Yokohama Publishers
ISSN 1349-8169 ONLINE JOURNAL

Pacific Journal of Optimization
Vol. 6, No. 1, pp. 115–140, January 2010

MULTIOBJECTIVE MULTICLASS SUPPORT VECTOR MACHINES MAXIMIZING GEOMETRIC MARGINS

Keiji Tatsumi, Kenji Hayashida, Ryo Kawachi and Tetsuzo Tanino

Abstract: In this paper, we focus on the *all together* model of the support vector machine (SVM) for multiclass classification, which constructs a piece-wise linear discriminant function. It is formulated as a single-objective optimization problem maximizing the sum of margins between all pairs of classes, which is defined as the distance between two normalized support hyperplanes parallel to the corresponding discriminant hyperplane where any sample is not contained. However, it is not necessarily equal to the geometric margin defined as the minimal distance of patterns in a pair of classes to the corresponding discriminant hyperplanes. Then, we formulate the proposed model as a multiobjective problem which maximizes all of the margins simultaneously. Moreover, we derive two kinds of single-objective second order cone programming (SOCP) problems based on scalarization approaches, Benson's method and ε-constraint method to solve the proposed multiobjective model, and show that the methods can find Pareto optimal solutions of the model. Furthermore, through numerical experiments we verify the generalization ability of discriminant functions obtained by the proposed SOCP problems.

Key words: *multiclass classification, support vector machine, maximization of margins, geometric margin, multiobjective optimization*

Mathematics Subject Classification: *62H30, 90C22, 90C29*

1 Introduction

The support vector machine (SVM) is a powerful machine learning method for binary-class classification problems. Some kinds of extensions to multiclass classification have been investigated [1, 9], which can be mainly classified into two kinds of approaches. One is constructing a discriminant function by training multiple binary SVMs and combining them and the other is finding a discriminant function directly by solving an optimization problem with all patterns. As the former approach, *one against others* and *one against one* methods are more commonly used [4, 11, 14], while as the latter approach, *all together* method is most popular [5, 8, 17, 18]. In this paper we focus on the all together method, where all patterns are classified into the corresponding classes by using a piece-wise linear function. Moreover, several improved or decomposition methods have been proposed [6, 9, 10, 15]. This model is formulated as a single-objective optimization problem of maximizing the sum of margins between all of the pairs of classes. The margin between each pair of classes is defined as the distance between two normalized support hyperplanes parallel to the corresponding discriminant hyperplane where any pattern is not contained.

However, as we point out in this paper, the margin is not always equal to the geometric margin which is defined as the minimal distance of patterns in a pair of classes to the

corresponding discriminant hyperplane classifying all patterns in both classes correctly, and thus, the geometric margin can exactly indicate the relation between each pattern and the discriminant function. Therefore, in this paper, we emphasize that maximizing the geometric margins is important for the generalization of multiclass classification, and propose a SVM model which maximizes all of the geometric margins of all pairs of classes. Moreover, since the multiclass classification can be essentially regarded as an optimization problem of maximizing all of the margins simultaneously, we formulate the proposed model as a multiobjective problem. However, since the multiobjective model is difficult to solve directly, we derive two kinds of single-objective optimization problems by using two scalarization approaches for multiobjective optimization, Benson's method and ε-constraint method, and transform them into single-objective second-order cone programming (SOCP) problems, respectively, which are solvable convex programming problems. Furthermore, we show theoretically that the proposed models can find Pareto optimal solutions of the multiobjective problem and apply them to some examples to demonstrate that the proposed models can achieve maximization of the geometric margins and to verify their generalization abilities.

This paper consists of six sections. In Section 2, we introduce the multiclass classification problem and the existing *all together* model. Next, in Section 3 we propose a new multiobjective SVM model, and in Section 4 we derive the proposed SOCP models based on the scalarization approaches. In Section 5, we verify the results shown in Sections 4 and 5 through numerical examples. Finally, we conclude in Section 6.

2 Multiclass Classification

2.1 All Together Model

In this paper, we consider the following multiclass classification problem: For given data: $D = \{x^i, y_i\}, i = 1, \ldots, m$, where $x^i \in \Re^n$ is an input pattern and $y_i \in K := \{1, \ldots, k\}$ denotes the corresponding class, we construct a classifier which divides all patterns into the corresponding classes:

$$f(x) = \arg\max_p \{w^{p\top} x + b^p\}.$$

where $w^p \in \Re^n$ and b^p, $p \in K$ are decision variables and the linear function $w^{p\top} x + b^p$ indicates the degree of confidence when a point x is classified into class p. Now, suppose that data D are piecewise linearly separable, which means that for all $q \neq p$, $p, q \in K$, there exists $w = (w^{1\top}, \ldots, w^{k\top})^\top$, $b = (b^1, \ldots, b^k)^\top$ such that

$$(w^p - w^q)^\top x^i + (b^p - b^q) > 0,\ i \in I^p, \tag{2.1}$$

where I^p denotes an index set defined by $I^p := \{i \in \{1, \ldots, m\} \mid y^i = p\}$. Here,

$$(w^p - w^q)^\top x + (b^p - b^q) = 0,\ q \neq p,\ p, q \in K, \tag{2.2}$$

is the discriminant hyperplane which distinguishes between classes p and q. Note that the representation of discriminant hyperplanes (2.2) is not unique. For any constants $t(\neq 0)$, $s \in \Re$ and any vector $v \in \Re^n$, $(w^{1\top}, \ldots, w^{k\top})$, $(b^1, \ldots, b^k)$ and $(tw^{1\top} + v^\top, \ldots, tw^{k\top} + v^\top)$, $(tb^1 + s, \ldots, tb^k + s)$ are different representations of the same discriminant function.

Furthermore, there exist an infinite number of discriminant functions to distinguish all classes correctly. In the binary classification, the discriminant hyperplane (2.2) is selected by maximizing $1/\|w^1 - w^2\|$ subject to $w^1 + w^2 = 0$, which is equivalent to minimization of $\|w^1\|^2$, that is, the standard binary SVM model. Then, it is guaranteed that the discriminant

hyperplane (2.2) maximizes the margin defined as the minimal distance of patterns in two classes 1 and 2 to the hyperplane, and thus the obtained discriminant function has the high generalization ability [13].

Therefore, in the multiclass classification, on the analogy of the binary SVM, the model maximizing $1/\|w^p - w^q\|$ for each pair $\{p, q\}$ of all classes was proposed [17].

$$\text{(O)} \quad \begin{array}{ll} \min\limits_{w,b} & f_{\mathrm{o}}(w) = \dfrac{1}{2}\displaystyle\sum_{p=1}^{k}\sum_{q=1,q\neq p}^{k}\|w^p - w^q\|^2 \\ \text{s.t} & (w^p - w^q)^\top x^i + (b^p - b^q) \geqq 1, \quad i \in I_p, \quad q \neq p, \quad p, q \in K. \end{array}$$

In addition, different models have been investigated from similar viewpoints, respectively [5, 8, 18], however, it is shown that these models are equivalent to the model (O) [8]. In this paper, we discuss only the model (O).

The model (O) can be interpreted as maximizing the margins called the *functional margins* in this paper:

$$d^{\mathrm{f}}_{pq}(w,b) := \left\{ \frac{1}{\|w^p - w^q\|} \middle| \begin{array}{l} (w^p - w^q)^\top x^i + (b^p - b^q) \geqq 1, \ x^i \in I_p, \\ (w^q - w^p)^\top x^i + (b^q - b^p) \geqq 1, \ x^i \in I_q \end{array} \right\}, \ q \neq p, \ p, q \in K.$$

Note that if D is piecewise linearly separable, (2.1) guarantees that $\|w^p - w^q\| > 0, \quad q \neq p, \ p, q \in K$ and thus the margins $d^{\mathrm{f}}_{pq}(w,b)$ are bounded. The functional margin denotes a half of the distance between the following two normalized support hyperplanes,

$$(w^p - w^q)^\top x + (b^p - b^q) = 1 \ \text{ and } \ (w^q - w^p)^\top x + (b^q - b^p) = 1, \tag{2.3}$$

where any pattern is not contained between the hyperplanes. However, the functional margin is not necessarily equal to the *geometric margin* defined as the distance of the nearest pattern in a pair of classes to the corresponding discriminant hyperplane classifying all patterns in both classes correctly.

$$d^{\mathrm{g}}_{pq}(w,b) := \min\left\{ \min_{i\in I_p} \frac{|(w^p - w^q)^\top x^i + (b^p - b^q)|}{\|w^p - w^q\|}, \ \min_{i\in I_q} \frac{|(w^p - w^q)^\top x^i + (b^p - b^q)|}{\|w^p - w^q\|} \right\}, \quad q > p, \ \ p, q \in K,$$

which denote a half of the distance between the following two support hyperplanes

$$\begin{aligned} (w^p - w^q)^\top x + (b^p - b^q) &= \sigma_{pq}(w,b), \\ (w^q - w^p)^\top x + (b^q - b^p) &= \sigma_{pq}(w,b), \end{aligned}$$

and $\sigma_{pq}(w,b)$ is defined by

$$\sigma_{pq}(w,b) := \min\left\{ \min_{i\in I_p} |(w^p - w^q)^\top x^i + (b^p - b^q)|, \ \min_{i\in I_q} |(w^q - w^p)^\top x^i + (b^q - b^p)| \right\}, \quad q > p, \ \ p, q \in K.$$

Note that the right-hand sides of these equalities are different from those in (2.3). Thus, although by minimizing $\|w^p - w^q\|, q \neq p \in K$ in the model (O), we can obtain the discriminant function having at least one pair (r, s) such that $d^{\mathrm{f}}_{rs}(w,b)$ is equal to the geometric margin $d^{\mathrm{g}}_{rs}(w,b)$, it cannot guarantee that all $d^{\mathrm{f}}_{pq}(w,b)$ are equal to the corresponding $d^{\mathrm{g}}_{pq}(w,b)$, which can be shown in the following theorem.

Theorem 2.1. *For any discriminant function* (2.1) *which correctly classifies a piecewise linearly separable data D,*

$$d^{\mathrm{g}}_{pq}(w,b) \geqq d^{\mathrm{f}}_{pq}(w,b), \quad q > p,\ p,q \in K,$$

holds. We have the equalities if and only if the following normalization condition holds:

$$\sigma_{pq}(w,b) = 1, \quad q > p, \ \ p,q \in K. \tag{2.4}$$

Proof. This result is easily verified by noticing the definitions of two kinds of margins. □

If $k = 2$, that is, in the binary classification, the condition (2.4) always holds for any discriminant function. On the other hand, in the multiclass classification there often exist discriminant functions having no representation satisfying (2.4), as we will show in the next subsection. Therefore, it often cannot guarantee that the model (O) achieves maximization of the geometric margins. Meanwhile, the geometric margin can exactly represent the distance of each class to the corresponding discriminant hyperplane in comparison with the functional margin.

Therefore, in this paper we emphasize that maximizing geometric margins is important for the generalization of multiclass classification and propose a new SVM which maximizes them. Moreover, since the multiclass classification means maximizing all margins simultaneously, it should be essentially regarded as a multiobjective optimization problem. Hence, we formulate the proposed model as a multiobjective problem.

2.2 Examples

Example 1:

Consider data $D^1 = \{x^1 = (0,1)^\top, y_1 = 1,\ x^2 = (1,0)^\top, y_2 = 1,\ x^3 = (2,0)^\top, y_3 = 2,\ x^4 = (0,2)^\top, y_4 = 3\ \}$. Then, the optimal solution of model (O) for D^1 is $w^1_{\mathrm{o}} = (-1,-1)^\top$, $w^2_{\mathrm{o}} = (1,0)^\top$, $w^3_{\mathrm{o}} = (0,1)^\top$ and $b_{\mathrm{o}} = (2,-1,-1)^\top$. The obtained discriminant hyperplanes are shown in Figure 1, where the dashed lines denote the obtained discriminant hyperplanes and the circle, square and triangle denote patterns with label 1, 2 and 3, respectively. Now,

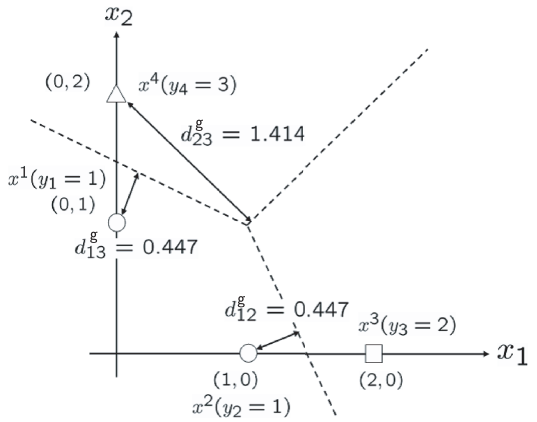


Figure 1: Model (O) for Example 1

the corresponding functional margins for the obtained discriminant function are

$$d^{\mathrm{f}}_{12}(w_{\mathrm{o}}, b_{\mathrm{o}}) = 0.447,\ d^{\mathrm{f}}_{13}(w_{\mathrm{o}}, b_{\mathrm{o}}) = 0.447,\ d^{\mathrm{f}}_{23}(w_{\mathrm{o}}, b_{\mathrm{o}}) = 0.707,$$

while its geometric margins are

$$d^{\mathrm{g}}_{12}(w_{\mathrm{o}}, b_{\mathrm{o}}) = 0.447,\ d^{\mathrm{g}}_{13}(w_{\mathrm{o}}, b_{\mathrm{o}}) = 0.447,\ d^{\mathrm{g}}_{23}(w_{\mathrm{o}}, b_{\mathrm{o}}) = 1.414.$$

Hence, we can see $d^{\mathrm{g}}_{23}(w_{\mathrm{o}}, b_{\mathrm{o}}) > d^{\mathrm{f}}_{23}(w_{\mathrm{o}}, b_{\mathrm{o}})$, which indicates that two kinds of margins are not equal. At the same time, we can observe that there exists no representation of these discriminant hyperplanes satisfying (2.4).

