数理解析研究所講究録
第1685巻 2010年 210-220

Multiobjective Multiclass Support Vector Machines Using Kernel Functions

大阪大学大学院工学研究科 河内 諒 (Ryo Kawachi)
大阪大学大学院工学研究科 巽 啓司 (Keiji Tatsumi)
大阪大学大学院工学研究科 谷野 哲三 (Tetsuzo Tanino)
Graduate School of Engineering, Osaka University

Abstract

The support vector machines (SVM) is originally designed for binary classification having high generalization ability, and thus, many kinds of extended models have been investigated for multiclass classification. In this paper, we deal with the *all together* model, which classifies all patterns into the corresponding classes at once, and especially focus on a multiobjective multiclass SVM model which was proposed as a novel all together model maximizing all of the geometric margins simultaneously. Although the model is reported to have high generalization ability, it is formulated as a linear model. Hence, the model can be applied to only some kinds of classification problems.

Therefore, in this paper, we extend the linear model into a nonlinear one to which the kernel method can be applied, where weight vectors of the discriminant function are represented by linear sums of the training patterns in the feature space. Moreover, we introduce a single-objective optimization model by exploiting the ε-constraint method and the coordinate transformation in order to solve the proposed multiobjective model. The single-objective model can be regarded as a second-order cone programming (SOCP) problem and its optimal solution is Pareto optimal for the proposed nonlinear multiobjective SVM. Furthermore through numerical experiments we verify that the proposed nonlinear model maximizes the geometric margins in the sense of multiobjective optimization and has good generalization ability.

1. Introduction

The support vector machine (SVM) has become greatly popular in the machine learning community because it has high generalization ability to solve binary classification problems, however, extending it for multiclass classification is still an ongoing research issue. Among the extended SVM models, all together model finds a discriminant function by solving directly a single optimization problem with all patterns, where all patterns are classified into the corresponding classes by using a piecewise linear function [2, 4, 9, 10]. In this paper, we focus on the model. The existing all together model is a formulated as a single-objective optimization problem of maximizing the sum of functional margins between all of the pairs of classes, where the functional margin is defined as the distance between two normalized support hyperplanes parallel to the corresponding discriminant hyperplane. However, as we point out it in [6], there exists a gap between the functional margin and the geometric margin which is defined as the minimal

distance of patterns to the corresponding discriminant hyperplane in some examples, and the geometric margin can exactly indicate the relation between each pattern and the discriminant function. Therefore, in [6], we emphasized that maximizing the geometric margins is important for the generalization of multiclass classification, and proposed a linear multiobjective multiclass SVM model which maximizes all of the geometric margins simultaneously. Moreover, we derived a single-objective second-order cone programming (SOCP) problem by using scalarization approaches for multiobjective optimization, which are solvable convex programming problems, and showed theoretically that the optimal solution of the SOCP is Pareto optimal of the proposed multiobjective model. Moreover, we applied them to some examples to demonstrate that the proposed models can achieve maximization of the geometric margins.

However, since the model uses a piecewise linear classifier, it is difficult to discriminate piecewise linearly inseparable data correctly, which are often seen in the real-world problems. In this paper, we extend the linear multiobjective multiclass model into a nonlinear one to which the kernel method can be applied. In the proposed model, weight vectors of the discriminant function are represented by linear sums of the data in the feature space. Then, similarly to the linear models, we derive a single-objective optimization model based on ε-constraint method to obtain a Pareto optimal solution of the proposed nonlinear model, and show that into a standard SOCP problem is obtained from the single-objective model by transforming its coordinate system. Finally, through numerical experiments we verify that the SOCP model maximizes the geometric margins in the sense of multiobjective optimization and compare the classification abilities of the proposed and the existing models.

2. Multiclass classification

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

$$f(x) = \arg\max_p \{w^{p\top} x + b^p\}, \tag{1}$$

where $w^p \in R^n$ and b^p, $p \in P$ 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. Then,

$$(w^p - w^q)^\top x + b^p - b^q = 0, q \neq p, p, q \in P, \tag{2}$$

is the discriminant hyperplane which distinguishes between classes p and q. Note that the representation of discriminant hyperplane (2) is not unique. For any constants $t(\neq 0), s \in R$ and any vector $v \in R^n$, $(w^{1\top}, ..., w^{\nu\top}), (b^1, ..., b^\nu)$ and $(tw^{1\top} + v^\top, ..., tw^{\nu\top} + v^\top), (tb^1 + s, ..., tb^k + s)$ are different representations of the same discriminant hyperplanes.

