contains the "witness" instances in positive bags and all instances in negative bags, and $K(\cdot,\cdot)$ is an appropriately chosen kernel. Note that the hyperplanes can be obtained as a special case of (11), by using a linear kernel $K(x^\top, C^\top) = C^\top x$, and defining $w_1 = Cu_1$ and $w_2 = Cu_2$.

In order to introduce the degree of an instance x belonging to the positive class, consider the separating surfaces

$$\begin{aligned} \frac{u_1^\top K(x,C^\top)+b_1}{\sqrt{u_1^\top K(C,C^\top)u_1^\top}} + \frac{u_2^\top K(x,C^\top)+b_2}{\sqrt{u_2^\top K(C,C^\top)u_2^\top}} = 0, \\ \frac{u_1^\top K(x,C^\top)+b_1}{\sqrt{u_1^\top K(C,C^\top)u_1^\top}} - \frac{u_2^\top K(x,C^\top)+b_2}{\sqrt{u_2^\top K(C,C^\top)u_2^\top}} = 0, \end{aligned} \tag{12}$$

generated by positive and negative surfaces (11). Define the distance between an instance x and the boundary is

$$D(x) = \min\ \{D_+(x), D_-(x)\}, \tag{13}$$

where

$$\begin{aligned} D_+(x) &= \frac{\left|\frac{u_1^\top K(x,C^\top)+b_1}{\sqrt{u_1^\top K(C,C^\top)u_1^\top}} + \frac{u_2^\top K(x,C^\top)+b_2}{\sqrt{u_2^\top K(C,C^\top)u_2^\top}}\right|}{\sqrt{2 + \frac{2u_1^\top K(C,C^\top)u_2^\top}{u_1^\top K(C,C^\top)u_1^\top u_2^\top K(C,C^\top)u_2^\top}}}, \\ D_-(x) &= \frac{\left|\frac{u_1^\top K(x,C^\top)+b_1}{\sqrt{u_1^\top K(C,C^\top)u_1^\top}} - \frac{u_2^\top K(x,C^\top)+b_2}{\sqrt{u_2^\top K(C,C^\top)u_2^\top}}\right|}{\sqrt{2 - \frac{2u_1^\top K(C,C^\top)u_2^\top}{u_1^\top K(C,C^\top)u_1^\top u_2^\top K(C,C^\top)u_2^\top}}}. \end{aligned} \tag{14}$$

We also define the degree of an instance x belonging to the positive class by:

$$d(x) = \begin{cases} D(x) & \frac{|u_1^\top K(x,C^\top)+b_1|}{|u_2^\top K(x,C^\top)+b_2|} \frac{\sqrt{u_2^\top K(C,C^\top)u_2^\top}}{\sqrt{u_1^\top K(C,C^\top)u_1^\top}} < 1; \\ 0 & \frac{|u_1^\top K(x,C^\top)+b_1|}{|u_2^\top K(x,C^\top)+b_2|} \frac{\sqrt{u_2^\top K(C,C^\top)u_2^\top}}{\sqrt{u_1^\top K(C,C^\top)u_1^\top}} = 1; \\ -D(x) & \frac{|u_1^\top K(x,C^\top)+b_1|}{|u_2^\top K(x,C^\top)+b_2|} \frac{\sqrt{u_2^\top K(C,C^\top)u_2^\top}}{\sqrt{u_1^\top K(C,C^\top)u_1^\top}} > 1. \end{cases} \tag{15}$$