Next, we show the case that model (O) cannot maximize the geometric margins.

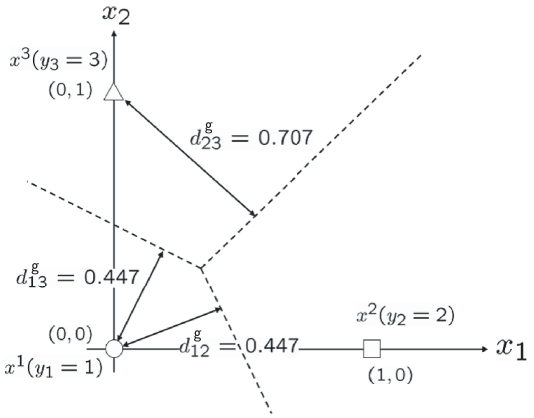


Figure 2: Model (O) for Example 2

Example 2:

Let us consider data $D^2 = \{x^1 = (0,0)^\top, y_1 = 1,\ x^2 = (1,0)^\top, y_2 = 2,\ x^3 = (0,1)^\top, y_3 = 3\}$. The optimal solution for model (O) for D^2 is $w^1_{\mathrm{o}} = (-1,-1)^\top$, $w^2_{\mathrm{o}} = (1,0)^\top$, $w^3_{\mathrm{o}} = (0,1)^\top$ and $b_{\mathrm{o}} = (2,-1,-1)^\top$. Figure 2 shows an obtained discriminant hyperplanes, where functional and geometric margins are given by

$$d^{\mathrm{f}}_{12}(w_{\mathrm{o}}, b_{\mathrm{o}}) = d^{\mathrm{g}}_{12}(w_{\mathrm{o}}, b_{\mathrm{o}}) = 0.447,\ d^{\mathrm{f}}_{13}(w_{\mathrm{o}}, b_{\mathrm{o}}) = d^{\mathrm{g}}_{13}(w_{\mathrm{o}}, b_{\mathrm{o}}) = 0.447,$$
$$d^{\mathrm{f}}_{23}(w_{\mathrm{o}}, b_{\mathrm{o}}) = d^{\mathrm{g}}_{23}(w_{\mathrm{o}}, b_{\mathrm{o}}) = 0.707.$$

In this case two kinds of margins are equal at $(w_{\mathrm{o}}, b_{\mathrm{o}})$. However, there exists another discriminant function with larger geometric margins given by $w^{1*} = \frac{2}{3}(-1\ \ -1)^\top$, $w^{2*} = \frac{2}{3}(2\ \ -1)^\top$, $w^{3*} = \frac{2}{3}(-1\ \ 2)^\top$ and $b^* = \frac{1}{3}(2,-1,-1)^\top$, as shown in Figure 3, where the functional and geometric margins are given by

$$d^{\mathrm{f}}_{12}(w^*, b^*) = 0.5,\ d^{\mathrm{f}}_{13}(w^*, b^*) = 0.5,\ d^{\mathrm{f}}_{23}(w^*, b^*) = 0.354,$$
$$d^{\mathrm{g}}_{12}(w^*, b^*) = 0.5,\ d^{\mathrm{g}}_{13}(w^*, b^*) = 0.5,\ d^{\mathrm{g}}_{23}(w^*, b^*) = 0.707.$$

This fact shows that two kinds of margins are not equal at (w^*, b^*) and that $(w_{\mathrm{o}}, b_{\mathrm{o}})$ is

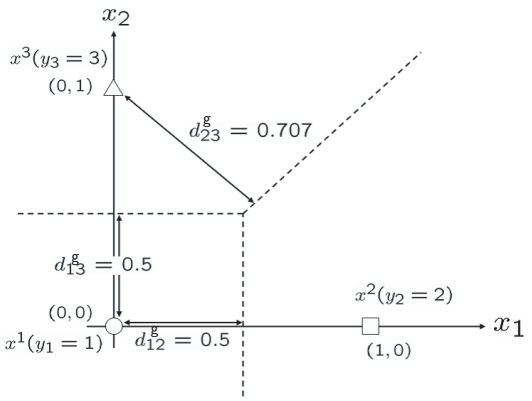


Figure 3: The best discriminant hyperplanes for Example 2

not optimal in the sense of maximization of the sum of geometric margins, which can be also verified from $f_{\rm o}(w_{\rm o}) = 12 < 16 = f_{\rm o}(w^*)$. Furthermore, note that the solution (w^*, b^*) is complete optimal in the sense of Pareto optimality for maximizing geometric margins and thus it dominates the solution $(w_{\rm o}, b_{\rm o})$, which implies that the multiclass classification problem should be formulated as a multiobjective optimization problem.

Therefore, in the next section, we propose a new model which can maximize the geometric margins in terms of the multiobjective optimization.

3 Multiobjective Model Maximizing Geometric Margins

In this and the following sections, we shall use the following notations for the orders of x, $y \in \Re^n$:

$$\begin{aligned} x \leqq y, \quad & \text{if } x_i \leqq y_i,\ i = 1, \ldots, n, \\ x \leq y, \quad & \text{if } x_i \leqq y_i,\ i = 1, \ldots, n, \text{ and } x \neq y, \\ x < y, \quad & \text{if } x_i < y_i,\ i = 1, \ldots, n. \end{aligned}$$

First, as we mentioned in the previous section, we formulate the multiclass classification problem as the following multiobjective optimization problem which maximizes multiple geometric margins.

$$\text{(M1)} \quad \begin{array}{ll} \max\limits_{w,b} & d(w, b) \\ \text{s.t.} & (w^p - w^q)^\top x^i + (b^p - b^q) \geqq 1, \quad i \in I_p, \quad q \neq p, \quad p, q \in K, \end{array}$$

where $d(w, b)$ is defined by

$$d(w, b) = \left(d^{\rm g}_{12}(w, b), d^{\rm g}_{13}(w, b), \ldots, d^{\rm g}_{(k-1)k}(w, b) \right)^\top.$$

The model (M1) maximizes the geometric margins of all pairs of classes subject to correct classification for all patterns. Although this formulation is natural, it is difficult to solve

it directly because of its complexity. Thus, we propose the following model (M2) using a vector $\sigma \in \Re^{k(k-1)/2}$ and a function $\theta(w, \sigma)$:

$$
\text{(M2)} \quad \begin{array}{ll} \max\limits_{w,b,\sigma} & \theta(w,\sigma) \\ \text{s.t.} & (w^p - w^q)^\top x^i + (b^p - b^q) \geqq \sigma_{pq}, \ \ i \in I_p, \ \ q > p, \ \ p, q \in K, \\ & (w^q - w^p)^\top x^i + (b^q - b^p) \geqq \sigma_{pq}, \ \ i \in I_q, \ \ q > p, \ \ p, q \in K, \\ & \sigma_{pq} \geqq 1, \ \ q > p, \ \ p, q \in K, \end{array}
$$

where the objective function is defined by

$$
\theta(w, \sigma) = \left(\theta_{12}(w, \sigma), \theta_{13}(w, \sigma), \ldots, \theta_{(k-1)k}(w, \sigma)\right)^\top,
$$

and

$$
\theta_{pq}(w, \sigma) = \frac{\sigma_{pq}}{\|w^p - w^q\|}, \ \ q > p, \ \ p, q \in K.
$$

In addition, for convenience, $F(\text{M1})$ and $F(\text{M2})$ represent the feasible regions of (M1) and (M2), respectively. Notice that if data D is piecewise linearly separable, $F(\text{M1})$ and $F(\text{M2})$ are not empty, respectively, and thus, $\|w^p - w^q\| > 0, \ q > p, \ p, q \in K$, as mentioned in the previous section.

Now, let us consider the relation between two models (M1) and (M2). We begin to discuss the boundedness of objective function $d(w, b)$ of (M1). Thus, we define the minimal distance between classes p and q by

$$
d^{\mathrm{m}}_{pq} := \min\{\|x^i - x^j\| \mid i \in I_p, \ j \in I_q\}, \ q > p, \ p, q \in K.
$$

Then, it is easily shown that $d^{\mathrm{m}}_{pq}/2$ is the upper bound of $d^{\mathrm{g}}_{pq}(w, b)$ for feasible solutions (w, b) of (M1) for any $q > p$, $p, q \in K$.

Next, we show some lemmas to see the boundedness of $\theta(w, \sigma)$ for feasible solutions (w, b, σ) of (M2).

Lemma 3.1. *If (w, b, σ) is feasible for* (M2)*, then (w, b) is feasible for* (M1)*.*

Proof. Since $(w, b, \sigma) \in F(\text{M2})$, we have

$$
(w^p - w^q)^\top x^i + (b^p - b^q) \geqq \sigma_{pq} \geqq 1, \ i \in I_p, \ q > p, \ p, q \in K,
$$

and

$$
(w^q - w^p)^\top x^i + (b^q - b^p) \geqq \sigma_{pq} \geqq 1, \ i \in I_p, \ q > p, \ p, q \in K.
$$

Thus, (w, b) is feasible for (M1). □

Now, we define $\sigma(w, b)$ by using $\sigma_{pq}(w, b)$ as

$$
\sigma(w, b) := \left(\sigma_{12}(w, b), \sigma_{13}(w, b), \ldots, \sigma_{(k-1)k}(w, b)\right)^\top.
$$

Lemma 3.2. *If (w, b) is feasible for* (M1)*, then $(w, b, \sigma(w, b))$ is feasible for* (M2) *and we have $d(w, b) = \theta(w, \sigma(w, b))$.*

Proof. Since (w, b) is feasible for (M1), (w, b) satisfies

$$
(w^p - w^q)^\top x^i + (b^p - b^q) \geqq 1, \ i \in I_p, \ q \neq p, p, q \in K.
$$

Thus, from the definition of $\sigma(w,b)$ we have that for any $x_i, i \in I^p,\ q > p,\ p,q \in K$,

$$\begin{aligned}
&(w^p - w^q)^\top x^i + (b^p - b^q) \\
&\geqq \min\left\{\min_{i\in I^p}\left|(w^p - w^q)^\top x^i + (b^p - b^q)\right|,\ \min_{i\in I^q}\left|(w^q - w^p)^\top x^i + (b^q - b^p)\right|\right\} \\
&= \sigma_{pq}(w,b) \geqq 1, \quad i \in I_p,
\end{aligned}$$

and similarly we have

$$(w^q - w^p)^\top x^i + (b^q - b^p) \geqq \sigma_{pq}(w,b) \geqq 1, \quad i \in I_q,\ q > p,\ p,q \in K.$$

Therefore, $(w,b,\sigma(w,b))$ is feasible for (M2). Moreover, we have for any $q > p$, $p,q \in K$

$$\begin{aligned}
d^{\mathrm{g}}_{pq}(w,b) &= \min\left\{\min_{i\in I_p}\frac{|(w^p - w^q)^\top x^i + (b^p - b^q)|}{\|w^p - w^q\|},\ \min_{i\in I_q}\frac{|(w^q - w^p)^\top x^i + (b^q - b^p)|}{\|w^q - w^p\|}\right\} \\
&= \frac{\sigma_{pq}(w,b)}{\|w^p - w^q\|} = \theta_{pq}(w,\sigma(w,b)).
\end{aligned}$$

□

Lemma 3.3. *If (w,b,σ) is feasible for* (M2)*, then $(w,b,\sigma(w,b))$ is also feasible for* (M2) *and $\theta(w,\sigma(w,b)) \geqq \theta(w,\sigma)$.*

Proof. If $(w,b,\sigma) \in F$(M2), then $(w,b) \in F$(M1) from Lemma 3.1. In addition, from Lemma 3.2 $(w,b,\sigma(w,b))$ is feasible for F(M2), and from constraints of (M2) we can derive

$$\begin{aligned}
\min_{i\in I_p}\{|(w^p - w^q)^\top x^i + (b^p - b^q)|\} &\geqq \sigma^{pq},\ q > p \in K, \\
\min_{i\in I_q}\{|(w^q - w^p)^\top x^i + (b^q - b^p)|\} &\geqq \sigma^{pq},\ q > p \in K,
\end{aligned}$$

which, together with the definition of $\sigma_{pq}(w,b)$, yields that

$$\frac{\sigma_{pq}(w,b)}{\|w^p - w^q\|} \geqq \frac{\sigma_{pq}}{\|w^p - w^q\|},\ q > p,\ p,q \in K.$$

Therefore, we have $\theta(w,\sigma(w,b)) \geqq \theta(w,\sigma)$. □

By applying these lemmas, the boundedness of $\theta(w,\sigma)$ for feasible solutions (w,b,σ) of (M2) can be shown as follows.