Now, we suppose that data D are piecewise linearly separable. Then, there exist an infinite number of discriminant functions to distinguish all classes correctly. In the multiclass classification, the model maximizing $1/\|w^p - w^q\|$ for all pairs $(p, q), q \neq p, p, q \in P$ similar to the binary

SVM was proposed [2, 4, 9, 10],

$$\text{(O)} \quad \begin{array}{ll} \min\limits_{w,b} & \dfrac{1}{2}\sum\limits_{p=1}^{\nu}\sum\limits_{q>p}\|w^p - w^q\|^2 \\ \text{s.t} & (w^p - w^q)^\top x^i + (b^p - b^q) \geq 1, \quad i \in I_p, \quad q > p, \quad p, q \in P, \end{array}$$

where I_p denotes an index set defined by $I_p := \{i \in \{1, ..., m\} | y^i = p\}$. However, the margin in model (O) 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, as we pointed out it in [6].

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

Thus, it cannot guarantee that margins obtained by minimizing $\|w^p - w^q\|, q \neq p \in P$ in model (O) are equal to the corresponding geometric margins $d^g_{pq}(w,b)$. Therefore, in [6] we proposed a linear hard-margin multiobjective SVM (M1) which maximizes all geometric margins simultaneously.

$$\text{(M1)} \quad \begin{array}{ll} \max\limits_{w,b} & \left(d^g_{12}(w,b), d^g_{13}(w,b), \ldots, d^g_{(\nu-1)\nu}(w,b)\right) \\ \text{s.t.} & (w^p - w^q)^\top x^i + (b^p - b^q) \geq 1, \quad i \in I_p, \quad q \neq p, \quad p, q \in P. \end{array}$$

Moreover, since model (M1) is difficult to solve directly, we proposed the following model (M2) using a vector $\sigma \in R^{k(k-1)/2}$:

$$\text{(M2)} \quad \begin{array}{ll} \max\limits_{w,b,\sigma} & \left(\dfrac{\sigma_{12}}{\|w^1 - w^2\|}, \ldots, \dfrac{\sigma_{(\nu-1)\nu}}{\|w^{(\nu-1)} - w^\nu\|}\right) \\ \text{s.t.} & (w^p - w^q)^\top x^i + (b^p - b^q) \geq \sigma_{pq}, \quad i \in I_p, \quad q > p, \quad p, q \in P, \\ & (w^q - w^p)^\top x^i + (b^q - b^p) \geq \sigma_{pq}, \quad i \in I_q, \quad q > p, \quad p, q \in P, \\ & \sigma_{pq} \geq 1, \quad q > p, \quad p, q \in P. \end{array}$$

Then, we showed that if there exist Pareto optimal solutions of (M2), the optimal solutions of (M2) can be considered to be equivalent to those of (M1) as follows [7].

Theorem 1 *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), *where an element of* σ *is defined by*

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

In addition, we derived two kinds of single-objective optimization problems by scalarization approaches to multiobjective optimization, ε-constraint approach and Benson's method, and transform them into single-objective second-order cone programming (SOCP) problems. 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 [1]. Moreover, several commercial and noncommercial solvers have been developed [5]. Furthermore, we showed theoretically that Pareto optimal solutions of the multiobjective problem (M2) can be obtained by solving SOCP models, and applied them to some examples to demonstrate that they can achieve maximization of the geometric margins [7]. Furthermore, we extended the models into the soft-margin ones for piecewise linearly inseparable data in the real-world classification problem [6].

However, there exist various kinds of classification data for which the linear models have poor performance. Thus, we extend the multiobjective multiclass SVM to a nonlinear one for the data in the next section.

3. Nonlinear multiobjective model maximizing geometric margins

In this section, we focus on the classification problem for data which cannot be correctly discriminated by the linear SVMs. To such data, nonlinear SVMs have been applied, which classify the data by a linear discriminant function in a high dimensional feature space F by using an appropriate mapping $\phi : R^n \to F$. Namely, a nonlinear classifier is trained as a linear model for images of the data, $\phi(x^1), ..., \phi(x^m)$. Therefore, our aim is finding a suitable classifier for the data:

$$f(x) = \arg\max_p \{w^{p\top}\phi(x) + b^p\}. \tag{3}$$

In this and subsequent sections, although we only discuss hard-margin nonlinear models, note that they can be easily extended into soft margin ones. Thus, we suppose that given data D are piecewise linearly inseparable in R^n, but piecewise linearly separable in an appropriate F. Then, the classification problem can be formulated as follows, which is derived by replacing x^i of (O) with $\phi(x^i)$ [4].