In line with the arguments in above section, our minimization problem for generating kernel-based nonlinear surfaces becomes the bilevel programming problem (BLPP):

$$\begin{aligned}
\min_{u_1,u_2,b_1,b_2,\eta,\xi,x_i} \quad & \sum_{i=1}^{l}((u_1^\top K(x_i,C^\top)+b_1)^2+c_2\eta_i)+ \\
& \sum_{i=l+1}^{l+m}\sum_{k=1}^{n_i}((u_2^\top K(x_{ik},C^\top)+b_2)^2+c_1\xi_{ik}) \qquad (16) \\
\text{s.t.} \quad & u_1^\top K(x_{ik},C^\top)+b_1 \geq 1-\xi_{ik},\xi_{ik}\geq 0, \\
& k=1,...,n_i, i=l+1,...m, \qquad (17) \\
& u_2^\top K(x_i,C^\top)+b_2 \geq 1-\eta_i,\eta_i\geq 0, i=1,...l, \qquad (18) \\
& x_i = \arg\max_{x_{ij}\in B_i}\{d(x_{ij})\}, i=1,...l, \qquad (19)
\end{aligned}$$

where $c_1 > 0$ and $c_2 > 0$ are positive parameters, $C^\top$ is composed of the “witness" instances x_i $(i=1,...,l)$ in positive bags and all instances x_{ik} $(k=1,...,n_i,\ i=l+1,...,m)$ in negative bags, $d(x)$ is defined by (15).

Once the u_1, u_2, b_1, b_2 and the “witness" instances x_i $(i=1,...,l)$ are known from the (16)–(19), the bag label information of a particular bag B is deduced by

$$F(B) = \text{sgn } (1-\max_{x\in B}(\frac{|u_1^\top K(x,C^\top)+b_1|}{|u_2^\top K(x,C^\top)+b_2|}\cdot\frac{\sqrt{u_2^\top K(C,C^\top)u_2^\top}}{\sqrt{u_1^\top K(C,C^\top)u_1^\top}})). \qquad (20)$$

3 Optimization Heuristics

The formulations (6)–(9) and (16)–(19) can be casted as mixed-integer programming. In deriving optimization heuristics, we exploit the fact that for given integer variables, i.e. the selected $x_i \in B_i$ $(i=1,...,l)$, the problem is reduced to a convex QP problems that can be solved easily. Thus for solving the linear model problem, we arrive the following algorithm.

Linear MI – TWSVM :

(1) Given the positive bags represented by $B_1,\cdots,B_l$ with the label of $+1$ and the negative bags by $B_{l+1},\cdots,B_{l+m}$ with the label of -1, where $B_i=\{x_{i1},x_{i2},\cdots,x_{in_i}\}$ is the *ith* bag containing instances $x_{ij}\in R^n, j=1,...,n_i, i=1,...,l,...,l+m$;

(2) Select proper kernel function $K(\cdot,\cdot)$ and parameters $c_1>0, c_2>0$, randomly select initial instance x_i^1 in each positive bag B_i, i.e., $x_i^1 = x_{i1}$, $i=1,...,l$; set $t=1$;

(3) Compute w_1^t, b_1^t, w_2^t and b_1^t from instances $\{x_i^t\}_{i=1}^l$ and all instances in negative bags by solve the convex quadratic programming

$$\begin{aligned}
\min_{w_1^t,b_1^t,\xi} \quad & \sum_{i=1}^{l}(w_1^{t\top}x_i^t+b_1^t)^2+c_1\sum_{i=l+1}^{l+m}\sum_{k=1}^{n_i}\xi_{ik}, \qquad (21) \\
\text{s.t.} \quad & w_1^{t\top}x_{ik}+b_1^t \geq 1-\xi_{ik}, \qquad (22) \\
& \xi_{ik}\geq 0, k=1,...,n_i, i=l+1,...,l+m, \qquad (23)
\end{aligned}$$

and

$$\min_{w_2^t, b_2^t, \eta} \quad \sum_{i=l+1}^{l+m}\sum_{k=1}^{n_i}(w_2^{t\top}x_{ik}+b_2^t)^2 + c_2\sum_{i=1}^{l}\eta_i, \tag{24}$$

$$\text{s.t.} \quad (w_2^{t\top}x_i^t + b_2^t) \geq 1-\eta_i, \eta_i \geq 0, i=1,...,l. \tag{25}$$