Lemma 3.4. *A set $\{\theta(w,b)\ |(w,b,\sigma) \in F$(M2)$\}$ is bounded.*

Proof. From Lemmas 3.1–3.3, we have that for any feasible solution (w,b,σ) of (M2), $(w,b) \in F$(M1) and $\theta(w,\sigma) \leqq \theta(w,\sigma(w,b)) = d(w,b)$. Here, since $d^{\mathrm{g}}_{pq}(w,b)$ for all feasible solutions of (M1) is bounded above by the constant $d^m_{pq}/2$ for any $q > p$, $p,q \in K$ and $\theta(w,b) \geqq 0$, we can see that $\{\theta(w,b)|\ (w,b,\sigma) \in F$(M2)$\}$ is also bounded. □

Next, we present the conditions which guarantee the existence of Pareto optimal solutions of (M2). Hence, let us define $F_{\bar{w},\bar{b},\bar{\sigma}}$(M2) for a feasible solution $(\bar{w},\bar{b},\bar{\sigma})$ of (M2) by

$$F_{\bar{w},\bar{b},\bar{\sigma}}(\mathrm{M2}) := \{(w,b,\sigma) \in F(\mathrm{M2}) \mid \theta(\bar{w},\bar{\sigma}) \leqq \theta(w,\sigma)\},$$

and consider the following assumption.

Assumption 3.5. For any feasible solution $(\bar{w}, \bar{b}, \bar{\sigma})$ of (M2), a set $\{\theta(w,\sigma) \mid (w,b,\sigma) \in F_{\bar{w},\bar{b},\bar{\sigma}}(\mathrm{M2})\}$ is closed.

Then, we can show the following theorem.

Theorem 3.6. *Suppose that Assumption 3.5 holds. Then there exist Pareto optimal solutions of* (M2).

Proof. Lemma 3.4 and Assumption 3.5 yield that $\{\theta(w,b) |\ (w,b,\sigma) \in F_{\bar{w},\bar{b},\bar{\sigma}}(\mathrm{M2})\}$ is closed and bounded, which guarantees the existence of Pareto optimal solutions of (M2) [7]. □

In practice, we can expect that the assumption holds for almost all classification problems. Hence, throughout this and the following sections we suppose that Assumption 3.5 holds.

Finally, we show that the optimal solutions of (M2) can be considered to be equivalent to those of (M1).

Lemma 3.7. *If* (w^*, b^*, σ^*) *is Pareto optimal for* (M2)*, then we have* $\theta(w^*, \sigma^*) = \theta(w^*, \sigma(w^*, b^*))$.

Proof. Since $(w^*, b^*, \sigma^*) \in F(\mathrm{M2})$, we have $\theta(w^*, \sigma^*) \leqq \theta(w^*, \sigma(w^*, b^*))$ from Lemma 3.3. Moreover, if $\theta(w^*, \sigma^*) \leq \theta(w^*, \sigma(w^*, b^*))$ holds, then it contradicts the Pareto optimality of (w^*, b^*, σ^*) . Therefore, we have $\theta(w^*, \sigma^*) = \theta(w^*, \sigma(w^*, b^*))$. □

Theorem 3.8. *If* (w^*, b^*, σ^*) *is Pareto optimal for* (M2)*,* (w^*, b^*) *is Pareto optimal for* (M1)*. Conversely, if* (w^*, b^*) *is Pareto optimal for* (M1)*,* $(w^*, b^*, \sigma(w^*, b^*))$ *is Pareto optimal for* (M2).

Proof. Frist, we show that (w^*, b^*) is Pareto optimal for (M1) if (w^*, b^*, σ^*) is Pareto optimal for (M2). We assume that (w^*, b^*) is not Pareto optimal for (M1). In view of Lemma 3.1, (w^*, b^*) is feasible for (M1) and hence there exists a feasible solution (w, b) for (M1) such that

$$d(w^*, b^*) \leq d(w, b). \tag{3.1}$$

Then, Lemma 3.2 leads to

$$d(w^*, b^*) = \theta(w^*, \sigma(w^*, b^*)). \tag{3.2}$$

From the Pareto optimality of (w^*, b^*, σ^*) for (M2) and Lemma 3.7, we have

$$\theta(w^*, \sigma^*) = \theta(w^*, \sigma(w^*, b^*)). \tag{3.3}$$

By using the feasibility of (w, b) and Lemma 3.2,

$$d(w, b) = \theta(w, \sigma(w, b)). \tag{3.4}$$

Then, from (3.1), (3.2), (3.3) and (3.4) we can derive

$$\theta(w, \sigma(w, b)) = d(w, b) \geq d(w^*, b^*) = \theta(w^*, \sigma(w^*, b^*)) = \theta(w^*, \sigma^*).$$

This fact contradicts the Pareto optimality of (w^*, b^*, σ^*). Therefore, we conclude that (w^*, b^*) is Pareto optimal for (M1).

Secondly, we show that $(w^*, b^*, \sigma(w^*, b^*))$ is Pareto optimal for (M2) if (w^*, b^*) is Pareto optimal for (M1). We assume that $(w^*, b^*, \sigma(w^*, b^*))$ is not Pareto optimal for (M2). From

the feasibility of (w^*, b^*) for (M1) and Lemma 3.2, $(w^*, b^*, \sigma(w^*, b^*))$ is feasible for (M2). Then, there exists a feasible solution (w, b, σ) for (M2) such that

$$\theta(w^*, \sigma(w^*, \sigma^*)) \leq \theta(w, \sigma), \tag{3.5}$$

and

$$d(w^*, b^*) = \theta(w^*, \sigma(w^*, b^*)). \tag{3.6}$$

Since (w, b) is feasible for (M1), by Lemma 3.2 we have

$$d(w, b) = \theta(w, \sigma(w, b)). \tag{3.7}$$

Moreover, by Lemma 3.3 we have

$$\theta(w, \sigma) \leqq \theta(w, \sigma(w, b)). \tag{3.8}$$

Thus, from (3.5), (3.6), (3.7) and (3.8), we can derive

$$d(w^*, b^*) = \theta(w^*, \sigma(w^*, b^*)) \leq \theta(w, \sigma) \leqq \theta(w, \sigma(w, b)) = d(w, b).$$

This fact contradicts the Pareto optimality of (w^*, b^*). Therefore, we conclude that $(w^*, b^*, \sigma(w^*, b^*))$ is Pareto optimal for (M2). □

Theorem 3.9. *If (w^*, b^*, σ^*) is weakly Pareto optimal for* (M2)*, (w^*, b^*) is weakly Pareto optimal for* (M1)*. Conversely, if (w^*, b^*) is weakly Pareto optimal for* (M1)*, $(w^*, b^*, \sigma(w^*, b^*))$ is weakly Pareto optimal for* (M2)*.*

Proof. This theorem can be easily shown similarly to Theorem 3.8. □

Theorems 3.6 and 3.8 show the existence of Pareto optimal solutions of (M1) and (M2). In addition, Theorems 3.8 and 3.9 imply that we can solve (M2) instead of (M1). However, since (M2) is multiobjective, in the next section we derive two kinds of single-objective optimization problems by scalarization approaches to multiobjective optimization, Benson's method and ε-constraint method, and furthermore transform them into solvable problems.

4 Single-objective Model

4.1 SOCP Model Based on Benson's Method

In this subsection, we first consider the following single-objective problem which is derived from a scalarization approach to multiobjective optimization called Benson's method.

$$\text{(Pmax-sum)} \quad \begin{array}{ll} \max\limits_{w,b,\sigma,l} & \sum\limits_{q\in K}\sum\limits_{q>p\in K} l_{pq} \\ \text{s.t.} & l_{pq} \geqq 0, \ \ q > p, \ \ p, q \in K, \\ & \dfrac{\sigma_{pq}}{\|w^p - w^q\|} - \dfrac{\bar{\sigma}_{pq}}{\|\bar{w}^p - \bar{w}^q\|} = l_{pq}, \ \ q > p, \ \ p, q \in K, \\ & (w^p - w^q)^\top x^i + (b^p - b^q) \geqq \sigma_{pq}, \ \ i \in I_p, \ \ q > p, \ \ p, q \in K, \\ & (w^q - w^p)^\top x^i + (b^q - b^p) \geqq \sigma_{pq}, \ \ i \in I_q, \ \ q > p, \ \ p, q \in K, \\ & \sigma_{pq} \geqq 1, \ \ q > p, \ \ p, q \in K. \end{array}$$

This method improves some initial feasible solution $(\bar{w}, \bar{b}, \bar{\sigma})$, by maximizing the sum of nonnegative deviation variables $l_{pq} = \theta_{pq}(w, b, \sigma) - \theta_{pq}(\bar{w}, \bar{b}, \bar{\sigma})$, $q > p$, $p, q \in K$. It is known that an optimal solution of (Pmax-sum) is Pareto optimal for (M2) [7].

However, (Pmax-sum) is still difficult to solve because of its fractional constraints. Thus, secondly, we consider the following solvable problem in which some constraints of (Pmax-sum), $\frac{\sigma_{pq}}{\|w^p - w^q\|} - \frac{\bar{\sigma}_{pq}}{\|\bar{w}^p - \bar{w}^q\|} = l_{pq}$ and $\sigma_{pq} \geqq 1$, are replaced with $\sigma_{pq} - \frac{\bar{\sigma}_{pq}}{\|\bar{w}^p - \bar{w}^q\|}\|w^p - w^q\| \geqq l_{pq}$ and $1 \leqq \sigma_{pq} \leqq c_{pq}$ by using constants c_{pq}, respectively.

$$
\begin{array}{ll}
\max\limits_{w,b,\sigma,l} & \sum\limits_{q\in K}\sum\limits_{q>p\in K} l_{pq} \\
\text{s.t.} & l_{pq} \geqq 0, \ \ q > p, \ \ p, q \in K, \\
\text{(P2max-sum)} & \sigma_{pq} - \|w^p - w^q\| \dfrac{\bar{\sigma}_{pq}}{\|\bar{w}^p - \bar{w}^q\|} \geqq l_{pq}, \ \ q > p, \ \ p, q \in K, \\
& (w^p - w^q)^\top x^i + (b^p - b^q) \geqq \sigma_{pq}, \ \ i \in I_p, \ \ q > p, \ \ p, q \in K, \\
& (w^q - w^p)^\top x^i + (b^q - b^p) \geqq \sigma_{pq}, \ \ i \in I_q, \ \ q > p, \ \ p, q \in K, \\
& 1 \leqq \sigma_{pq} \leqq c_{pq}, \ \ q > p, \ \ p, q \in K.
\end{array}
$$

This model can be regarded as a second-order cone programming problem (SOCP), as shown in the following transformation. By defining

$$
\begin{array}{rcl}
v^{pq} & := & w^p - w^q, \ \ q > p, \ \ p, q \in K, \\
\rho_{pq} & := & \sigma_{pq} - 1, \ \ q > p, \ \ p, q \in K, \\
m_{pq} & := & \dfrac{\|\bar{w}^p - \bar{w}^q\|}{\bar{\sigma}_{pq}}(\sigma_{pq} - l_{pq}), \ \ q > p, \ \ p, q \in K, \\
\xi_{pqi} & := & (w^p - w^q)^\top x^i + (b^p - b^q) - \sigma_{pq}, \ \ i \in I_p, \ \ q > p, \ \ p, q \in K, \\
\xi_{qpi} & := & (w^q - w^p)^\top x^i + (b^q - b^p) - \sigma_{pq}, \ \ i \in I_q, \ \ q > p, \ \ p, q \in K,
\end{array}
$$