$$\text{(NO)} \quad \begin{array}{ll} \min\limits_{w,b} & \dfrac{1}{2}\sum\limits_{p=1}^{\nu}\sum\limits_{q>p} \|w^p - w^q\|^2 \\ \text{s.t} & (w^p - w^q)^\top \phi(x^i) + (b^p - b^q) \geq 1, \quad i \in I_p, \quad q > p, \quad p, q \in P. \end{array}$$

In the nonlinear model, the geometric margin is redefined by

$$d^g_{pq}(w,b) = \min\left\{ \min_{i\in I_p} \frac{|(w^p - w^q)^\top \phi(x^i) + (b^p - b^q)|}{\|w^p - w^q\|}, \ \min_{i\in I_q} \frac{|(w^p - w^q)^\top \phi(x^i) + (b^p - b^q)|}{\|w^p - w^q\|} \right\}, \quad q > p, \ \ p, q \in P.$$

In addition, we consider the following dual problems of (NO):

$$
\begin{aligned}
\min_{\alpha} \quad & \frac{1}{2}\sum_{p=1}^{\nu}\sum_{q\neq p}\sum_{r\neq p}\left\{\sum_{i\in I_p}\sum_{j\in I_p}\alpha_{pqi}\alpha_{prj}\phi(x^i)^\top\phi(x^j) - 2\sum_{i\in I_p}\sum_{j\in I_r}\alpha_{pqi}\alpha_{rpj}\phi(x^i)^\top\phi(x^j)\right. \\
\text{(DNO)} \quad & \left. + \sum_{i\in I_q}\sum_{j\in I_r}\alpha_{qpi}\alpha_{rpj}\phi(x^i)^\top\phi(x^j)\right\} - \sum_{p=1}^{\nu}\sum_{q>p}\sum_{i\in I_p}\alpha_{pqi} \\
\text{s.t.} \quad & \sum_{q\neq p}\sum_{i\in I_p}\alpha_{pqi} = \sum_{q\neq p}\sum_{i\in I_q}\alpha_{qpi}, \quad p\in P \\
& \alpha_{pqi}\geq 0, \quad i\in I_p, \quad q\neq p, \quad p,q\in P,
\end{aligned}
$$

where $\alpha_{pqi}, i\in I_p, p, q\in P$ is a dual variable. Then, by using the optimal solution α^* of (DNO), (3) is rewritten as follows:

$$
f(x) = \arg\max_{p}\left\{\sum_{q\neq p}\sum_{i\in I_p}\alpha^*_{pqi}\phi(x^i)^\top\phi(x) - \sum_{q\neq p}\sum_{i\in I_q}\alpha^*_{qpi}\phi(x^i)^\top\phi(x) + b^p\right\}. \tag{4}
$$

(DNO) has often been used because it can be applied the kernel method to. The kernel method enables us to solve (DNO) without calculating images of data $\phi(x^i), i = 1, ..., m$, but rather by simply calculating the inner products $\phi(x^i)^\top\phi(x^j)$ between images of all pairs $(x^i, x^j), i, j = 1, ..., m$. It is well-known that kernel functions $k : R^n \times R^n \to R$ satisfying Mercer's Theorem are guaranteed to have a function $\phi : R^n \to F$ such that $k(x, x') = \phi(x)^\top\phi(x')$ for any $x, x' \in R^n$ [8]. Moreover, $\phi(x)$ appears in the form of the inner product in (DNO) and (4). Therefore, in (DNO) all of the inner products can be replaced with the kernel functions. Then, even if the dimension of F is quite high, we can solve (DNO) without suffering from the curse of dimensionality. For major kernel functions, polynomial kernels $k(x, x') = (x^\top x' + 1)^d$ with a integer constant d, and RBF kernels $k(x, x') = \exp(-\gamma\|x - x'\|^2)$ with a real constant γ are commonly used.

