A NEW SEMI-SUPERVISED CLASSIFIER BASED ON
MAXIMUM VECTOR-ANGULAR MARGIN

LIMING YANG* AND YANNAN CHAO

College of Science, China Agricultural University
No.17 Tsing Hua East Road, Hai Dian District, Beijing 100083, China

(Communicated by Kok Lay Teo)

ABSTRACT. Semi-supervised learning is an attractive method in classification problems when insufficient training information is available. In this investigation, a new semi-supervised classifier is proposed based on the concept of maximum vector-angular margin, (called $S^3MAMC$), the main goal of which is to find an optimal vector $c$ as close as possible to the center of the dataset consisting of both labeled samples and unlabeled samples. This makes $S^3MAMC$ better generalization with smaller VC (Vapnik-Chervonenkis) dimension. However, $S^3MAMC$ formulation is a non-convex model and therefore it is difficult to solve. Following that we present two optimization algorithms, mixed integer quadratic program (MIQP) and DC (difference of convex functions) program algorithms, to solve the $S^3MAMC$. Compared with the supervised learning methods, numerical experiments on real and synthetic databases demonstrate that the $S^3MAMC$ can improve generalization when the labelled samples are relatively few. In addition, the $S^3MAMC$ has competitive experiment results in generalization compared to the traditional semi-supervised classification methods.

1. Introduction. In many classification problems, labelled data are scarce, while manually labelled data are often expensive. In contrast, unlabelled data are abundant and easy to collect. Moreover, when there are relatively few labelled data, a frequent drawback in classification is the possibility of overfitting the training data with a consequent loss of generality. Using both labelled and unlabelled data for the purpose of learning is called semi-supervised learning, the main goal of which is to improve generalization by incorporating unlabelled data in training when insufficient training information is available. Recently, semi-supervised learning has become an important topic in both theory and practice. Semi-supervised support vector machine ($S^3VM$) [4,6,8,14] is a popular semi-supervised learning framework and has demonstrated its effectiveness in machine learning. However, the parameters for the $S^3VM$ balance only the model complexity and misclassification errors, and thus do not permit other theoretical interpretation.

For different data sets, the amount of data points, the number of data dimensions and the shapes of data distributions are quite different. Therefore different models...
are used to solve problems in different data sets. Recently, a maximum vector-angular margin classifier (MAMC) [5] has been proposed based on a new concept of vector-angular margin. The main idea of MAMC is to find a vector $c$ as close as possible to the data center in the sense of smaller VC dimension [17]. Unlike the standard support vector machine (SVM) [16,17] where the number of support vectors cannot be controlled by certain parameters, MAMC can effectively control the number of support vectors by model parameter $\nu$. However, this supervised MAMC requires a large number of labelled data in order to construct accurate classifier.

In this work, we extend the MAMC to semi-supervised learning framework and a new semi-supervised classifier is proposed based on the maximum vector-angular margin (called $S^3$MAMC). The advantages of doing so are twofold:

(1) $S^3$MAMC is to find an optimal vector $c$ approximate to the center of the whole dataset including both labelled data and unlabelled data. Thus the $S^3$MAMC has better generalization with smaller VC dimension than the supervised MAMC since the upper bound of the VC dimension depends theoretically on the minimal radius of the smallest sphere enclosing all samples including labelled samples and unlabelled samples.

(2) $S^3$MAMC inherits the advantage of MAMC and $\nu$-support vector classifier ($\nu$-SVC) [12]. It can effectively control the number of support vectors by model parameter $\nu$, which make the parameter $\nu$ in $S^3$MAMC have theoretical better interpretation than the penalty factors in SVM and $S^3$VM models.

However, $S^3$MAMC model is nonconvex and nonsmooth, which makes the problem difficult to solve. In this work, we first give a mixed integer quadratic programming (MIQP) algorithm to obtain exact solution of $S^3$MAMC. Following that, we reformulate $S^3$MAMC as a DC (difference of convex functions) program [1,2,5,11,13]. Moreover, the resulting DCA (DC algorithm) converges finitely and only requires solving one quadratic program at each iteration.

2. Background. We consider here binary classification problem. Suppose the training set consists of $m$ labelled samples $\{(x_i, y_i), x_i \in \mathbb{R}^n, y_i = \pm 1, i = 1, 2, \ldots, m\}$ and the test set consists of $p$ unlabelled samples $\{x_j \in \mathbb{R}^n, j = m+1, m+2, \ldots, m+p\}$.