(P2max-sum) can be transformed into the following second-order cone programming problem (SOCP1):

$$
\begin{array}{ll}
\max\limits_{v,m,b,\rho,l,\xi,\eta} & \sum\limits_{q\in K}\sum\limits_{q>p\in K} l_{pq} \\
\text{s.t.} & m_{pq} \geqq \|v^{pq}\|, \ \ q > p, \ \ p, q \in K, \\
& m_{pq} - \dfrac{\|\bar{w}^p - \bar{w}^q\|}{\bar{\sigma}_{pq}}(1 + \rho_{pq} - l_{pq}) = 0, \ \ q > p, \ \ p, q \in K, \\
& (v^{pq})^\top x^i + (b^p - b^q) - \rho_{pq} - \xi_{pqi} = 1, \ \ i \in I_p, \\
& \qquad\qquad q > p, \ \ p, q \in K, \\
\text{(SOCP1)} & (-v^{pq})^\top x^i + (b^q - b^p) - \rho_{pq} - \xi_{qpi} = 1, \ \ i \in I_q, \\
& \qquad\qquad q > p, \ \ p, q \in K, \\
& c_{pq} - \rho_{pq} - \eta_{pq} = 1, \ \ q > p, \ \ p, q \in K, \\
& v^{pq} = v^{p\kappa} + v^{\kappa q}, \kappa \in K \setminus \{p, q\}, \ q > p, \ p, q \in K, \\
& \xi_{pqi} \geqq 0, \ \ i \in I^p, \ \ p \neq q, \ \ p, q \in K, \\
& l_{pq} \geqq 0, \ \rho_{pq} \geqq 0, \ \eta_{pq} \geqq 0, \ \ q > p, \ \ p, q \in K, \\
& b_p \geqq 0, \ \ p \in K.
\end{array}
$$

The SOCP is a convex programming problem having a linear objective function and linear and second-order cone constraints, which can be efficiently solved by a number of methods such as the primal-dual interior point method within the almost same time as a quadratic programming problem of the same size [2]. Moreover, several commercial and noncommercial solvers have been developed [3, 12].

Next, let us consider what kind of solution we can obtain by solving (P2max-sum). Here, by focusing on the constraints of (P2max-sum) and (M2), we can easily confirm that for any feasible solution (w, b, σ, l) of (P2max-sum), (w, b, σ) is always feasible for (M2), while

for any feasible solution (w, b, σ) of (M2), σ does not necessarily satisfy the constraints of (P2max-sum), $1 \leqq \sigma_{pq} \leqq c_{pq},\ q > p,\ \ p, q \in K$. Thus, let us consider the relation between feasible solutions of (M2) and (P2max-sum).

Now, we define $t(\sigma)$ by

$$t(\sigma) := \max\{1/\sigma_{pq} \mid q > p,\ p, q \in K\}.$$

Then, for any feasible solution (w, b, σ) for (M2), $t(\sigma)$ is the minimal $t > 0$ such that $(tw, tb, t\sigma) \in F(\mathrm{M2})$ and $\theta(t(\sigma)w, t(\sigma)\sigma) = \theta(w, \sigma)$. Next, we define $c^M_{\bar{w},\bar{b},\bar{\sigma}}$ by using $F_{\bar{w},\bar{b},\bar{\sigma}}(\mathrm{M2})$

$$c^M_{\bar{w},\bar{b},\bar{\sigma}} := \sup\{t(\sigma)\sigma_{pq} \mid q > p,\ p, q \in K, (w, b, \sigma) \in F_{\bar{w},\bar{b},\bar{\sigma}}(\mathrm{M2})\}.$$

Then, the relation between two feasible solutions is shown in the following lemma.

Lemma 4.1. *If parameters c_{pq} in* (P2max-sum) *satisfy $c_{pq} \geqq c^M_{\bar{w},\bar{b},\bar{\sigma}}$ for any $q > p,\ p, q \in K$, then for any solution $(w, b, \sigma) \in F_{\bar{w},\bar{b},\bar{\sigma}}(\mathrm{M2})$, $(t(\sigma)w, t(\sigma)b, t(\sigma)\sigma, l(w, \sigma))$ is feasible for* (P2max-sum)*, where $l(w, \sigma)$ is defined by*

$$l_{pq}(w, \sigma) := t(\sigma)\sigma_{pq} - \|t(\sigma)(w^p - w^q)\| \frac{\bar{\sigma}_{pq}}{\|\bar{w}^p - \bar{w}^p\|},\ q > p,\ p, q \in K,$$

and we have $\theta(w, \sigma) = \theta(t(\sigma)w, t(\sigma)\sigma)$.

Proof. Since $(w, b, \sigma) \in F_{\bar{w},\bar{b},\bar{\sigma}}(M2)$, we have $\theta(w, \sigma) \geqq \theta(\bar{w}, \bar{\sigma})$. Thus,

$$\begin{aligned} l_{pq}(w, \sigma) &= \|t(\sigma)(w^p - w^q)\| \left(\frac{\sigma_{pq}}{\|(w^p - w^q)\|} - \frac{\bar{\sigma}_{pq}}{\|\bar{w}^p - \bar{w}^p\|} \right) \\ &= \|t(\sigma)(w^p - w^q)\| \left(\theta_{pq}(w, \sigma) - \theta_{pq}(\bar{w}, \bar{\sigma})\right) \geqq 0,\ q > p,\ p, q \in K. \end{aligned}$$

From the definitions of $t(\sigma)$ and $c^M_{\bar{w},\bar{b},\bar{\sigma}}$, and the assumption of the lemma, we have $1 \leqq t(\sigma)\sigma_{pq} \leqq c^M_{\bar{w},\bar{b},\bar{\sigma}} \leqq c_{pq}$, for any $q > p,\ p, q \in K$. In addition, since $(t(\sigma)w, t(\sigma)b, t(\sigma)\sigma, l(w, \sigma))$ satisfies other constraints of (P2max-sum), it is feasible for (P2max-sum). Moreover, from the definition of $\theta(w, b)$ we have $\theta(w, b) = \theta(t(\sigma)w, t(\sigma)b)$. □

From this lemma we can see that $c^M_{\bar{w},\bar{b},\bar{\sigma}} < \infty$ is required and c_{pq} should be selected appropriately. Hence, let us consider the following assumption.

Assumption 4.2. For any feasible $(\bar{w}, \bar{b}, \bar{\sigma}) \in F(\mathrm{M2})$, $c^M_{\bar{w},\bar{b},\bar{\sigma}} < \infty$ and c_{pq} in (P2max-sum) satisfy $c_{pq} \geqq c^M_{\bar{w},\bar{b},\bar{\sigma}}$ for any $q > p,\ p, q \in K$.

In general, it can be expected that $c^M_{\bar{w},\bar{b},\bar{\sigma}} < \infty$ for any feasible $(\bar{w}, \bar{b}, \bar{\sigma}) \in F(\mathrm{M2})$ in all classification problems because any feasible solution (w, b, σ) is constrained to classify all patterns correctly. Throughout this subsection, we suppose that Assumption 4.2 is satisfied.

Theorem 4.3. *If the optimal value of* (P2max-sum) *is* 0 *and its optimal solution is $(w^*, b^*, \sigma^*, l^*)$, then (w^*, b^*, σ^*) is Pareto optimal for* (M2) *and $\theta(w^*, \sigma^*) = \theta(\bar{w}, \bar{\sigma})$. Conversely, if $(\bar{w}, \bar{b}, \bar{\sigma})$ is Pareto optimal for* (M2)*, then the optimal value of* (P2max-sum) *is* 0*, and $(t(\bar{\sigma})\bar{w}, t(\bar{\sigma})\bar{b}, t(\bar{\sigma})\bar{\sigma}, 0)$ is optimal for* (P2max-sum)*.*

Proof. First, we show that if the optimal value of (P2max-sum) is 0 and its optimal solution is $(w^*, b^*, \sigma^*, l^*)$, then (w^*, b^*, σ^*) is Pareto optimal for (M2). Thus, assume that (w^*, b^*, σ^*)

is not Pareto optimal for (M2). Then, since (w^*, b^*, σ^*) is feasible for (M2), there exists a feasible solution $(\hat{w}, \hat{b}, \hat{\sigma})$ of (M2) such that $\theta(w^*, \sigma^*) \leq \theta(\hat{w}, \hat{\sigma})$. Now, let us define

$$\hat{l}_{pq} := t(\hat{\sigma})\hat{\sigma}_{pq} - \|t(\hat{\sigma})(\hat{w}^p - \hat{w}^q)\| \frac{\bar{\sigma}_{pq}}{\|\bar{w}^p - \bar{w}^p\|},\ q > p,\ p, q \in K.$$

Then, we have $\sum_{q\in K} \sum_{q>p\in K} \hat{l}_{pq} > 0$. In addition, $(t(\hat{\sigma})\hat{w}, t(\hat{\sigma})\hat{b}, t(\hat{\sigma})\hat{\sigma}, \hat{l})$ is feasible for (P2max-sum) from Lemma 4.1. These facts contradict that the optimal value of (P2max-sum) is 0. Therefore, (w^*, b^*, σ^*) is Pareto optimal for (M2). Moreover, since $\sigma^*_{pq} - \frac{\bar{\sigma}_{pq}}{\|\bar{w}^p - \bar{w}^q\|}\|w^{p*} - w^{q*}\| = 0,\ p < q,\ p, q \in K$, we have $\theta(\bar{w}, \bar{\sigma}) = \theta(w^*, \sigma^*)$.

Next, we show that if $(\bar{w}, \bar{b}, \bar{\sigma})$ is Pareto optimal for (M2), then the optimal value of (P2max-sum) is 0. Assume that $(w^*, b^*, \sigma^*, l^*)$ is an optimal solution for (P2max-sum) and $\sum_{q\in K} \sum_{q>p\in K} l^*_{pq} > 0$. Then, for some $s > r \in K$

$$\sigma^*_{rs} - \|w^{r*} - w^{s*}\| \frac{\bar{\sigma}_{rs}}{\|\bar{w}^r - \bar{w}^s\|} \geqq l^*_{rs} > 0,$$

and for any $q > p,\ p, q \in K$

$$\sigma^*_{pq} - \|w^{p*} - w^{q*}\| \frac{\bar{\sigma}_{pq}}{\|\bar{w}^p - \bar{w}^q\|} \geqq l^*_{pq} \geqq 0,$$

These inequalities yield that $\theta(\bar{w}, \bar{\sigma}) \leq \theta(w^*, \sigma^*)$ and (w^*, b^*, σ^*) is feasible for (M2). Thus, the facts contradict the Pareto optimality of $(\bar{w}, \bar{b}, \bar{\sigma})$. Therefore, the optimal value $\sum_{q\in K} \sum_{q>p\in K} l^*_{pq}$ of (P2max-sum) is 0. Moreover, we define $\bar{l}_{pq} := t(\bar{\sigma})\bar{\sigma}_{pq} - \|t(\bar{\sigma})(\bar{w}^p - \bar{w}^q)\| \frac{\bar{\sigma}_{pq}}{\|\bar{w}^p - \bar{w}^p\|},\ q > p,\ p, q \in K$. Then, from $\bar{l} = 0$ and Lemma 4.1, $(t(\bar{\sigma})\bar{w},\ t(\bar{\sigma})\bar{b},\ t(\bar{\sigma})\bar{\sigma},\ 0)$ is optimal for (P2max-sum). □

Theorem 4.4. *Let* $(w^*, b^*, \sigma^*, l^*)$ *be an optimal solution of* (P2max-sum). *If its optimal value* $\sum_{q\in K} \sum_{q>p\in K} l^*_{pq}$ *is greater than* 0, *then* $\theta(\bar{w}, \bar{\sigma}) \leq \theta(w^*, \sigma^*)$.

Proof. From the assumption of theorem, we have $l^* \geq 0$. Since we have $\sigma^*_{pq} - \frac{\bar{\sigma}}{\|\bar{w}^p - \bar{w}^q\|}\|w^{p*} - w^{q*}\| = l^*_{pq},\ q > p,\ p, q \in K$ from optimality of (w^*, b^*, σ^*), we can derive the result of the theorem. □

Theorems 4.3 and 4.4 imply that if an obtained optimal value for (P2max-sum) is 0, the obtained solution is Pareto optimal for (M2), and otherwise, the obtained solution (w^*, b^*, σ^*) dominates the initial solution $(\bar{w}, \bar{b}, \bar{\sigma})$. Furthermore, we propose the following iterative method of solving (M2) by exploiting these properties of (P2max-sum).

Iterative method based on Benson's method: IMB

Step 0. Set $\tau := 0$ and $(w^{(0)}, b^{(0)}, \sigma^{(0)}) = (\bar{w}, \bar{b}, \bar{\sigma})$.