However, since (DNO) maximizes functional margins $1/\|w^p - w^q\|$ for all pairs $(p, q), q \neq p, p, q \in P$ in the same way as (O), we derive a nonlinear multiobjective model maximizing all geometric margins. Now, we extend the linear model (M2) into a nonlinear one similarly to (NO) as follows:

$$
\begin{aligned}
\max_{w,b,\sigma} \quad & \left(\frac{\sigma_{12}}{\|w^1 - w^2\|}, \cdots, \frac{\sigma_{(\nu-1)\nu}}{\|w^{\nu-1} - w^\nu\|}\right) \\
\text{(MN)} \quad \text{s.t.} \quad & (w^p - w^q)^\top\phi(x^i) + (b^p - b^q) \geq \sigma_{pq}, \quad i\in I_p, \quad q > p, \quad p, q\in P, \\
& (w^q - w^p)^\top\phi(x^i) + (b^q - b^p) \geq \sigma_{pq}, \quad i\in I_q, \quad q > p, \quad p, q\in P, \\
& \sigma_{pq}\geq 1, \quad q > p, \quad p, q\in P.
\end{aligned}
$$

(MN) is a nonlinear multiobjective SVM which maximizes all geometric margins in F simultaneously. However, since any $\phi(x^i)$ does not appear in the form of inner product differently to (NO), we cannot straightforwardly apply the kernel method to (MN). Thus, we propose a model by limiting the feasible region of (MN) to which the kernel method can be applied. The following constraints are imposed on (MN) by introducing variables β_i^p, $i = 1, ..., m$, $p\in P$.

$$
w^p = \sum_{i=1}^{m}\beta_i^p\phi(x^i), \quad p\in P, \tag{5}
$$

which means that each w^p can be represented by a linear sum of $\phi(x^i)$. This kind of representation has been used in many binary SVM, and it is reported that they are high generalization ability. Then, in (MN) with (5), $\phi(x)$ appears in the form of the inner product. Similarly to (DNO) they can be replaced with kernel functions. Here, we define $\Phi(x) = (\phi(x^1), ..., \phi(x^m))$, $\beta^p = (\beta_1^p, ..., \beta_m^p)^\top, p \in P$ and $K(x,x) = \Phi(x)^\top \Phi(x)$, $K(x,x)$ is called a kernel matrix. The denominator of objective function in (MN) is transformed as follows:

$$\|w^p - w^q\| = \|\Phi(x)(\beta^p - \beta^q)\| = \|\beta^p - \beta^q\|_{\Phi(x)^\top \Phi(x)} = \|\beta^p - \beta^q\|_{K(x,x)}, \tag{6}$$

where for any semi-definite positive symmetric matrix A, $\|u\|_A$ is defined by $\|u\|_A = \sqrt{u^\top A u}$. Then, the substitution of (5) and (6) into (MN) gives us the following model:

$$\begin{array}{lll} & \max\limits_{\beta,b,\sigma} & \left(\dfrac{\sigma_{12}}{\|\beta^1 - \beta^2\|_{K(x,x)}}, ..., \dfrac{\sigma_{(k-1)k}}{\|\beta^{k-1} - \beta^k\|_{K(x,x)}} \right) \\ \text{(MN2)} & \text{s.t.} & (\beta^p - \beta^q)^\top \kappa(x, x^i) + (b^p - b^q) \geq \sigma_{pq}, \ i \in I_p, \ q > p, \ p, q \in P, \\ & & (\beta^q - \beta^p)^\top \kappa(x, x^i) + (b^q - b^p) \geq \sigma_{pq}, \ i \in I_q, \ q > p, \ p, q \in P, \\ & & \sigma_{pq} \geq 1, \ \ q > p, \ \ p, q \in P, \end{array}$$

where $\kappa(x, x^i)$ is defined by $\kappa(x, x^i) = (k(x^1, x^i), ..., k(x^m, x^i))^\top, \ \ i = 1, ..., m$. Then, we can obtain the discriminant function for an unseen sample $\bar{x}$ by using the optimal solution (β^*, b^*, σ^*) of (MN2) as follows:

$$f(\bar{x}) = \arg\max_p \left\{ \beta^{p*\top} \kappa(x, \bar{x}) + b^{p*} \right\}. \tag{7}$$

Now, we obtain the nonlinear multiobjective model (MN2) maximizing geometric margins. Next, let us consider solving (MN2) to obtain its Pareto optimal solutions.