(4) Compute x_i^{t+1}, $i=1,...,l$ where x_i^{t+1} is the optimal solution of the following problem:

$$\max_{x_{ij}\in B_i} d^t(x_{ij}), \tag{26}$$

where

$$d^t(x) = \begin{cases} \min\ \{D_+^t(x), D_-^t(x)\} & \frac{|w_1^{t\top}x+b_1^t|}{|w_2^{t\top}x+b_2^t|}\cdot\frac{||w_2^t||}{||w_1^t||} < 1; \\ 0 & \frac{|w_1^{t\top}x+b_1^t|}{|w_2^{t\top}x+b_2^t|}\cdot\frac{||w_2^t||}{||w_1^t||} = 1; \\ -\min\ \{D_+^t(x), D_-^t(x)\} & \frac{|w_1^{t\top}x+b_1^t|}{|w_2^{t\top}x+b_2^t|}\cdot\frac{||w_2^t||}{||w_1^t||} > 1; \end{cases} \tag{27}$$

where

$$D_+^t(x) = \frac{|\frac{w_1^{t\top}x+b_1^t}{||w_1^t||} + \frac{w_2^{t\top}x+b_2^t}{||w_2^t||}|}{\sqrt{2+\frac{2(w_1^t\cdot w_2^t)}{||w_1^t||||w_2^t||}}}, \quad D_-^t(x) = \frac{|\frac{w_1^{t\top}x+b_1^t}{||w_1^t||} - \frac{w_2^{t\top}x+b_2^t}{||w_2^t||}|}{\sqrt{2-\frac{2(w_1^t\cdot w_2^t)}{||w_1^t||||w_2^t||}}}. \tag{28}$$

Denote the corresponding optimal value as $g^t(x_{ij})$.

(5) Compare $\sum_{i=1}^l g^t(x_{ij})$ with $\sum_{i=1}^l g^{t+1}(x_{ij})$. When their difference is small enough, set $x_i^{t+1}=x_i^*$, $w_1^{t+1}=w_1^*$, $b_1^{t+1}=b_1^*$, $w_2^{t+1}=w_2^*$ and $b_2^{t+1}=b_2^*$, stop; Otherwise, set $t=t+1$, go to step (3);

(6) Construct the decision function

$$F(B) = \text{sgn}\ (1-\max_{x\in B}(\frac{|w_1^{*\top}x+b_1^*|}{|w_2^{*\top}x+b_2^*|}\cdot\frac{||w_2^*||}{||w_1^*||})). \tag{29}$$

In line with the Linear MI-TWSVM, for solving the nonlinear model problem, we arrive the following algorithm.

Nonlinear MI − TWSVM :

(1) Given the positive bags represented by $B_1,\cdots,B_l$ with the label of $+1$ and the negative bags by $B_{l+1},\cdots,B_{l+m}$ with the label of -1, where $B_i=\{x_{i1},x_{i2},\cdots,x_{in_i}\}$ is the ith bag containing instances $x_{ij}\in R^n, j=1,...,n_i, i=1,...,l,...,l+m$;

(2) Select proper kernel function $K(\cdot,\cdot)$ and parameter $c_1>0, c_2>0$, randomly select initial instance x_i^1 in each positive bag B_i, i.e., $x_i^1=x_{i1}$, set $t=1$;