Step 1. Solve (P2max-sum) using $(w^{(\tau)}, b^{(\tau)}, \sigma^{(\tau)})$ as an initial solution and $c^{(\tau)}_{pq} > 0,\ q > p \in K$, and obtain the optimal solution $(w^*, b^*, \sigma^*, l^*)$.

Step 2. Set $(w^{(\tau+1)}, b^{(\tau+1)}, \sigma^{(\tau+1)}, l^{(\tau+1)}) := (w^*, b^*, \sigma^*, l^*)$.

If $\sum_{q\in K} \sum_{q>p\in K} l^{(\tau+1)}_{pq} \leqq \delta$, then terminate. Otherwise, $\tau := \tau + 1$ and go to **Step 1**.

Here, δ is a sufficiently small positive constant.

If $\sum_{q\in K}\sum_{q>p\in K} l_{pq}^{(\tau+1)} = 0$ holds at some τ, then IMB obtains a Pareto optimal solution for (M2). Otherwise, the method may generate an infinite sequence $\{w^{(\tau)}, b^{(\tau)}, \sigma^{(\tau)}\}$, $\tau = 0, \ldots$. Thus, let us consider the case where the condition $\sum_{q\in K}\sum_{q>p\in K} l_{pq}^{(\tau)} = 0$ does not hold at any τ.

Theorem 4.5. *Assume that* $c_{pq}^{(\tau)}$ *in* IMB *satisfies that* $c_{pq}^{(\tau)} \geqq c^{M}_{w^{(\tau)},b^{(\tau)},\sigma^{(\tau)}}$, $q > p \in K$ *for any* $\tau \geqq 0$. *If* $\delta = 0$ *and* $l^{(\tau)} \geq 0$ *for any* τ *in* IMB, *then a sequence* $\{\theta(w^{(\tau)}, \sigma^{(\tau)})\}$ *generated by* IMB *converges to a point* $\theta(\hat{w}, \hat{\sigma})$ *such that* $(\hat{w}, \hat{b}, \hat{\sigma}) \in F_{\bar{w},\bar{b},\bar{\sigma}}$(M2) *is Pareto optimal for* (M2). *In addition,* $\left\{\sum_{q\in K}\sum_{q>p\in K} l_{pq}^{(\tau)}\right\}$ *converges to 0.*

Proof. First, we show the convergence of the sequences $\{\theta(w^{(\tau)}, \sigma^{(\tau)})\}$ and $\left\{\sum_{q\in K}\sum_{q>p\in K} l_{pq}^{(\tau)}\right\}$. Since (P2max-sum) solved at iteration τ in IMB uses $(w^{(\tau)}, b^{(\tau)}, \sigma^{(\tau)})$ as an initial solution and the obtained optimal solutions is given by $(w^{(\tau+1)}, b^{(\tau+1)}, \sigma^{(\tau+1)})$, we have $\theta(w^{(\tau)}, \sigma^{(\tau)}) \leq \theta(w^{(\tau+1)}, \sigma^{(\tau+1)})$ from Theorem 4.4. In addition, the sequence $\{\theta(w^{(\tau)}, \sigma^{(\tau)})\}$ is monotone nondecreasing and included in $\{\theta(w, \sigma) \mid (w, b, \sigma) \in F_{\bar{w},\bar{b},\bar{\sigma}}(\text{M2})\}$, which is bounded and closed from Lemma 3.4 and Assumption 3.5. Therefore, $\{\theta(w^{(\tau)}, \sigma^{(\tau)})\}$ converges to a point $\theta(\hat{w}, \hat{\sigma})$ such that $(\hat{w}, \hat{b}, \hat{\sigma}) \in F_{\bar{w},\bar{b},\bar{\sigma}}$(M2).

Furthermore, since $\theta(\bar{w}, \bar{\sigma}) = \theta(w^{(0)}, \sigma^{(0)}) \leq \theta(w^{(\tau)}, \sigma^{(\tau)})$ and $\sigma_{pq}^{(\tau)} \leqq c_{pq}^{(\tau)}$ from the feasibility of $(w^{(\tau)}, b^{(\tau)}, \sigma^{(\tau)}, l^{(\tau)})$ for (P2max-sum), we have

$$\|w^{p(\tau)} - w^{q(\tau)}\| \leqq \frac{\sigma_{pq}^{(\tau)}}{\theta_{pq}(\bar{w}, \bar{\sigma})} \leqq \frac{\max_{\tau} c_{pq}^{(\tau)}}{\theta_{pq}(\bar{w}, \bar{\sigma})},\ q > p,\ p, q \in K.$$

Thus, $\|w^{p(\tau)} - w^{q(\tau)}\|$ is bounded from above. At the same time, we have

$$\begin{aligned} l_{pq}^{(\tau)} &= \sigma^{(\tau)} - \|w^{p(\tau)} - w^{q(\tau)}\| \frac{\sigma_{pq}^{(\tau-1)}}{\|w^{p(\tau-1)} - w^{q(\tau-1)}\|} \\ &= \|w^{p(\tau)} - w^{q(\tau)}\| \left(\theta_{pq}(w^{(\tau)}, \theta^{(\tau)}) - \theta_{pq}(w^{(\tau-1)}, \theta^{(\tau-1)})\right), \\ &\quad q < p,\ p, q \in K,\ \tau = 1, \ldots, \end{aligned}$$

which, together with the upper boundedness of $\|w^{p(\tau)} - w^{q(\tau)}\|$ and the convergence of $\{\theta(w^{(\tau)}, \sigma^{(\tau)})\}$, yields that $\sum_{q\in K}\sum_{p<q\in K} l_{pq}^{(\tau)} \to 0$ as $\tau \to \infty$.

Next, we show $(\hat{w}, \hat{b}, \hat{\sigma})$ is Pareto optimal for (M2). Assume that $(\hat{w}, \hat{b}, \hat{\sigma})$ is not Pareto optimal. Then, let us consider the problem (P2max-sum) using $(\hat{w}, \hat{b}, \hat{\sigma})$ as an initial solution and suppose that (w, b, σ, l) is its optimal solution. Then, we have $\sum_{q\in K}\sum_{p<q\in K} l_{pq} > 0$, $l \geq 0$ and

$$\sigma_{pq} - \|w^p - w^q\| \frac{\hat{\sigma}_{pq}}{\|\hat{w}^p - \hat{w}^q\|} = l_{pq},\ q > p,\ p, q \in K.$$

Moreover, from Theorem 4.4 we have $\hat{\sigma}_{pq}/\|\hat{w}^p - \hat{w}^q\| \geqq \sigma_{pq}^{(\tau)}/\|w^{p(\tau)} - w^{q(\tau)}\|$ for any $\tau \geq 0$ and any $q > p \in K$. Thus, we have

$$\sigma_{pq} - \|w^p - w^q\| \frac{\sigma_{pq}^{(\tau)}}{\|w^{p(\tau)} - w^{q(\tau)}\|} \geqq l_{pq},\ \tau \geq 0,\ q > p,\ p, q \in K,$$

which means (w, b, σ, l) satisfies the second constraints of (P2sum-max) using $(w^{(\tau)}, b^{(\tau)}, \sigma^{(\tau)})$ as an initial solution. In addition, since (w, b, σ, l) satisfies other constraints, it is feasible. Furthermore, since $\sum_{q\in K}\sum_{p<q\in K} l_{pq}^{(\tau+1)} \to 0$ as $\tau \to \infty$, we have $\sum_{q\in K}\sum_{p<q\in K} l_{pq}^{(\tau+1)} < \sum_{q\in K}\sum_{p<q\in K} l_{pq}$ for a sufficiently large τ. However, the result contradicts the fact that $(w^{(\tau+1)}, b^{(\tau+1)}, \sigma^{(\tau+1)})$ is optimal for (P2sum-max) using $(w^{(\tau)}, b^{(\tau)}, \sigma^{(\tau)})$ as an initial solution. Therefore, $(\hat{w}, \hat{b}, \hat{\sigma})$ is Pareto optimal for (M2). □

Theorem 4.5 implies that if the constant δ is small positive, then IMB terminates within a finite number of iterations. Additionally if $\sum_{q\in K}\sum_{q>p\in K} l_{pq}^{(\tau+1)} = 0$, the obtained solution is Pareto optimal for (M2). Otherwise, the obtained solution is approximately Pareto optimal. Moreover, since $\theta(w^{(\tau)}, b^{(\tau)}, \sigma^{(\tau)}) \leq \theta(w^{(\tau+1)}, b^{(\tau+1)}, \sigma^{(\tau+1)})$, we have $c^M_{w^{(\tau)}, b^{(\tau)}, \sigma^{(\tau)}} \geqq c^M_{w^{(\tau+1)}, b^{(\tau+1)}, \sigma^{(\tau+1)}}$ from the definition of $c^M_{w,b,\sigma}$ for any $\tau \geqq 0$. Thus, we can use the same constant vector $\bar{c}$ such that $\bar{c} \geqq c^M_{\bar{w}, \bar{b}, \bar{\sigma}}$ as $c^{(\tau)}$ at each iteration τ.

In this subsection, we have shown that the proposed method IMB can obtain a Pareto optimal solution. In order to obtain various Pareto optimal solutions, we can extend (P2max-sum) by replacing the objective function with $\sum_{q\in K}\sum_{q>p\in K} \omega_{pq} l_{pq}$, where ω_{pq} is a positive weight for each l_{pq}, $q > p \in K$. In the next subsection, we discuss another scalarization approach which is more suitable to finding many kinds of Pareto optimal solutions.

4.2 SOCP Model Based on ε-constraint Method

Here, we propose another model based on the ε-constraint method. By applying the ε-constraint approach to (M2), the following problem can be derived:

$$
(\varepsilon\text{-P}) \quad \begin{array}{ll} \max\limits_{w,b,\sigma} & g_1(w,\sigma) = \dfrac{\sigma_{rs}}{\|w^r - w^s\|} \\ \text{s.t.} & \dfrac{\sigma_{pq}}{\|w^p - w^q\|} \geqq \varepsilon_{pq}, \quad q > p, \quad (p,q) \neq (r,s), \quad p, q \in K, \\ & (w^p - w^q)^\top x^i + (b^p - b^q) \geqq \sigma_{pq}, \quad i \in I_p, \quad q > p, \quad p, q \in K, \\ & (w^q - w^p)^\top x^i + (b^q - b^p) \geqq \sigma_{pq}, \quad i \in I_q, \quad q > p, \quad p, q \in K, \\ & \sigma_{pq} \geqq 1, \quad q > p, \quad p, q \in K, \end{array}
$$

where a pair (r, s) and constants ε_{pq}, $q > p$, $(p, q) \neq (r, s)$, $p, q \in K$ are appropriately selected such that the feasible reagin of (ε-P) is not empty. This method maximizes only one of the objectives of (M2) while the others are transformed to constraints with ε_{pq}. Then, the following theorems are known about ε-constraint method.

Theorem 4.6 ([7]). *Let (w, b, σ) be an optimal solution of (ε-P) for some (r, s). Then (w, b, σ) is weakly Pareto optimal for (M2).*

Theorem 4.7 ([7]). *(w, b, σ) is Pareto optimal for (M2) if and only if there exists an ε_{-rs} such that (w, b, σ) is optimal for (ε-P) for any $(r, s) \subset K$.*

Here ε_{-rs} denotes a vector in which the element ε_{rs} is removed from ε. These theorems show that we can obtain any Pareto optimal solution of (M2) by solving (ε-P) with an appropriate choice of ε_{-rs}.

However, (ε-P) is also difficult to solve because of its fractional constraints and objective functions. Hence, by making use of one degree of freedom of (ε-P), we add a constraint

$\sigma_{rs} = c_{rs}$ with an appropriate constant $c_{rs} \geqq 1$ to obtain the following model:

$$
(\varepsilon\text{-P2}) \quad
\begin{array}{ll}
\max\limits_{w,b,\sigma_{-rs}} & g_2(w) = \dfrac{c_{rs}}{\|w^r - w^s\|} \\
\text{s.t.} & \dfrac{\sigma_{pq}}{\|w^p - w^q\|} \geqq \varepsilon_{pq}, \ \ q > p, \ \ (p,q) \neq (r,s), \ \ p,q \in K, \\
& (w^r - w^s)^\top x^i + (b^r - b^s) \geqq c_{rs}, \ \ i \in I_r, \\
& (w^s - w^r)^\top x^i + (b^s - b^r) \geqq c_{rs}, \ \ i \in I_s, \\
& (w^p - w^q)^\top x^i + (b^p - b^q) \geqq \sigma_{pq}, \ \ i \in I_p, \ \ q > p, \\
& \qquad\qquad (p,q) \neq (r,s), \ \ p,q \in K, \\
& (w^q - w^p)^\top x^i + (b^q - b^p) \geqq \sigma_{pq}, \ \ i \in I_q, \ \ q > p, \\
& \qquad\qquad (p,q) \neq (r,s), \ \ p,q \in K, \\
& \sigma_{pq} \geqq 1, \ \ q > p, \ \ (p,q) \neq (r,s), \ \ p,q \in K,
\end{array}
$$

where (w, b, σ_{-rs}) denotes the vector in which the element σ_{rs} is removed from (w, b, σ). Moreover, for a solution (w, b, σ_{-rs}) of (ε-P2), we define a vector $(w, b, (\sigma_{-rs}, c_{rs}))$ whose element σ_{rs} is c_{rs} and other elements are equal to (w, b, σ_{-rs}).