4. SOCP model based on ε-constraint method

In this section, we propose a method of obtaining a Pareto optimal solution of (MN2). First, we derive the following single-objective model by using a scalarization approach, ε-constraint method, to (MN2) similarly to the linear multiobjective multiclass model:

$$\begin{array}{lll} & \max\limits_{\beta,b,\sigma} & \dfrac{\sigma_{rs}}{\|\beta^r - \beta^s\|_{K(x,x)}} \\ & \text{s.t.} & \dfrac{\sigma_{pq}}{\|\beta^p - \beta^q\|_{K(x,x)}} \geq \varepsilon_{pq}, \ \ q > p, \ \ (p,q) \neq (r,s), \ \ p, q \in P, \\ (\varepsilon - \text{MN}) & & (\beta^p - \beta^q)^\top \kappa(x, x^i) + (b^p - b^q) \geq \sigma_{pq}, \ \ i \in I_p, \ \ q > p, \ \ p, q \in P, \\ & & (\beta^q - \beta^p)^\top \kappa(x, x^i) + (b^q - b^p) \geq \sigma_{pq}, \ \ i \in I_q, \ \ q > p, \ \ p, q \in P, \\ & & \sigma_{pq} \geq 1, \ \ q > p, \ \ p, q \in P, \end{array}$$

where a pair (r, s) and constants $\varepsilon_{pq}, q > p, (p, q) \neq (r, s), p, q \in P$ are appropriately selected such that the feasible region of (ε-MN) is not empty. This method maximizes only one of the objectives of (MN2) while the others are transformed to constraints with ε_{pq}. Then, the following theorems are known.

Theorem 2 *[3] Let (β, b, σ) be an optimal solution of* (ε-MN) *for some (r, s). Then (β, b, σ) is weakly Pareto optimal for* (MN2).

Theorem 3 *[3] (β, b, σ) is Pareto optimal for* (MN2) *if and only if there exists an ε_{-rs} such that (β, b, σ) is optimal for* (ε-MN) *for any (r, s), $r, s \in P$.*

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 (MN2) can be obtained by solving (ε-MN) with an appropriate choice of ε_{-rs}.

Although (ε-MN) is a single-objective model, (ε-MN) is also difficult to solve because of its fractional constraints and objective functions. Hence, by making use of one degree of freedom of (ε-MN), we add a constraint $\sigma_{rs} = c_{rs}$ with an appropriate constant c_{rs} to obtain the following model:

$$
(\varepsilon - \mathrm{MN2}) \quad
\begin{array}{ll}
\max\limits_{\beta,b,\sigma} & \dfrac{c_{rs}}{\|\beta^r - \beta^s\|_{K(x,x)}} \\
\text{s.t.} & \dfrac{\sigma_{pq}}{\|\beta^p - \beta^q\|_{K(x,x)}} \geq \varepsilon_{pq}, \quad q > p, \quad (p,q) \neq (r,s), \quad p, q \in P, \\
& (\beta^r - \beta^s)^\top \kappa(x, x^i) + (b^r - b^s) \geq c_{rs}, \quad i \in I_r, \\
& (\beta^s - \beta^r)^\top \kappa(x, x^i) + (b^s - b^r) \geq c_{rs}, \quad i \in I_s, \\
& (\beta^p - \beta^q)^\top \kappa(x, x^i) + (b^p - b^q) \geq \sigma_{pq}, \quad i \in I_p, \quad q > p, \quad (p,q) \neq (r,s), \quad p, q \in P, \\
& (\beta^p - \beta^q)^\top \kappa(x, x^i) + (b^q - b^p) \geq \sigma_{pq}, \quad i \in I_q, \quad q > p, \quad (p,q) \neq (r,s), \quad p, q \in P, \\
& \sigma_{pq} \geq 1, \quad q > p, \quad (p,q) \neq (r,s), \quad p, q \in P,
\end{array}
$$

where (β, b, σ_{-rs}) denotes the vector in which the element σ_{rs} is removed from (β, b, σ). Moreover, for a solution (β, b, σ_{-rs}) of (ε-MN2), we define a vector $(\beta, b, (\sigma_{-rs}, c_{rs}))$ whose element σ_{rs} is c_{rs} and other elements are equal to (β, b, σ_{-rs}). Now, we can show the following theorems about relations between problems (ε-MN) and (ε-MN2).

Theorem 4 *Let $(\hat{\beta}, \hat{b}, \hat{\sigma})$ be an optimal solution of* (ε-MN) *and $c_{rs} = t\hat{\sigma}_{rs}$ for $t \geq 1$. If $(\beta^*, b^*, \sigma^*_{-rs})$ is an optimal solution of* (ε-MN2), *then $(\beta^*, b^*, (\sigma^*_{-rs}, c_{rs}))$ is optimal for* (ε-MN).