(3) Compute u_1^t, b_1^t, u_2^t and b_2^t from instances $\{x_i^t\}_{i=1}^l$ and all instances in negative bags by solve the convex quadratic programming

$$\min_{u_1^t, b_1^t, \xi} \quad \sum_{i=1}^{l}(u_1^{t\top}K(x_i^t, C^{t\top})+b_1^t)^2 + c_1\sum_{i=l+1}^{l+m}\sum_{k=1}^{n_i}\xi_{ik}, \tag{30}$$

$$\text{s.t.} \quad u_1^{t\top}K(x_{ik}, C^{t\top}) + b_1^t \geq 1-\xi_{ik}, \tag{31}$$

$$\xi_{ik}\geq 0, k=1,...,n_i, i=l+1,...,l+m, \tag{32}$$

and

$$\min_{u_2^t, b_2^t, \eta} \quad \sum_{i=l+1}^{l+m} \sum_{k=1}^{n_i} (u_2^{t\top} K(x_{ik}, C^{t\top}) + b_2^t)^2 + c_2 \sum_{i=1}^{l} \eta_i, \tag{33}$$

$$\text{s.t.} \quad (u_2^{t\top} K(x_i^t, C^{t\top}) + b_2^t) \geq 1 - \eta_i, \eta_i \geq 0, i = 1, ..., l, \tag{34}$$

where $C^{t\top}$ is composed of the current "witness" instances x_i^t $(i = 1, ..., l)$ in positive bags and all instances in negative bags.

(4) Compute x_i^{t+1} $(i = 1, ..., l)$, where x_i^{t+1} is the optimal solution of the following problem:

$$\max_{x_{ij} \in B_i} d^t(x_{ij}), \tag{35}$$

where

$$d^t(x) = \begin{cases} \min \ \{D_+^t(x), D_-^t(x)\} & \frac{|u_1^{t\top} K(x, C^{t\top}) + b_1^t|}{|u_2^{t\top} K(x, C^{t\top}) + b_2^t|} \frac{\sqrt{u_2^{t\top} K(C^t, C^{t\top}) u_2^{t\top}}}{\sqrt{u_1^{t\top} K(C^t, C^{t\top}) u_1^{t\top}}} < 1; \\ 0 & \frac{|u_1^{t\top} K(x, C^{t\top}) + b_1^t|}{|u_2^{t\top} K(x, C^{t\top}) + b_2^t|} \frac{\sqrt{u_2^{t\top} K(C^t, C^{t\top}) u_2^{t\top}}}{\sqrt{u_1^{t\top} K(C^t, C^{t\top}) u_1^{t\top}}} = 1; \\ -\min \ \{D_+^t(x), D_-^t(x)\} & \frac{|u_1^{t\top} K(x, C^{t\top}) + b_1^t|}{|u_2^{t\top} K(x, C^{t\top}) + b_2^t|} \frac{\sqrt{u_2^{t\top} K(C^t, C^{t\top}) u_2^{t\top}}}{\sqrt{u_1^{t\top} K(C^t, C^{t\top}) u_1^{t\top}}} > 1; \end{cases} \tag{36}$$

where

$$D_+^t(x) = \frac{\left| \frac{u_1^{t\top} K(x, C^{t\top}) + b_1^t}{\sqrt{u_1^{t\top} K(C^t, C^{t\top}) u_1^{t\top}}} + \frac{u_2^{t\top} K(x, C^{t\top}) + b_2^t}{\sqrt{u_2^{t\top} K(C^t, C^{t\top}) u_2^{t\top}}} \right|}{\sqrt{2 + \frac{2u_1^{t\top} K(C^t, C^{t\top}) u_2^{t\top}}{u_1^{t\top} K(C^t, C^{t\top}) u_1^{t\top} u_2^{t\top} K(C^t, C^{t\top}) u_2^{t\top}}}},$$

$$D_-^t(x) = \frac{\left| \frac{u_1^{t\top} K(x, C^{t\top}) + b_1^t}{\sqrt{u_1^{t\top} K(C^t, C^{t\top}) u_1^{t\top}}} - \frac{u_2^{t\top} K(x, C^{t\top}) + b_2^t}{\sqrt{u_2^{t\top} K(C^t, C^{t\top}) u_2^{t\top}}} \right|}{\sqrt{2 - \frac{2u_1^{t\top} K(C^t, C^{t\top}) u_2^{t\top}}{u_1^{t\top} K(C^t, C^{t\top}) u_1^{t\top} u_2^{t\top} K(C^t, C^{t\top}) u_2^{t\top}}}}. \tag{37}$$

Denote the corresponding optimal value as $g^t(x_{ij})$.