Then, similarly to (P2max-sum), we can show that (ε-P2) is equivalent to the SOCP. Now, we define

$$
\begin{array}{rcl}
v^{pq} & := & w^p - w^q, \ \ q > p, \ \ p,q \in K \\
\rho_{pq} & := & \sigma_{pq} - 1, \ \ q > p, \ \ p,q \in K \\
\xi_{pqi} & := & (w^p - w^q)^\top x^i + (b^p - b^q) - \sigma_{pq}, \ \ i \in I_p, \ \ q > p, \ \ p,q \in K, \\
\xi_{qpi} & := & (w^q - w^p)^\top x^i + (b^q - b^p) - \sigma_{pq}, \ \ i \in I_q, \ \ q > p, \ \ p,q \in K.
\end{array}
$$

By using these variables, (ε-P2) can be transformed to the following problem:

$$
(\text{SOCP2}) \quad
\begin{array}{ll}
\min\limits_{v,b,\sigma_{-rs},l,\rho_{-rs},\xi} & l \\
\text{s.t.} & l \geqq \|v^{rs}\|, \\
& \sigma_{pq} \geqq \varepsilon_{pq}\|v^{pq}\|, \ \ (p,q) \neq (r,s), \ \ q > p, \ \ p,q \in K, \\
& \sigma_{pq} - \rho_{pq} = 1, \ \ (p,q) \neq (r,s), \ \ q > p, \ \ p,q \in K, \\
& (v^{pq})^\top x^i + (b^p - b^q) - \sigma_{pq} - \xi_{pqi} = 0, \ \ i \in I_p, \\
& \qquad\qquad (p,q) \neq (r,s), \ \ q > p, \ \ p,q \in K, \\
& (-v^{pq})^\top x^i + (b^q - b^p) - \sigma_{pq} - \xi_{qpi} = 0, \ \ i \in I_q, \\
& \qquad\qquad (p,q) \neq (r,s), \ \ q > p, \ \ p,q \in K, \\
& (v^{rs})^\top x^i + (b^r - b^s) - \xi_{rsi} = c_{rs}, \ \ i \in I_r, \\
& (-v^{sr})^\top x^i + (b^s - b^r) - \xi_{sri} = c_{rs}, \ \ i \in I_s, \\
& v^{pq} = v^{p\kappa} + v^{\kappa q}, \kappa \in K \setminus \{p,q\}, \ q > p, \ p,q \in K, \\
& \xi_{pqi} \geqq 0, \ \ i \in I^p, \ \ p \neq q, \ \ p,q \in K, \\
& \rho_{pq} \geqq 0, \ \ (p,q) \neq (r,s), \ \ q > p, \ \ p,q \in K, \\
& b_p \geqq 0, \ \ p \in K.
\end{array}
$$

Therefore, (ε-P2) can be effectively solved similarly to (P2max-sum).

Next, let us consider the properties of optimal solutions of (ε-P2). Here, we define c_ε^M by using $t(\sigma)$.

$$
c_\varepsilon^M := \sup\{t(\sigma)\sigma_{pq} \mid q > p, \ p,q \in K, (w,b,\sigma) \in \Omega(\varepsilon\text{-P})\},
$$

where $\Omega(\varepsilon$-P) denotes the set of all optimal solutions of (ε-P). Moreover, we consider the following assumption:

Assumption 4.8. For any ε_{-rs} such that $\Omega(\varepsilon\text{-P}) \neq \emptyset$, $c_\varepsilon^M < \infty$ and c_{rs} in (ε-P2) satisfy $c_{rs} \geqq c_\varepsilon^M$.

Note that this assumption is similar to Assumption 4.2 mentioned in the previous subsection. Since we can also expect that $c_\varepsilon^M < \infty$ for any classification problem, we suppose that this assumption holds throughout this subsection. Then we can show the following theorems.

Theorem 4.9. *Let (w^*, b^*, σ^*) be an optimal solution of (ε-P), then $\frac{c_{rs}}{\sigma^*_{rs}}(w^*, b^*, \sigma^*_{-rs})$ is optimal for (ε-P2).*

Proof. First, we show that $\frac{c_{rs}}{\sigma^*_{rs}}(w^*, b^*, \sigma^*_{-rs})$ is feasible for (ε-P2). From the feasibility of (w^*, b^*, σ^*) for (ε-P), Assumption 4.8 and the definition of c_ε^M, we have $t(\sigma^*)\sigma^*_{rs} \leqq c_\varepsilon^M \leqq c_{rs}$. From the definition of $t(\sigma)$, we have

$$t(\sigma^*)\sigma^*_{pq} = \max_{q>p\in K}\left\{\frac{1}{\sigma^*_{pq}}\right\}\sigma^*_{pq} \geqq 1,\ q > p,\ p, q \in K.$$

Thus,

$$\frac{c_{rs}}{\sigma^*_{rs}}\sigma^*_{pq} = \frac{c_{rs}}{t(\sigma^*)\sigma^*_{rs}}\ t(\sigma^*)\sigma^*_{pq} \geqq 1,\ (p, q) \neq (r, s),\ q > p,\ p, q \in K.$$

Moreover, the feasibility of (w^*, b^*, σ^*) for (ε-P) yields that

$$\begin{aligned}
&\frac{c_{rs}}{\sigma^*_{rs}}(w^{*r} - w^{*s})^\top x^i + \frac{c_{rs}}{\sigma^*_{rs}}(b^{*r} - b^{*s}) \geqq \frac{c_{rs}}{\sigma^*_{rs}}\ \sigma^*_{rs} = c_{rs},\ \ i \in I_r,\\
&\frac{c_{rs}}{\sigma^*_{rs}}(w^{*s} - w^{*r})^\top x^i + \frac{c_{rs}}{\sigma^*_{rs}}(b^{*s} - b^{*r}) \geqq \frac{c_{rs}}{\sigma^*_{rs}}\ \sigma^*_{rs} = c_{rs},\ \ i \in I_s.
\end{aligned}$$

In addition, it is easily shown that other constraints of (ε-P2) are satisfied by $\frac{c_{rs}}{\sigma^*_{rs}}(w^*, b^*, \sigma^*_{-rs})$. Therefore, $\frac{c_{rs}}{\sigma^*_{rs}}(w^*, b^*, \sigma^*_{-rs})$ is feasible for (ε-P2).

Next, we show that $\frac{c_{rs}}{\sigma^*_{rs}}(w^*, b^*, \sigma^*_{-rs})$ is optimal for (ε-P2). It is easily confirmed that for any feasible solution (w, b, σ_{-rs}) of (ε-P2), $(w, b, (\sigma_{-rs}, c_{rs}))$ is feasible for (ε-P) and

$$g_1(w, (\sigma_{-rs}, c_{rs})) = \frac{c_{rs}}{\|w^r - w^s\|} = g_2(w). \tag{4.1}$$

In addition, since (w^*, b^*, σ^*) is optimal for (ε-P), we have

$$\frac{\sigma^*_{rs}}{\|w^{*r} - w^{*s}\|} = g_1(w^*, \sigma^*) \geqq g_1(w, (\sigma_{-rs}, c_{rs})). \tag{4.2}$$

At the same time, we have

$$g_2\left(\frac{c_{rs}}{\sigma^*_{rs}}w^*\right) = \frac{\sigma^*_{rs}}{\|w^{*r} - w^{*s}\|}, \tag{4.3}$$

Therefore, from (4.1), (4.2) and (4.3) we can derive that

$$g_2\left(\frac{c_{rs}}{\sigma^*_{rs}}w^*\right) \geqq g_2(w),$$

for any feasible solution (w, b, σ_{-rs}) of (ε-P2). Therefore, $\frac{c_{rs}}{\sigma^*_{rs}}(w^*, b^*, \sigma^*)$ is optimal for (ε-P2). □

Theorem 4.10. *Let* $(w^*, b^*, \sigma^*_{-rs})$ *be an optimal solution of* (ε-P2). *Then* $(w^*, b^*, (\sigma^*_{-rs}, c_{rs}))$ *is optimal for* (ε-P).

Proof. It is obvious that if $(w^*, b^*, \sigma^*_{-rs})$ is feasible for (ε-P2), $(w^*, b^*, (\sigma^*_{-rs}, c_{rs}))$ is feasible for (ε-P). Now, let us suppose that $(w^*, b^*, (\sigma^*_{-rs}, c_{rs}))$ is not optimal for (ε-P). Then, there exists an optimal solution $(\hat{w}, \hat{b}, \hat{\sigma})$ such that $g_1(\hat{w}, \hat{\sigma}) > g_1(w^*, (\sigma^*_{-rs}, c_{rs}))$, which means

$$\frac{\hat{\sigma}_{rs}}{\|\hat{w}^r - \hat{w}^s\|} > \frac{c_{rs}}{\|w^{*r} - w^{*s}\|}. \tag{4.4}$$

Now, since $\frac{c_{rs}}{\hat{\sigma}_{rs}}(\hat{w}, \hat{b}, \hat{\sigma}_{-rs})$ is optimal for (ε-P2) from Theorem 4.9 and $(w^*, b^*, \sigma^*_{-rs})$ is also optimal for (ε-P2), we have

$$\frac{\hat{\sigma}_{rs}}{\|\hat{w}^r - \hat{w}^s\|} = g_2\left(\frac{c_{rs}}{\hat{\sigma}_{rs}}\hat{w}\right) = g_2(w^*) = \frac{c_{rs}}{\|w^{*r} - w^{*s}\|}. \tag{4.5}$$

Thus, (4.4) and (4.5) contradict. Therefore, $(w^*, b^*, (\sigma^*{}_{-rs}, c_{rs}))$ is optimal for (ε-P). □

Theorem 4.10 shows that for an optimal solution $(w^*, b^*, \sigma^*_{-rs})$ of (ε-P2), $(w^*, b^*, (\sigma^*_{-rs}, c_{rs}))$ is optimal for (ε-P). Thus, Theorem 4.6 implies that the optimal solution is weakly Pareto optimal for (M2). In addition, the result, together with Theorems 4.7 and 4.9, suggests that we can obtain any (weakly) Pareto optimal solution of (M2) by solving (ε-P2) with an appropriate choice of ε_{-rs}. Consequently, we can conclude that various discriminant functions maximizing the geometric margins can be obtained by solving (ε-P2) as a pair (r, s) and the corresponding parameter ε_{-rs} are varied.

Finally, in the next section, we apply the proposed models to some examples as mentioned in Section 3 and other examples.

5 Numerical Examples

In this section, we report the results of numerical experiments, where we compared the proposed models based on Benson's and ε-constraint methods with the existing model (O). We used the optimization tools in MathWorks Matlab 7.0.1 * and Mosek version 5.0 † to solve the QP and SOCP problems.

We applied the existing and proposed models to examples mentioned in Section 2, an example having no complete optimal solution, and real-world data sets.

5.1 Examples 1 and 2

For Examples 1 and 2, we classified D^1 and D^2 by using IMB and (ε-P2). For both examples, we set $(r, s) = (1, 2)$ and $c_{12} = 10$ in (ε-P2). Parameters ε_{13} and ε_{23} were set as $\varepsilon_{-12} = \theta_{-12}(w_{\rm o}, \sigma_{\rm o})$, where $(w_{\rm o}, b_{\rm o}, \sigma_{\rm o})$ is the solution obtained by the existing model (O) for D^1 and D^2, respectively, as shown in Section 2.2. In addition, IMB used $\delta = 10^{-6}$, $c_{pq} = 10$, $q > p,\ p, q \in \{1, 2, 3\}$ and $(w_{\rm o}, b_{\rm o}, \sigma_{\rm o})$ as an initial solution $(\bar{w}, \bar{b}, \bar{\sigma})$ for both examples.