2.1. $\nu$-SVC. The standard SVM is a supervised learning method. It classifies two-class datasets by constructing a separating hyperplane $H(w, b) = w^T x + b$ with maximum distance (margin) $\frac{2}{\|w\|_2}$, and the corresponding primal optimization problem is mathematically modeled as

$$\begin{align*}
\min_{w, b, \xi} & \quad \frac{1}{2} \|w\|_2^2 + \lambda \sum_{i=1}^{m} \xi_i \\
\text{s.t.} & \quad y_i(w^T x_i + b) + \xi_i \geq 1 \\
& \quad \xi_i \geq 0, i = 1, \ldots, m
\end{align*}$$

where $\lambda > 0$ does not permit any other theoretical interpretation except for trade off between model complexity and empirical risk. Scholkopf et al propose a modification of the SVM model, called $\nu$-SVC [12], by introducing a new parameter $\nu$ and an
additional variable $\rho$:

$$\min_{w, b, \xi, \rho} \frac{1}{2} \|w\|_2^2 - \nu \rho + \lambda \sum_{i=1}^{m} \xi_i$$

s.t. $y_i (w^T x_i + b) + \xi_i \geq \rho$  \hspace{1cm} (2)

\[\xi_i, \rho \geq 0, 1 \leq i \leq m\]

Where the parameter $\nu$ controls the number of support vectors and the bound of the classification errors. The corresponding decision hyperplane takes still the form: $f(x) = \text{sgn}(w^T x + b)$, but the separating margin is changed to $\frac{2\rho}{\|w\|_2^2}$, which is different from the traditional SVM. The dual of problem (2) is usually solved.

2.2. Maximum vector-angular margin classifier (MAMC). The MAMC is a supervised classification method. By finding an optimal vector $c$ using labelled samples, MAMC separates two-class samples in terms of the maximum vector-angular margin between the vector $c$ and the labelled samples. Hu et al define MAMC as the following optimization [10]:

$$\min_{\rho, c} -\nu \rho + \frac{1}{m} \sum_{i=1}^{m} (x_i - c)^T (x_i - c) + \lambda \sum_{i=1}^{m} \xi_i$$

s.t. $y_i (x_i^T c) \geq \rho - \xi_i, \xi_i \geq 0, 1 \leq i \leq m$  \hspace{1cm} (3)

where $\xi_i$ is a slack variable and the second item in objective function is to make vector $c$ as close as possible to the center of the labelled data. The $\rho$ is known as the margin in the sense of vector-angular distance, which is not a Euclidean distance.

According to the principle of Euclidean geometry, $x_i^T c = \cos(\theta_i)\|x_i\|\|c\|$, where $\theta_i$ is the angle between $x_i$ and $c$. Therefore, $x_i^T c$ actually reflects the information about the angular and Euclidean distance between $x_i$ and $c$. The corresponding decision function is $f(x) = \text{sgn}(x^T c)$.

The dual problem of MAMC (3) has the form:

$$\max_{\alpha} \sum_{i=1}^{m} \alpha_i y_i (\frac{4}{m} \sum_{j=1}^{m} x_j^T x_i - \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j x_i^T x_j$$

s.t. $\sum_{i=1}^{m} \alpha_i = \nu, 0 \leq \alpha_i \leq \lambda, i = 1, \cdots m$  \hspace{1cm} (4)

We see that parameter $\nu > 0$ exactly controls the sum of support vectors, which implies that MAMC inherits the main advantage of $\nu$-SVC. Furthermore, parameter $\nu$ has the following theoretical interpretations [12]

Suppose $\rho > 0$, then the following two statements hold:

1. $\nu / (m\lambda)$ is the lower bound of the fraction of support vectors.
2. $\nu / (m\lambda)$ is the upper bound of the fraction of misclassified samples.

2.3. DC programming. We outline the main algorithmic results for DC programming [1,2,5,11,13]. DC programming and DCA, introduced by Pham Dinh Tao in 1985, constitute the backbone of nonconvex continuous programming. The key to DC programs is to decompose an objective function into the difference of two convex (DC) functions, from which a sequence of approximations of the objective function yields a sequence of solutions converging to a stationary point, possibly an optimal
solution. Generally speaking, a so-called DC program ($P_{dc}$) is to minimize a DC function:

$$f(x) = g(x) - h(x), x \in \mathbb{R}^n \ (P_{dc})$$  \hspace{1cm} (5)

where $g(x)$ and $h(x)$ are all convex functions. Let $g^*(y) = \sup\{x^T y - g(x), x \in \mathbb{R}^n\}$ denote the conjugate function of $g$. The Fenchel-Rockafellar dual of $(P_{dc})$ is defined as

$$\inf\{h^*(y) - g^*(y), y \in \mathbb{R}^n\} \ (D_{dc})$$  \hspace{1cm} (6)

A DC program is called a polyhedral DC program when either $g(x)$ or $h(x)$ is a polyhedral convex function (i.e., the pointwise supremum of a finite collection of affine functions).

The DCA is an iterative algorithm based on local optimality conditions and duality. The idea of DCA is simple (for simplicity, we omit here the dual part): at each iteration, one replaces in the primal DC problem $(P_{dc})$ the second component $h$ by its affine minorization: $h(x_k) + (x - x_k)^T y_k$, to generate the convex program:

$$\minimize \ : \ \{g(x) - h(x_k) - (x - x_k)^T y_k, x \in \mathbb{R}^n, y_k \in \partial h(x_k)\}$$  \hspace{1cm} (7)

Where $\partial h$ is the subdifferential of convex function $h$. In practice, a simplified form of the DCA is used. Two sequences $\{x_k\}$ and $\{y_k\}$ satisfying $y_k \in \partial h(x_k)$ are constructed, and $x_{k+1}$ is a solution to the convex program [7].