(5) Compare $\sum_{i=1}^{l} g^t(x_{ij})$ with $\sum_{i=1}^{l} g^{t+1}(x_{ij})$. When their difference is small enough, set $x_i^{t+1} = x_i^*$, $u_1^{t+1} = u_1^*$, $b_1^{t+1} = b_1^*$, $u_2^{t+1} = u_2^*$ and $b_2^{t+1} = b_2^*$, stop; Otherwise, set $t = t+1$, go to step (3);

(6) Construct the decision function

$$F(B) = \text{sgn}\ (1 - \max_{x \in B}(\frac{|u_1^{*\top} K(x, C^{*\top}) + b_1^*|}{|u_2^{*\top} K(x, C^{*\top}) + b_2^*|} \cdot \frac{\sqrt{u_2^{*\top} K(C^*, C^{*\top}) u_2^{*\top}}}{\sqrt{u_1^{*\top} K(C^*, C^{*\top}) u_1^{*\top}}})). \tag{38}$$

4 Experimental Results

To demonstrate the capabilities of our formulation we conduct experiments on the same datasets with the ones in [1]. Two of these datasets are from the UCI machine learning repository[7], and ten from [1]. The two datasets from the UCI repository [7] are the Musk datasets, which are commonly used in multiple instance classification. We report results on these datasets for our nonlinear classification algorithm. We use the datasets from [1] to evaluate our linear classification algorithm. Three of these datasets are from an image annotation task in which the goal is to determine whether or not a given animal appears in an image. The other seven datasets are from the OHSUMED data and the task is to learn binary concepts associated with the Medical Subject Headings of MEDLINE documents.

The testing accuracy of Algorithm are calculated using the standard 10-fold cross validation method[6]. The regularization parameters c_1 and c_2 are selected from the set $\{2^i|i=-8,\cdots,8\}$ by 10-fold cross validation on the tuning set comprising of random 10% of the training data, and the RBF kernel parameter γ is selected from the set $\{2^i|i=-12,\cdots,4\}$. Algorithms terminate if the difference between the convex combination coefficients is smaller than 10^{-3} or the iterations $k>50$.

Table 1. Ten-fold testing accuracy on the Musk-1 and Musk-2 datasets. The best accuracy is shown in bold type figure.

Data sets	IAPR	SVM-CC	MICA	mi-SVM	MI-SVM	MI-TWSVM
MUSK1	92.4	88.9	84.4	87.4	77.9	**94.6**
MUSK2	89.2	-	**90.5**	83.6	84.3	88.2

We compare our MI-TWSVM with IAPR[2], EM-DD[12], MICA[4], mi-SVM[1], MI-SVM[1] and SVM-CC[11]. The 10-fold cross validation accuracy on “Musk1" and “Musk2" is listed in Table 1, where the results for mi-SVM and MI-SVM, for IAPR and MICA, and SVM-CC are taken from [1], [4] and [11] respectively. It can be seen from Table 1 that our method gives the best correctness on the “Musk1" and is competitive with the best one on “Musk2".

Table 2. Ten-fold testing accuracy on the Image datasets. The best accuracy is shown in bold type figure.

Data sets	EM-DD	SVM-CC	MICA	mi-SVM	MI-SVM	MI-TWSVM
Fox	56.1	-	58.7	58.2	57.8	**62.5**
Tiger	72.1	83.0	82.6	78.4	**84.0**	79.0
Elephant	78.3	81.5	80.5	82.2	81.4	**83.5**

The 10-fold cross validation accuracy on Image datasets is listed in Table 2, where the results for EM-DD, mi-SVM and MI-SVM, for MICA, and SVM-CC are taken from [1], [4] and [11] respectively. It can be seen from Table 2 that our method gives the best correctness on the “Fox" and “Elephant". This implies that our MI-TWSVM are comparable methods with previous MIL.