For Example 1, (ε-P2) obtained a solution: $w^1 = (-6.6552, -6.6804)^\top$, $w^2 = (13.3448, -6.6806)^\top$, $w^3 = (-6.6896, 13.3610)^\top$, $b = (20.0024, -9.9976, -10.0048)^\top$, $\sigma_{13} = 9.4524$, $\sigma_{23} = 40.0757$. The corresponding geometric margins for the solution were given by

$$d^{\rm g}_{12}(w, b) = 0.5000,\ d^{\rm g}_{13}(w, b) = 0.4973,\ d^{\rm g}_{23}(w, b) = 1.4142.$$

*http://www.mathworks.com/

†http://www.mosek.com/

Besides, for Examples 1, IMB obtained the following solution in four iterations. $w^1 = (-0.6673,\ -0.6673)^\top$, $w^2 = (1.3345, -0.6673)^\top$, $w^3 = (-0.6673, 1.3345)^\top$, $b = (2.0018, -1.0009, -1.0009)^\top$, $\sigma = (1.0009, 1.0009, 4.0036)^\top$. The corresponding geometric margins of the solution were

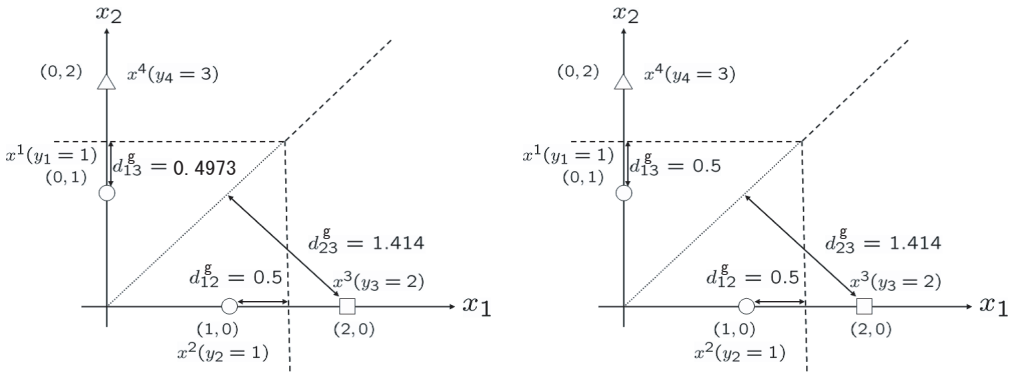
$$d_{12}^{\mathrm{g}}(w,b) = 0.5000,\ d_{13}^{\mathrm{g}}(w,b) = 0.5000,\ d_{23}^{\mathrm{g}}(w,b) = 1.4142.$$

Table 1 shows the geometric margins obtained at each iteration in IMB, which indicates that margins obtained finally were achieved at iteration 2.

Table 1: Obtained geometric margins at each iteration in IMB

Iteration	d_{12}^{g}	d_{13}^{g}	d_{23}^{g}
1	0.4900	0.4960	1.4142
2	0.5000	0.5000	1.4142
3	0.5000	0.5000	1.4142
4	0.5000	0.5000	1.4142

The discriminant hyperplanes obtained in two models are shown in Figures 4(a) and 4(b), where the dashed lines denote the discriminant hyperplanes and the circle, square and triangle denote patterns with label 1, 2 and 3, respectively, similarly to Figures 1-3. We can observe that the geometric margins and the corresponding discriminant hyperplanes for the solution obtained by both methods are almost same. Moreover note that the obtained solution is the complete optimal solution of (M1), which dominates the geometric margins obtained by solving model (O) (cf. Figure 1).



(a) (ε-P2) (b) IMB

Figure 4: Proposed models for Example 1

For Example 2, (ε-P2) obtained $w^1 = (-6.6603, -6.6746)^\top$, $w^2 = (13.3397, -6.6742)^\top$, $w^3 = (-6.6794, 13.3488)^\top$, $b = (6.6673, -3.3327, -3.3347)^\top$, $\sigma_{13} = 9.3116$, $\sigma_{23} = 20.0209$. The corresponding geometric margins for the solution are

$$d_{12}^{\mathrm{g}}(w,b) = 0.5000,\ d_{13}^{\mathrm{g}}(w,b) = 0.4995,\ d_{23}^{\mathrm{g}}(w,b) = 0.7071.$$

Besides, IMB obtained the following solution in five iterations, $w^1 = (-0.6667,\ -0.6668)^\top$, $w^2 = (1.3333, -0.6665)^\top$, $w^3 = (-0.6667, 1.3333)^\top$, $b = (0.6667,\ -0.3333,\ -0.3334)^\top$,

$\sigma = (1.0000, 1.0001, 2.0000)^\top$. The corresponding geometric margins for the solution are given by

$$d^{\mathrm{g}}_{12}(w,b) = 0.5000,\ d^{\mathrm{g}}_{13}(w,b) = 0.4995,\ d^{\mathrm{g}}_{23}(w,b) = 0.7071.$$

We can see that the geometric margins and the corresponding discriminant hyperplanes for the solution obtained by both methods are same, additionally, which are the same ones shown in the Figure 3. Note that the obtained solution is also the complete optimal solution of (M1).

These results show that the proposed model can obtain a Pareto optimal solution of (M1) which is better than the solution obtained by solving (O) in the sense of maximizing the geometric margins.

5.2 Example 3

Next, we consider Example 3 given by $D^3 = \{x^1 = (0,1)^\top, y_1 = 1,\ x^2 = (2,0)^\top,\ y_2 = 1,\ x^3 = (3,0)^\top,\ y_3 = 2,\ x^4 = (0,3)^\top, y_4 = 3\}$, which has also only three points but has no complete optimal solution. Thus, the problem has many Pareto optimal solutions.

We applied the existing model (O) and proposed models to this problem. In both proposed models, similarly to the previous subsection, we set $c_{pq} = 10$, $q > p$, $p, q \in \{1,2,3\}$ and used the solution obtained by the model (O) as an initial solution $(\bar{w}, \bar{b}, \bar{\sigma})$ in IMB, while we executed (ε-P) with all combinations of pairs of classes as the fixed pair (r,s), that is, $(r,s) = (1,2)$, $(1,3)$, $(2,3)$, and set $\varepsilon_{-rs} = \theta_{-rs}(\bar{w}, \bar{b}, \bar{\sigma})$.

Obtained results are shown in Figures 5–6 and Table 2, where (ε-P2)$_{(1,2)}$ denotes the model (ε-P2) with $(r,s) = (1,2)$. The table shows that the geometric margins for the discriminant hyperplanes obtained by the model (O) are smaller than those obtained by any proposed model. In particular, the margin d^{g}_{23} obtained by the model (O) is considerably small, while there is no large difference of obtained margins among proposed models.

Table 2: Results of each methods for Example 3

Model	Parameters	d^{g}_{12}	d^{g}_{13}	d^{g}_{23}
(O)	–	0.4851	0.8944	0.9487
IMB	$c = 10$	0.4919	0.9996	2.1181
(ε-P2)$_{(1,2)}$	$c_{12} = 10$	0.5000	0.9940	1.7645
(ε-P2)$_{(1,3)}$	$c_{13} = 10$	0.4991	1.0000	1.8990
(ε-P2)$_{(2,3)}$	$c_{23} = 10$	0.4980	0.9856	2.1213

Furthermore, we applied (ε-P2) to Example 3 so as to obtain various solutions, where we set $(r,s) = (2,3)$ and $c_{23} = 10$, and varied $(\varepsilon_{12}, \varepsilon_{13})$ in $[0.01, 0.5] \times [0.01, 1.0]$. Figure 7 (a) indicates that many kinds of weakly Pareto optimal solutions were obtained by (ε-P2). Figure 7 (b) shows the magnified region of the Pareto curve near to the point $(0.5, 1.0, 2.12)$, which is a set of Pareto optimal solutions. We can observe that many Pareto optimal solutions were obtained. Therefore, it can be concluded that (ε-P2) can obtain many kinds of Pareto optimal solutions by appropriate choices of ε_{-rs}.

5.3 Real-world Data Sets

Finally, in this subsection, we applied models (O), IMB and (ε-P2) to two real-world data sets from the UCI machine learning repository [16], Wine and DNA. For model (ε-P2)

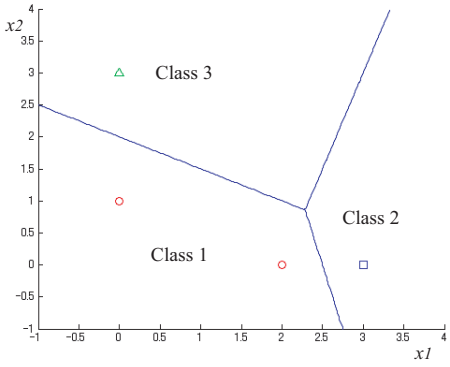
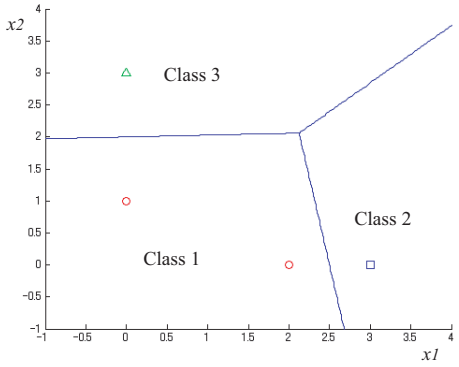
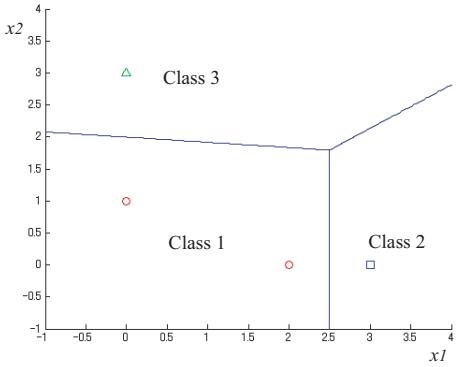


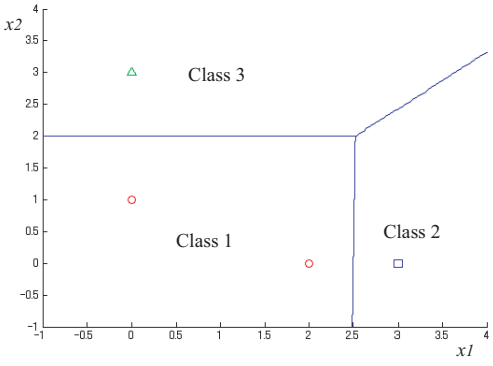
Figure 5: Model (O) for Example 3



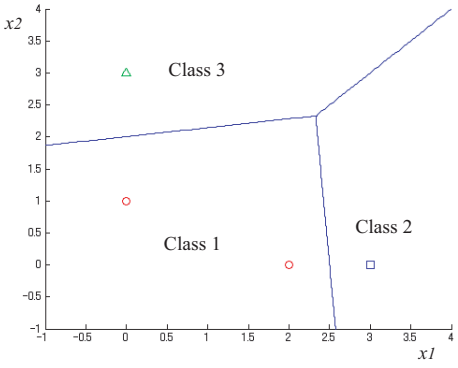
(a) IMB



(b) $(\varepsilon\text{-P2})_{(1,2)}$

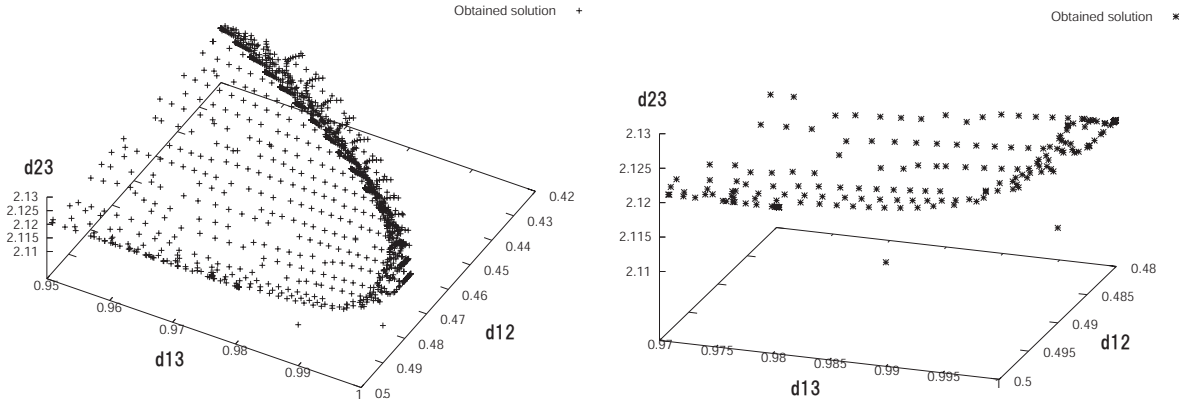


(c) $(\varepsilon\text{-P2})_{(1,3)}$



(c) $(\varepsilon\text{-P2})_{(2,3)}$

Figure 6: Proposed models, IMB and (ε-P2), for Example 3



(a) Overall view of obtained solutions (b) Pareto curve of obtained solutions

Figure 7: Proposed model (ε-P2) for Example 3

we set $(r, s) = (1, 2), (1, 3)$ or $(2, 3)$ and $c_{rs} = 10$, and determined constants ε_{-rs} by $\varepsilon_{-rs} = \theta_{-rs}(w_{\rm o}, \sigma(w_{\rm o}, b_{\rm o}))$, while for model IMB we used $(c_{12}, c_{13}, c_{23}) = (100, 100, 100)$ and $(w_{\rm o}, b_{\rm o}, \sigma(w_{\rm o}, b_{\rm o}))$ as an initial solution $(\bar{w}, \bar{b}, \bar{\sigma})$, where $(w_{\rm o}, b_{\rm o})$ was a solution obtained by the model (O) for each problem. In addition, we executed IMB using a solution obtained by model (ε-P2) as an initial solution, which is called the hybrid model. If a solution obtained by (ε-P2) is weakly Pareto optimal, this model can improve the solution to obtain a Pareto otpimal solution. However, we could not obtain the satisfactory solution by IMB for DNA data, where IMB required solving a large number of problems (P2max-sum) iteratively, and the optimal solutions of some problems were not properly obtained. That is attributed to the fact the problem (P2max-sum) gradually becomes ill-conditioned because DNA data has a large number of instances and attributes. Meanwhile, the difficulty was not observed in model (ε-PS2) because a weakly Pareto optimal solution is obtained by solving a single problem.