DCA is a descent method without line search, and it converges linearly for general DC programs. In particular, for polyhedral DC programs, the sequence $\{x_k\}$ contains finitely many elements, and in a finite number of iterations the algorithm converges to a stationary point satisfying the necessary optimality condition.

3. Semi-supervised maximum vector-angular margin classifier (S$^3$MAMC). When the dataset lacks of labelled samples, the supervised MAMC is difficult to apply. Therefore, motivated by the idea from the supervised MAMC and the traditional S$^3$VM [4], we here propose a new semi-supervised classification method based on the concept of the maximum vector-angular margin, called S$^3$MAMC, the main idea of which is to find a optimal vector $c$ as close as possible to the center of the whole dataset consisting of labelled data and unlabelled data. This implies that the S$^3$MAMC has better generalization than MAMC and $\nu$-SVC in the perspective of VC dimension since the upper bound of the VC dimension depends theoretically on the minimal radius of the smallest sphere enclosing all samples including labelled samples and unlabelled samples.

Specifically, for each unlabelled sample $x_j$, we add two constraints and introduce two variables $r_j$ and $s_j$ to represent two possible misclassification errors. One constraint calculates the misclassification error $r_j$ as if the sample $x_j$ is in positive class and the other constraint calculates the misclassification error $s_j$ as if the sample $x_j$ is in negative class. The objective function defines the minimum of the two possible misclassification errors $\min\{r_j, s_j\} (j = m + 1, \ldots, m + p)$. The final class of the unlabelled sample $x_j$ corresponds to the one that results in the smallest error. The above motivation can be formulated as the following optimization problem,
called $S^3$MAMC

$$\begin{align*}
\min_{c, \rho, \xi, r, s, \lambda, \mu} & \quad \ - \nu \rho + \frac{1}{m+p} \sum_{i=1}^{m+p} (x_i - c)^T (x_i - c) + \lambda \sum_{i=1}^{m} \xi_i + \mu \sum_{j=m+1}^{m+p} \min \{r_j, s_j\} \\
\text{s.t.} & \quad y_i(x_i^T c) + \xi_i \geq \rho, \quad \xi_i \geq 0 \\
& \quad x_j^T c + r_j \geq \rho, \quad r_j \geq 0 \\
& \quad -x_j^T c + s_j \geq \rho, \quad s_j \geq 0 \\
& \quad i = 1, \ldots, m, \quad j = m+1, \ldots, m+p
\end{align*}$$

where $\lambda, \mu > 0$ are penalty parameters for misclassification of labelled and unlabelled samples. The variable $\rho$ is vector-angular margin between two-class samples. The first three terms of objective function together with the first two constraints are corresponding to a supervised MAMC. The last item of objective function together with the last four constraints are to classify each unlabelled sample to the category with minimum misclassification. The decision function is $f(x) = \text{sgn}(x^T c)$.

However, $S^3$MAMC (8) is a nonconvex and nonsmooth problem owing to the last term in the objective function, which precludes the use of convex and smooth methods. Thus, this makes the problem difficult to optimize.

**The following should be noted:**

1. $S^3$MAMC extends the supervised MAMC to the semi-supervised learning setting. When the unlabelled sample set is null, $S^3$MAMC is equivalent to the supervised MAMC. Thus $S^3$MAMC includes and extends the MAMC.
2. The goal of $S^3$MAMC is to find a vector $c$ as close as possible to the center of the whole dataset including labelled data and unlabelled data, and therefore it has better generalization with smaller VC dimension than $\nu$-SVC and the supervised MAMC.
3. Similar to the MAMC and $\nu$-SVC, $S^3$MAMC can well control the number of support vectors and the classification error by introducing an additional variable $\rho$ and a new parameter $\nu$.
4. Compared with the popular semi-supervised methods, $S^3$VMs, the main benefit of $S^3$MAMC is that the parameter $\nu$ in the $S^3$MAMC has a better theoretical interpretation than the penalty factors in $S^3$VMs.

### 3.1. Solving $S^3$MAMC via mixed integer quadratic programming (MIQP-$S^3$MAMC)

In this section, we present a combinatorial optimization technique to globally solve the $S^3$MAMC. Specifically, by introducing integer variable $d_j \in \mathbb{R}^p$ with component $d_{j} = 0$ or 1 for each unlabelled sample $x_j, (j = m+1, \ldots, m+p)$, the $S^3$MAMC (8) can be posed as the following mixed integer quadratic programming, called MIQP-$S^3$MAMC:

$$\begin{align*}
\min_{c, \rho, \xi, r, s, d} & \quad \ - \nu \rho + \frac{1}{m+p} \sum_{i=1}^{m+p} (x_i - c)^T (x_i - c) + \lambda \sum_{i=1}^{m} \xi_i + \mu \sum_{j=m+1}^{m+p} (r_j + s_j) \\
\text{s.t.} & \quad y_i(x_i^T c) + \xi_i \geq \rho, \quad \xi_i \geq 0, \quad i = 1, \ldots, m \\
& \quad x_j^T c + r_j + M(1 - d_j) \geq \rho, \quad r_j \geq 0 \\
& \quad -x_j^T c + s_j + Md_j \geq \rho, \quad s_j \geq 0 \\
& \quad d_j = \{0, 1\}, \quad j = m+1, \ldots, m+p
\end{align*}$$

where $M > 0$ is a sufficiently large constant such that if $d_j = 0$ then $r_j = 0$ is feasible for any optimal $c$, which attempts to classify the unlabelled sample $x_j$ to
negative class. Likewise if \( d_j = 1 \) then \( s_j = 0 \), which attempts to classify the sample \( x_j \) to positive class. Solving the problem \([9]\) obtains the exact solution of S\(^3\)MAMC \([8]\) using the popular YALMIP Toolbox \([15]\). In general, this global optimization can be computationally very demanding.