Table 3. Ten-fold testing accuracy on the document categorization datasets. The best accuracy is shown in bold type figure.

Data sets	EM-DD	MICA	mi-SVM	MI-SVM	MI-TWSVM
TST1	85.8	**94.5**	93.6	93.9	90.5
TST2	84.0	85.5	78.2	84.5	**86.3**
TST3	69.0	86.0	**87.0**	82.2	81.8
TST4	80.5	**87.7**	82.8	82.4	83.0
TST7	75.4	78.9	**81.3**	78.0	77.5
TST9	65.5	61.4	67.5	60.2	**70.5**
TST10	78.5	**82.3**	79.6	79.5	79.0

The 10-fold cross validation accuracy on on document categorization datasets is listed in Table 3, where the results for EM-DD, mi-SVM and MI-SVM,and for MICA are taken from [1], [4] respectively. The results in Table 3 are similar with that appeared in Table 2, and therefore confirm the above conclusion further.

5 Conclusion

We have introduced a bilevel programming method to multiple instance classification, termed MI-TWSVM. Different from maximizing the “margin" in other SVM-type methods in MIL, MI-TWSVM generates two non-parallel hyperplanes (positive hyperplane and negative hyperplane) such that the positive hyperplane is close to “witness" instances in positive bags and is distant from the all negative instances; at the same time, the negative hyperplane is close to negative instances and is distant from “witness" instances in positive bags. The “witness" instance in positive bag is the “most positive" instance by defining the degree of the distance belonging to positive class. Furthermore, our method can be easily extended to nonlinear classifiers by using the nonlinear kernels. Results on previously published datasets indicate that our approach is effective. Improvements on better optimization and evaluation using a wide variety of datasets and algorithms, such as those in [9], are promising avenues of future research.

References

[1] S. Andrews, I. Tsochantaridis, and T. Hofmann. Support vector machines for multiple-instance learning. In *In Neural Information Processing Systems*, pages 561–568, MIT Press, 2003.

[2] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez. Decentralized adaptive output feedback design for large-scale nonlinear systems. *Artificial Intelligence*, 89:31ĺC71, 1997.

[3] Jayadeva, R. Khemchandani, and S. Chandra. Twin support vector machines for pattern classification. *IEEE Trans.PatternAnal. Machine Intell*, 29(5):905–910, 2007.

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

[5] O. Maron and A. L. Ratan. Multiple-instance learning for natural scene classification. In *15th International Conference on Machine Learning*, San Francisco, CA, 1998.

[6] T. M. Mitchell. *Machine learning*. McGraw-Hill International, ingapore, 1997.

[7] P. M. Murphy and D. W. Aha. *UCI machine learning repository*, www.ics.uci.edu/ mlearn/mlrepository.html edition, 1992.

[8] J. Ramon and L. De Raedt. Multi-instance neural networks. In *Proceedings of International Conference on Machine Learning*, 2000.

[9] S. Ray and M. Craven. Supervised versus multiple instance learnining: An emperical comparison. In *Proceedings of the 22nd International Conference on Machine Learning*, Bonn, Germany, 2005.

[10] O. Seref and O. E. Kundakcioglu. Multiple instance classification with relaxed support vector machines. In *Proceedings of the 3rd INFORMS Workshop on Data Mining and Health Informatics (DM-HI)*, 2008.

[11] Z. X. Yang and N.Y. Deng. Multi-instance support vector machine based on convex combination. In *The Eighth International Symposium on Operations Research and Its Applications*, pages 481–487, 2009.

[12] Q. Zhang and S. Goldman. Em-dd: an improved multiple instance learning technique. In *In Neural Information Processing Systems*, 2001.