Tables 3 shows classification rates and objective function values of solutions obtained by (O), IMB, (ε-P2) and the hybrid model for Wine, and those obtained by (O) and (ε-P2) for DNA, where "# of Ite." denotes the number of iterations in which problems (P2max-sum) were solved by IMB and the hybrid model, and hybrid$_{(1,2)}$ denotes the hybrid model, that is, IMB using a solution obtained by (ε-PS2)$_{(1,2)}$ as an initial solution. We can observe that all solutions obtained by the proposed three models dominate ones by the existing model and that test classification rates of proposed models are better than or equal to those of the existing model. On the other hand, although Theorem 4.6 guarantees that a solution obtained by the hybrid model dominate the corresponding initial solution, which is obtained by (ε-PS2), the results of Table 3 are slightly inconsistent with it. They can be considered to be due to the numerical instability of IMB. However, the numerical error is small enough to neglect in practical use, and the hybrid model boosts the test classification rate of (ε-PS2).

In addition, we evaluated the 10-fold cross-validation estimate of four models for Wine, and two models for DNA to compare the generalization abilities of them as shown in Tables 4 and 5, respectively, where the figure in parenthesis denotes the average number of iterations required in IMB. The tables indicate that (ε-P2) and hybrid model are superior to (O) and IMB in the sense of the generalization, and that in the hybrid model, IMB improved the solution obtained by (ε-P2) for Wine. Therefore, we can conclude that there exist

Table 3: Comparison of results obtained by four models

Data	Model	Classification rate (training)	Classification rate (test)	Geometric margins d^g_{12}	Geometric margins d^g_{13}	Geometric margins d^g_{23}	# of Ite.
Wine	(O)	100.00	88.89	0.3812	0.4858	0.4368	-
	IMB	100.00	88.89	0.3812	0.8451	0.4368	2
	$(\varepsilon\text{-P2})_{(1,2)}$	100.00	88.89	0.3875	0.5130	0.4479	-
	$(\varepsilon\text{-P2})_{(1,3)}$	100.00	88.89	0.3811	0.8452	0.4368	-
	$(\varepsilon\text{-P2})_{(2,3)}$	100.00	94.44	0.3816	0.4889	0.5127	-
	$\text{hybrid}_{(1,2)}$	100.00	94.44	0.3857	0.7155	0.4478	2
	$\text{hybrid}_{(1,3)}$	100.00	88.89	0.3811	0.8455	0.4368	2
	$\text{hybrid}_{(2,3)}$	100.00	94.44	0.3799	0.5186	0.5119	10
DNA	(O)	100.00	93.00	0.1148	0.1241	0.1194	-
	$(\varepsilon\text{-P2})_{(1,2)}$	100.00	93.00	0.1178	0.1241	0.1194	-
	$(\varepsilon\text{-P2})_{(1,3)}$	100.00	93.50	0.1148	0.1263	0.1194	-
	$(\varepsilon\text{-P2})_{(2,3)}$	100.00	93.00	0.1148	0.1241	0.1213	-

Table 4: 10-fold Cross-validation results of four models for Wine

(O)	$(\varepsilon\text{-PS2})_{(1,2)}$	$(\varepsilon\text{-PS2})_{(1,3)}$	$(\varepsilon\text{-PS2})_{(2,3)}$
95.51	96.07	95.51	96.63
IMB	$\text{hybrid}_{(1,2)}$	$\text{hybrid}_{(1,3)}$	$\text{hybrid}_{(2,3)}$
95.51 (2.1)	96.63 (4.2)	95.51 (3.4)	96.07 (5.2)

Table 5: 10-fold Cross-validation results of two models for DNA

(O)	$(\varepsilon\text{-PS2})_{(1,2)}$	$(\varepsilon\text{-PS2})_{(1,3)}$	$(\varepsilon\text{-PS2})_{(2,3)}$
92.15	92.25	92.25	92.15

discriminant functions better than those obtained by (O), and they can be found by (ε-P2) and the hybrid model. On the other hand, since IMB is numerically instable, especially for large-scale problem, (ε-P2) is comprehensively superior to other models.

Besides, in this experiment, the initial solutions $(\bar{w}, \bar{b}, \bar{\sigma})$ required in IMB were set by solutions obtained by (O) or (ε-P2), while constants ε_{-rs} for (ε-P2) were determined on the basis of the solutions obtained by (O). However, we can obtain many (weakly) Pareto optimal solutions of the proposed multiobjective model (M2) by using other various initial solutions or constants. In particular, we can find various kinds of weakly Pareto optimal solutions by (ε-P2) as constants ε_{-rs} are varied as mentioned at 4.2. Therefore, note that the proposed model may have better discriminant functions as (weakly) Pareto optimal solutions of (M2).

6 Conclusion

In this paper, we have focused on the *all together* model of the support vector machine (SVM) for multiclass classification, which uses a piece-wise linear function to construct a discriminant function. We have pointed out that for each pair of classes, the functional margin maximized in the original *all together* method is not necessarily equal to the geometric margin which is defined as the minimal distance of patterns of a pair of classes to the corresponding discriminant hyperplane classifying all patterns in both classes correctly, and that maximizing geometric margins is important for the generalization of multiclass classification. Moreover, although the sum of functional margins between all pairs of classes is maximized in the existing model, we have emphasized that the multiclass classification should be essentially formulated as a multiobjective optimization problem which maximizes all of the geometric margins simultaneously.

Therefore, we have proposed a multiobjective SVM model whose objective functions represent exactly the geometric margins of discriminant hyperplanes. In order to solve the multiobjective model, we have derived single-objective models by the scalarization approaches, ε-constraint and Benson's methods, and transformed them into solvable second-order cone programming (SOCP) problems, (ε-P2) and (P2max-sum), which can be efficiently solved by several interior point methods. Moreover, we have theoretically shown that a weakly Pareto optimal solution of the multiobjective problem is obtained by solving a single (ε-P2), while a Pareto optimal solution is obtained by solving (P2max-sum) iteratively, which is called IMB, and we have verified those results through some numerical examples. In particular, we have observed that many kinds of weakly Pareto optimal solutions can be found by solving (ε-P2) as the parameter vector ε is varied. In addition, we have applied the existing and two proposed models and a hybrid model of them to two classification problems using real-world data sets. The results show that the three proposed models can maximize geometric margins, and that (ε-P2) and the hybrid model are better than IMB and the existing model in the sense of generalization for those data sets, while IMB and the hybrid model are not suitable for a large-scaled problem. Thus, we can conclude that (ε-P2) is comprehensively superior to other models.

In this paper, we have mainly focused on the analysis of solutions obtained by the existing and proposed models in the sense of Pareto optimality. Therefore, for further tasks we should investigate various kinds of Pareto optimal solutions of the proposed multiobjective model to evaluate its potential ability properly, and, moreover, we should apply them to a wide variety of classification problems. Additionally, we have to develop the proposed models further in order to apply it to classification problems including noisy data or outliers. At the same time, through numerical experiments, we need to inspect the relation between the

geometric or functional margin and the generalization ability, which is an issue in the future.

Acknowledgment

The authors are grateful to the two anonymous referees whose comments and suggestions led to an improved version of this paper.

References

[1] S. Abe, *Support Vector Machines for Pattern Classification*, Springer, London, 2005

[2] F. Alizadeh and D. Goldfarb, Second-order cone programming, *Mathematical Programming* Ser. B, 95 (2003) 3–51.

[3] E. D. Andersen, C. Roos and T. Terlaky, On implementing a primal-dual interior-point method for conic quadratic optimization, *Mathematical Programming* Ser. B, 95 (2003) 249–277.

[4] L. Bottou, C. Cortes, J. Denker, H. Drucker, I. Guyon, L. Jackel, Y. LeCun, U. Muller, E. Sackinger, P. Simard, and V. Vapnik, Comparison of classifier methods: A case study in handwriting digit recognition, in *Proc. Int. Conf. Pattern Recognition*, IEEE Computer Society Press, 1994, pp. 77–87.

[5] E. J. Bredensteiner and K. P. Bennett, Multicategory classification by support vector machines, *Computational Optimization and Applications* 12 (1999) 53-79.

[6] K. Crammer and Y. Singer, On the learnability and design of output codes for multiclass problems, *Machine Learning* 47 (2002) 201–233.

[7] M. Ehrgott, *Multicriteria optimization*, Springer, Berlin, 2005.

[8] Y. Guermeur, Combining discriminant models with new multiclass SVMs, *Neuro COLT2 Technical Report Series* (2000)

[9] C. W. Hsh and C. J. Lin, A comparison of methods for multiclass support vector machines, *IEEE Transactions on Neural Networks* 13 (2002) 181–201.

[10] C. W. Hsu and C. J. Lin, A simple decomposition method for support vector machines, *Machine Learning* 46 (2002) 291–314.

[11] U. Kressel, Pairwise classification and support vector machines, in *Advances in kernel methods – Support vector learning*, B. Schölkopf, C. Burges, and A. J. Smola (eds), MIT Press, Cambridge, 1999, pp. 255–268,

[12] H.D. Mittelmann, An independent benchmarking of SDP and SOCP solvers, *Mathematical Programming* Ser. B, 95 (2003) 407–430.

[13] K. R. Müller, S. Mika, G. Rätsch, K. Tsuda and B. Shölkopf, An introduction to kernel-based learning algorithms, *IEEE Transactions on Neural Networks* 12 (2001) 181–201.

[14] J. C. Platt, N. Cristianini and J. Shawe-Taylor, Large margin DAG's for multiclass classification, in *Advances in Neural Information Processing Systems*, MA: MIT Press, Cambridge, 12 2000, pp. 547–553.

[15] K. Tatsumi, K. Hayashida, H. Higashi and T. Tanino, Multi-objective multiclass support vector machine for pattern recognition, in *Proceedings of SICE2007*, 2007, pp. 1095–1098.

[16] UCI benchmark repository of artificial and real data sets, University of California Irvine, `http://www.ics.uci.edu/~mlearn/databases/`

[17] V.N. Vapnik, *Statistical learning theory*, A Wiley-Interscience Publication, New York, 1998.

[18] J. Weston and C. Watkins, Multi-class support vector machines, Technical report CSD-TR-98-04, Univ. London, Royal Holloway, 1998.

Manuscript received 31 March 2008
revised 15 January 2009
accepted for publication 15 January 2009

Keiji Tatsumi
Division of Electrical, Electronic and Information Engineering, Graduate School of Engineering
Osaka University, 2-1 Yamada-oka, Suita, Osaka, Japan
E-mail address: `tatsumi@eei.eng.osaka-u.ac.jp`

Kenji Hayashida
Division of Electrical, Electronic and Information Engineering, Graduate School of Engineering
Osaka University, 2-1 Yamada-oka, Suita, Osaka, Japan
E-mail address: `hayashida@sa.eei.eng.osaka-u.ac.jp`

Ryo Kawachi
Division of Electrical, Electronic and Information Engineering, Graduate School of Engineering
Osaka University, 2-1 Yamada-oka, Suita, Osaka, Japan
E-mail address: `kawachi@sa.eei.eng.osaka-u.ac.jp`

Tetsuzo Tanino
Division of Electrical, Electronic and Information Engineering, Graduate School of Engineering
Osaka University, 2-1 Yamada-oka, Suita, Osaka, Japan
E-mail address: `tanino@eei.eng.osaka-u.ac.jp`