According to the above analysis, the algorithm for solving MIQP-S\(^3\)MAMC \([9]\) is described as follows

**Algorithm 1**

1. Choose a sufficiently large constant \( M > 0 \) and suitable parameters \( \lambda, \mu > 0 \).
2. Construct the nonconvex S\(^3\)MAMC \([8]\) and the corresponding MIQP-S\(^3\)MAMC \([9]\).
3. Solve MIQP-S\(^3\)MAMC \([9]\) to obtain point \((c^*, \rho^*, \xi^*, r^*, s^*, d^*)\).

**Theorem 3.1.** Suppose point \((c^*, \rho^*, \xi^*, r^*, s^*, d^*)\) is an optimal solution of MIQP-S\(^3\)MAMC \([9]\), where \( M > 0 \) is a sufficiently large constant, then the corresponding \((c^*, \rho^*, \xi^*, r^*, s^*)\) is an exact solution of S\(^3\)MAMC \([8]\).

**Proof.** Assume that \((c^*, \rho^*, \xi^*, r^*, s^*, d^*)\) is an optimal solution of the MIQP-S\(^3\)MAMC \([9]\), where \( M > 0 \) is a sufficiently large number.

If \( d_j = 0 \), then the constraint condition

\[
x_j^T c + r_j + M \geq \rho, \quad r_j \geq 0, \quad j = m + 1, \ldots, m + p
\]

is redundant and \( r_j^* = 0 \) is feasible for any optimal \( c \). Moreover, the \((c^*, \rho^*, \xi^*, r^*, s^*, d^*)\) is an optimal solution of the following problem:

\[
\begin{align*}
\min_{c, \rho, \xi, r, s} & \quad -\nu \rho + \frac{1}{m + p} \sum_{i=1}^{m+p} (x_i - c)^T(x_i - c) + \lambda \sum_{i=1}^{m} \xi_i + \mu \sum_{j=m+1}^{m+p} s_j \\
\text{s.t.} & \quad y_i(x_i^T c) + \xi_i \geq \rho, \quad \xi_i \geq 0, \quad i = 1, \ldots, m \\
& \quad -x_j^T c + s_j \geq \rho, \quad s_j \geq 0, \quad j = m + 1, \ldots, m + p
\end{align*}
\]

which implies to classify the unlabelled sample \( x_j \) to negative class.

If \( d_j = 1 \), then the constraint condition

\[
-x_j^T c + s_j + M \geq \rho, \quad s_j \geq 0, \quad j = m + 1, \ldots, m + p
\]

is redundant and \( s_j^* = 0 \) is feasible for any optimal \( c \). Moreover, the \((c^*, \rho^*, \xi^*, r^*, s^*)\) is an optimal solution of the following problem:

\[
\begin{align*}
\min_{c, \rho, \xi, r} & \quad -\nu \rho + \frac{1}{m + p} \sum_{i=1}^{m+p} (x_i - c)^T(x_i - c) + \lambda \sum_{i=1}^{m} \xi_i + \mu \sum_{j=m+1}^{m+p} r_j \\
\text{s.t.} & \quad y_i(x_i^T c) + \xi_i \geq \rho, \quad \xi_i \geq 0, \quad i = 1, \ldots, m \\
& \quad x_j^T c + r_j \geq \rho, \quad r_j \geq 0, \quad j = m + 1, \ldots, m + p
\end{align*}
\]

which attempts to classify the unlabelled sample \( x_j \) to positive class.

Final, two problems \([11]\) and \([13]\) have merged into one optimization problem

\[
\begin{align*}
\min_{c, \rho, r, s} & \quad -\nu \rho + \frac{1}{m + p} \sum_{i=1}^{m+p} (x_i - c)^T(x_i - c) + \lambda \sum_{i=1}^{m} \xi_i + \mu \sum_{j=m+1}^{m+p} \min \{r_j, s_j\} \\
\text{s.t.} & \quad y_i(x_i^T c) + \xi_i \geq \rho, \quad \xi_i \geq 0 \\
& \quad x_j^T c + r_j \geq \rho, \quad r_j \geq 0 \\
& \quad -x_j^T c + s_j \geq \rho, \quad s_j \geq 0 \\
& \quad i = 1, \ldots, m, \quad j = m + 1, \ldots, m + p
\end{align*}
\]
which is very $S^3$MAMC \cite{6}. Thus if $(c^*, \rho^*, \xi^*, r^*, s^*, d^*)$ is an optimal solution of the MIQP-$S^3$MAMC \cite{6}, then the corresponding $(c^*, \rho^*, \xi^*, r^*, s^*)$ is the exact solution of $S^3$MAMC \cite{6}.

3.2. Solving $S^3$MAMC via DC programming (DCA-$S^3$MAMC). In this section, we discuss an approximation algorithm for solving the $S^3$MAMC \cite{6}. By suitable decomposition of the objective function, the $S^3$MAMC \cite{6} can be transformed into a DC program.

Let $x = (\rho, c, \xi, r, s)$ and $\Omega$ denote the feasible set of constraints for problem \cite{6}. The $S^3$MAMC \cite{6} can be expressed as the following unconstrained optimization problem, called DCA-$S^3$MAMC:

$$\min_x \{F(x) = G(x) - H(x)\}$$

with

$$G(x) = -\nu \rho + \frac{1}{m + p} \sum_{i=1}^{m+p} (x_i - c)^T (x_i - c) + \lambda \sum_{i=1}^{m} \xi_i + \chi_\Omega(x)$$

$$H(x) = -\mu \sum_{j=m+1}^{m+p} \min\{r_j, s_j\}$$

where $\chi_\Omega(x)$ denotes the indicator function for the set $\Omega$: $\chi_\Omega(x) = 0$ if $x \in \Omega$, and $+\infty$ otherwise. Note that $G(x)$ is a convex function and $H(x)$ is a polyhedral convex function. Thus DCA-$S^3$MAMC \cite{6} is a polyhedral DC program.

According to the analysis in Section 2, performing DCA for problem \cite{6} amounts to computing two sequences $\{x^{k+1}\}$ and $\{y^k\}$, where $\{x^{k+1}\}$ is the solution to the following convex programming:

$$(P_k) \quad \min_{x \in \mathbb{R}^n} \{G(x) - ((H(x^k) + (x - x^k)^T y^k))\}$$

with $y^k \in \partial H(x^k)$. The subgradient of convex function $H(x)$ is below:

$$\partial H(x) = \partial(-\mu \sum_{j=m+1}^{m+p} \min\{r_j, s_j\})$$

$$= \mu \sum_{j=m+1}^{m+p} \left\{ \begin{array}{ll}
(0, 0, 0, -I_j, 0), & r_j < s_j \\
(0, 0, 0, 0, -I_j), & r_j > s_j \\
(0, 0, 0, 0, 0), & r_j = s_j
\end{array} \right\}$$

$I_j$ means the $j$th column of unit matrix $I$. To sum up, we get DCA for solving the problem DCA-$S^3$MAMC \cite{6}.

**Algorithm 2**

(1) $\forall \varepsilon > 0$ is a sufficiently small number, and set $k = 0$. Choose an initial point $x^0 \in \Omega$.

(2) Calculate $y^k \in \partial H(x^k)$ according to (19) and solve the quadratic programming

$$\min_{c, \rho, \xi, r, s} \quad -\nu \rho + \frac{1}{m + p} \sum_{i=1}^{m+p} (x_i - c)^T (x_i - c) + \lambda \sum_{i=1}^{m} \xi_i - (y^k)^T x$$

s.t.

$$y_i (x_i^T c) + \xi_i \geq \rho, \quad \xi_i \geq 0$$

$$x_i^T c + r_j \geq \rho, \quad r_j \geq 0$$

$$-x_j^T c + s_j \geq \rho, \quad s_j \geq 0$$

$$i = 1, \cdots, m, \quad j = m + 1, \cdots, m + p$$
to get $x^{k+1}$.

(3) If $\|x^{k+1} - x^k\| \leq \varepsilon$ or $G(x^{k+1}) - H(x^{k+1}) \geq G(x^k) - H(x^k) - \varepsilon$, then the iteration stops, and $x^{k+1}$ is the computed solution as required; otherwise, set $k = k + 1$, to step (2).

**Theorem 3.2.** (1) DCA generates two sequences $\{x^k\}$ and $\{y^k\}$ such that $G(x^k) - H(x^k)$ and $H^*(y^k) - G^*(y^k)$ decrease monotonically.

(2) After a finite number of iterations, the sequences $\{x^k\}$ and $\{y^k\}$ converge to $x^*$ and $y^*$ respectively. The point $x^*$ is the stationary point of the DC program (14).

**Proof.** The first conclusion is a direct consequence of the convergence properties of the general DC program. Note that DC program (14) is a DC polyhedral program, and hence algorithm 2 converges finitely, and its limit point is a stationary point of the DC program (14). The proof is then complete.

4. Numerical experiments. To evaluate the proposed framework, we run the proposed $S^3$MAMC on a synthetic dataset and some the real world data sets from the University of California Irvine (UCI) Machine Learning Repository [3]. Experiments use Matlab R2012a as a solver.

4.1. **Experimental design.** The performances of the proposed method is assessed by the following criteria:

ACC: classification accuracy of all samples from two classes, G-ACC: geometric accuracy, MCC: Mathew correlation coefficient and $F_1$-measure. The above values can be obtained from the decision function and are defined as [7]

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}, \quad G - ACC = \sqrt{a^+ \times a^-}$$

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

$$F_1 - measure = \frac{2 \times TP}{2 \times TP + FP + FN}$$

where TP and TN denote true positives and true negatives; FN and FP denote false negatives and false positives, respectively. And $a^+ = TP / (TP + FN), a^- = TN / (TN + FP)$. The G-ACC, MCC and $F_1$-measure are comprehensive measures of the quality of classification models. The higher the values of these values are, the better the models are.

For comprehensive evaluation, our experiment is composed of three parts.

(1) We compare the performance of $S^3$MAMC with the supervised learning methods.

(2) We compare the two proposed $S^3$MAMC algorithms: MIQP-$S^3$MAMC and DCA-$S^3$MAMC.

(3) We compare two $S^3$MAMC algorithms with other semi-supervised learning models.

In addition, we conduct 10 times trials for each dataset. In each trial, each dataset is randomly divided into two parts: 20% of the samples for training and the remaining 80% of the samples for testing. We remove the labels for the test set, and run the proposed semi-supervised algorithms to reclassify the test set using the learning results.

$S^3$MAMC parameter $\lambda$ balances the model complexity and the empirical risk of labelled samples, parameter $\mu$ for the unlabelled samples. The parameter $\nu$ controls the sum of support vectors and the bound of classification errors.
Figure 1. ACC versus $\mu=1$ for $\nu=1$ and 10 on Thyroid data

Figure 2. Training samples of the synthetic data

The accuracy of the proposed method depends usually on these parameter values. In this work, these parameters are adjusted from the set $\{10^i | i = -2, \ldots, 3\}$ by cross-validation. We also present a map (FIGURE 1) to illustrate the relationship between the ACC and the $\mu$ with $\nu = 1$ and $\nu = 10$, where the x-axis denotes the values of parameter $\mu$, and the y-axis denotes the accuracy (ACC). For fixed $\nu$, we find that the accuracy of the $S^3$MAMC increases when $\mu$ ranges from 1 to 100, while the $S^3$MAMC produces smaller accuracy when $\mu$ ranges from 100 to 1000. At the same time, FIGURE 1 shows that the ACC is not very sensitive with the choice of parameter $\nu$. These findings are helpful in the choice of parameters $\mu$ and $\nu$ in the following benchmark experiments.

According to the above analysis, $S^3$MAMC parameters $\nu = 10$ and $\lambda = \mu = 100$ are chosen in our experiments. The parameters in the other models are set to be the same as in $S^3$MAMC.

Moreover, a synthetic banana shaped dataset is shown in FIGURE 2, which is a linearly separable two-dimensional dataset. The goal of designing this dataset is to support the superiority of $S^3$MAMC more intuitively. Moreover, the positive samples are represented by "o" and the negative samples are represented by "*".
| data         | Classification models | G-ACC (%) | ACC (%) | MCC (%) | $F_1$-measure (%) |
|--------------|------------------------|-----------|---------|---------|------------------|
| DCA-$S^3$MAMC | 100                    | 100       | 100     | 100     |
| MIQP-$S^3$MAMC | 100                    | 100       | 100     | 100     |
| Wine         | MAMC                  | 94.31     | 94.39   | 89.06   | 94.16            |
|              | $\nu$-SVM             | 99.30     | 99.30   | 98.62   | 99.31            |
| DCA-$S^3$MAMC | 99.81                  | 99.81     | 99.63   | 99.81   |
| MIQP-$S^3$MAMC | 95.55                  | 95.65     | 91.64   | 95.45   |
| Tryroid      | MAMC                  | 94.87     | 95.00   | 90.45   | 95.24            |
|              | $\nu$-SVM             | 87.94     | 88.34   | 77.76   | 89.23            |
| DCA-$S^3$MAMC | 92.69                  | 93.76     | 86.73   | 93.49   |
| MIQP-$S^3$MAMC | 96.35                  | 96.35     | 91.20   | 96.16   |
| Cancer       | MAMC                  | 92.26     | 93.36   | 86.03   | 93.02            |
|              | $\nu$-SVM             | 91.27     | 91.46   | 85.30   | 90.93            |
| DCA-$S^3$MAMC | 62.40                  | 62.40     | 24.81   | 62.65   |
| MIQP-$S^3$MAMC | 62.42                  | 63.02     | 26.44   | 65.98   |
| Sonar        | MAMC                  | 60.11     | 60.41   | 20.97   | 62.67            |
|              | $\nu$-SVM             | 60.11     | 60.19   | 20.40   | 58.99            |
|              | $\nu$-SVM             | 60.11     | 60.19   | 20.40   | 58.99            |
| DCA-$S^3$MAMC | 86.65                  | 87.61     | 75.00   | 87.98   |
| MIQP-$S^3$MAMC | 85.34                  | 86.69     | 75.91   | 88.20   |
| Ionosphere   | MAMC                  | 80.94     | 81.25   | 63.14   | 82.49            |
|              | $\nu$-SVM             | 86.34     | 86.74   | 74.51   | 87.76            |
| DCA-$S^3$MAMC | 72.82                  | 73.66     | 48.51   | 76.28   |
| MIQP-$S^3$MAMC | 78.26                  | 78.65     | 58.00   | 80.19   |
| Hepatitis    | MAMC                  | 70.60     | 71.76   | 45.04   | 74.98            |
|              | $\nu$-SVM             | 70.00     | 71.21   | 43.93   | 74.98            |
| DCA-$S^3$MAMC | 86.39                  | 86.40     | 72.80   | 86.54   |
| MIQP-$S^3$MAMC | 84.72                  | 84.79     | 69.76   | 85.31   |
| Heart        | MAMC                  | 81.82     | 81.86   | 63.83   | 82.35            |
|              | $\nu$-SVM             | 84.70     | 84.72   | 69.46   | 84.95            |
| DCA-$S^3$MAMC | 94.78                  | 94.81     | 89.69   | 94.91   |
| MIQP-$S^3$MAMC | 95.61                  | 95.62     | 91.29   | 95.69   |
| Vote         | MAMC                  | 90.11     | 90.29   | 81.07   | 90.80            |
|              | $\nu$-SVM             | 94.51     | 94.54   | 89.09   | 94.60            |
| DCA-$S^3$MAMC | 92.50                  | 92.50     | 85.00   | 92.54   |
| MIQP-$S^3$MAMC | 93.25                  | 93.25     | 86.50   | 93.23   |
| Synthesis    | MAMC                  | 86.45     | 86.49   | 73.07   | 86.82            |
|              | $\nu$-SVM             | 82.50     | 82.50   | 65.00   | 82.41            |

4.2. **Comparisons of $S^3$MAMC and the supervised learning methods.** We first compare the $S^3$MAMC with supervised learning methods on eight UCI datasets and a synthesis data. We chose MAMC and $\nu$-SVM as the baseline methods. With the ratio of labelled to unlabelled samples being 2:8, experimental results are averaged over 10-time trials, and the results are reported in TABLE 1.
TABLE 2. Comparison of $S^3$MAMC, MAMC and $\nu$-SVM with the ratio of labelled to unlabelled samples being 1:9 in terms of accuracy (ACC)

| models  | DCA-$S^3$MAMC (%) | MAMC (%) | $\nu$-SVM (%) |
|---------|-------------------|----------|---------------|
| Tryroid | 92.59             | 85.19    | 86.29         |
| Ionosphere | 83.37         | 71.43    | 71.74         |
| Sonar  | 60.45             | 55.56    | 53.89         |
| Cancer | 93.15             | 89.88    | 84.52         |
| Heart  | 85.19             | 74.31    | 75.49         |
| Hepatitis | 73.33         | 64.44    | 71.11         |
| Vote   | 93.65             | 86.95    | 89.68         |
| Synthesis | 91.56         | 70.39    | 63.66         |

We find from TABLE 1 that $S^3$MAMC outperforms MAMC and $\nu$-SVM in five datasets, Wine, Cancer, Tryroid, Sonar and Synthesis datasets, especially synthetic dataset; in other four datasets, the performance of $S^3$MAMC is competitive with or better lightly than that of MAMC and $\nu$-SVM according to generalization. These results show that $S^3$MAMC either improves or shows no significant difference in generalization compared to the supervised approach when lacking of adequate labelled samples.

To further examine the performance of the proposed $S^3$MAMC for large unlabelled samples, we present another numerical experiment result. Specifically, each dataset is randomly separated into: 10% of samples as the labelled samples and 90% of samples as unlabelled. We perform the DCA-$S^3$MAMC, MAMC and $\nu$-SVM in eight datasets. With the ratio of labelled to unlabelled samples being 1:9, experimental results are averaged over 10-time trials, and the generalization of $S^3$MAMC and the supervised learning methods are reported in TABLE 2.

As expected, TABLE 2 shows that the performance of the DCA-$S^3$MAMC outperforms MAMC and $\nu$-SVM in all eight considered datasets. This suggests that incorporating unlabelled samples in training improves the generalization when the proportion of labelled samples in the dataset is relatively low.

4.3. Comparisons of the two proposed semi-supervised MAMC algorithms. We here compare the proposed two semi-supervised MAMC models: MIQP-$S^3$MAMC and DCA-$S^3$MAMC. The accuracy (ACC) and time-consuming (CPU-time, in seconds) of these two models are illustrated in FIGURE 3 and FIGURE 4 respectively.

We know from FIGURE 3 that MIQP-$S^3$MAMC requires more running time than DCA-$S^3$MAMC. At the same time, FIGURE 4 illustrates that the accuracy of these two algorithms has no very significant difference in the considered datasets.

4.4. Comparisons of $S^3$MAMC with other semi-supervised learning methods. Then we compared the two proposed $S^3$MAMC algorithms with two typical semi-supervised SVM algorithms: MILP-$S^3$VM [4] and VS$^3$VM [8]. The classification accuracy (ACC) of the 10-time trials on six datasets is reported in TABLE 3.

We know from TABLE 3 that the two proposed $S^3$MAMC models have obvious advantages over MIP-$S^3$VM and VS$^3$VM in Hepatitis and synthesis datasets; the performances of two $S^3$MAMC algorithms are inferior to MIP-$S^3$VM and VS$^3$VM
in Sonar and Ionosphere datasets; while in other two datasets, the performances of these four semi-supervised learning methods show no significant difference. However, for the synthetic banana shaped dataset, the $S^3$MAMC is superior to the MIP-$S^3$VM and VS$^3$VM by obtaining better generalization.
5. Conclusions. In this paper, we extend the supervised MAMC to semi-supervised learning framework. A novel semi-supervised classifier (called $S^3$MAMC) is proposed based on the concept of vector-angular margin. This makes $S^3$MAMC better generalization with smaller VC dimension. And $S^3$MAMC can control the number of support vectors and the classification errors by model parameter $\nu$. Following that, we show how the $S^3$MAMC can be converted to a mixed-integer quadratic program and then solved exactly. Moreover, the $S^3$MAMC is posed as a DC program by suitable decomposition of the objective function. The resulting DCA converges finitely and only requires solving one quadratic program at each iteration.

Computational results of $S^3$MAMC are compared with the supervised methods on real and synthetic databases. The results show that $S^3$MAMC can improve the generalization by incorporating unlabelled samples in training when there are relatively few labelled data. Moreover, comparing with other semi-supervised methods, $S^3$MAMC has competitive results in UCI datasets. For the synthetic banana shaped dataset, the $S^3$MAMC outperforms significantly the supervised methods (MAMC and $\nu$-SVM) and the traditional semi-supervised methods (MIP-$S^3$VM and VS$^3$VM) by obtaining better ability.

The proposed methods in this work can also be extended to formulate multi-class and nonlinear cases. In addition, it is worth noting that the proposed DCA depends greatly on two components ($g$ and $h$). The question of "good" DC decompositions for the $S^3$MAMC will be studied in the future work.

Acknowledgments. This work is supported by National Nature Science Foundation of China (11471010,11271367). Moreover, the authors thank the referees and the editors for their constructive comments. Their suggestions improved the paper significantly.

REFERENCES

[1] L. T. H. An and P. D. Tao, The DC (difference of convex functions) programming and DCA revisited with DC models of real world nonconvex optimization problems, Annals of Operations Research, 133 (2005), 23–46.
[2] L. T. H. An, H. M. Le, V. V. Nguyen and P. D. Tao, A DC programming approach for feature selection in support vector machines learning, Advances in Data Analysis and Classification, 2 (2008), 259–278.
[3] A. Asuncion and D. J. Newman, UCI machine learning repository, School of Information and Computer Sciences, University of California Irvine, 2007, http://www.ics.uci.edu/ mllearn/MLRepository.html.
[4] K. Bennett and A. Demiriz, Semi-supervised support vector machines, In Advances in Neural Information Processing Systems, MIT Press, Cambridge, 12 (1998), 368–374.
[5] W. Changzhi, L. Chaojie and L. Qiang, A DC programming approach for sensor network localization with uncertainties in anchor positions, Journal of Industrial and Management Optimization, 10 (2014), 817–826.
[6] O. Chapelle, V. Sindhwani and S. Keerthi, Optimization Techniques for Semi-Supervised Support Vector Machines, Journal of Machine Learning Research, 9 (2008), 203–233.
[7] T. Fawcett, An introduction to ROC analysis, Pattern Recognition Letters, 27 (2006), 861–874.
[8] G. Fung and O. Mangasarian, Semi-Supervised Support Vector Machines for Unlabeled Data Classification, Optimization methods & software, 15 (2001), 29–44.
[9] W. Guan and A. Gray, Sparse high-dimensional fractional-norm support vector machine via DC programming, Computational Statistics and Data Analysis, 67 (2013), 136–148.
[10] W. J. Hu, F. L. Chung and L. SH. Wang, The Maximum Vector-Angular Margin Classifier and its fast training on large datasets using a core vector machine, Neural Networks, 27 (2012), 60–73.
[11] P. D. Tao and L. T. T. An, Convex analysis approaches to DC programming: Theory, algorithms and applications, Acta Mathematica, 22 (1997), 287–367.

[12] B. Scholkopf, A. J. Smola, R. C. Williamson and P. L. Bartlett, New support vector algorithms, Neural Computation, 12 (2000), 1207–1245.

[13] X. Xiao, J. Gu, L. Zhang and S. Zhang, A sequential convex program method to DC program with joint chance constraints, Journal of Industrial and Management Optimization, 8 (2012), 733–747.

[14] L. M. Yang and L. SH. Wang, A class of smooth semi-supervised SVM by difference of convex functions programming and algorithm, Knowledge-Based Systems, 41 (2013), 1–7.

[15] YALMIP Toolbox. http://control.ee.ethz.ch/~joloef/wiki/pmwiki.php.

[16] Y. B. Yuan, Canonical duality solution for alternating support vector machine, Journal of Industrial and Management Optimization, 8 (2012), 611–621.

[17] V. N. Vapnik, Statistical Learning Theory, New York: Wiley.1998.

Received December 2014; 1st revision August 2015; final revision December 2015.

E-mail address: cauyanglm@163.com
E-mail address: chinachaoyn@163.com