Theorem 5 *If (β^*, b^*, σ^*) is an optimal solution of* (ε-MN), *then for any $t \geq 1$, $(t\beta^*, tb^*, t\sigma^*_{-rs})$ is an optimal solution of* (ε-MN2) *with $c_{rs} = t\sigma^*_{rs}$.*

Theorem 4 shows that for an optimal solution $(\beta^*, b^*, \sigma^*_{-rs})$ of (ε-MN2), $(\beta^*, b^*, (\sigma^*_{-rs}, c_{rs}))$ is optimal for (ε-MN). Thus, Theorem 2 implies that the optimal solution is weakly Pareto optimal for (MN2). In addition, the result together with Theorems 3 and 5 suggests that we can obtain any Pareto optimal solution of (MN2) is obtained by solving (ε-MN) with an appropriate choice of ε_{-rs}. Consequently, we can conclude that various discriminant functions which determined by Pareto optimal solutions of (MN2) can be obtained by solving (ε-MN2) as a pair (r, s) and the corresponding parameter ε_{-rs} are varied.

Next let us consider how to solve (ε-MN2). Although (ε-MN2) can be regarded as a SOCP problem, and thus, it is efficiently solvable as mentioned at Section 2, (ε-MN2) is not a standard form of SOCP. Therefore, we introduce a coordinate transformation for (ε-MN2). First, since

$K(x,x)$ is a positive semidefinite, it can be diagonalized by using an orthogonal matrix T as follows:

$$K(x,x) = T\Lambda T^{\top}, \tag{8}$$

where Λ is a diagonal matrix whose elements are the corresponding eigenvalues of $K(x,x)$. Now, the number of positive eigenvalues of $K(x,x)$ is represented by l. Then, let us define a diagonal matrix $\bar{\Lambda} \in R^{l\times l}$ whose elements are positive eigenvalues of $K(x,x)$, and $\bar{T} \in R^{m\times l}$ consisting of the corresponding l column vectors of T. Moreover, $\bar{t}_i^{\top}$ is the i-th row vectors of $\bar{T}$. Then, we have

$$\begin{aligned} K(x,x) &= T\Lambda T^{\top} = \bar{T}\bar{\Lambda}\bar{T}^{\top}, & (9)\\ \kappa(x,x^i) &= \bar{T}\bar{\Lambda}\bar{t}_i. & (10) \end{aligned}$$

Here, we define a new decision variable z^p by $z^p = \bar{\Lambda}^{\frac{1}{2}}\bar{T}^{\top}\beta^p$, $p \in P$, then we can obtain from (10) and (10)

$$\|\beta^p - \beta^q\|_{K(x,x)} = \|\beta^p - \beta^q\|_{\bar{T}\bar{\Lambda}\bar{T}^{\top}} = \|\bar{\Lambda}^{\frac{1}{2}}\bar{T}^{\top}(\beta^p - \beta^q)\| = \|z^p - z^q\|, \tag{11}$$

and,

$$(\beta^p - \beta^q)^{\top}\kappa(x,x^i) = (\beta^p - \beta^q)^{\top}\bar{T}\bar{\Lambda}\bar{t}^i = (z^p - z^q)^{\top}\bar{\Lambda}^{\frac{1}{2}}\bar{t}^i. \tag{12}$$

Thus, consequently we can transform (ε-MN2) into the following model from (11) and (12).

$$(\varepsilon - \text{MN3}) \quad \begin{array}{ll} \max\limits_{z,b,\sigma} & \dfrac{c_{rs}}{\|z^r - z^s\|} \\ \text{s.t.} & \dfrac{\sigma_{pq}}{\|z^p - z^q\|} \geq \varepsilon_{pq}, \quad q > p, \quad (p,q) \neq (r,s), \quad p,q \in P, \\ & (z^r - z^s)^{\top}\bar{\Lambda}^{\frac{1}{2}}\bar{t}^i + (b^r - b^s) \geq c_{rs}, \quad i \in I_r, \\ & (z^s - z^r)^{\top}\bar{\Lambda}^{\frac{1}{2}}\bar{t}^i + (b^s - b^r) \geq c_{rs}, \quad i \in I_s, \\ & (z^p - z^q)^{\top}\bar{\Lambda}^{\frac{1}{2}}\bar{t}^i + (b^p - b^q) \geq \sigma_{pq}, \quad i \in I_p, \quad q > p, \quad (p,q) \neq (r,s), \quad p,q \in P, \\ & (z^q - z^p)^{\top}\bar{\Lambda}^{\frac{1}{2}}\bar{t}^i + (b^q - b^p) \geq \sigma_{pq}, \quad i \in I_q, \quad q > p, \quad (p,q) \neq (r,s), \quad p,q \in P, \\ & \sigma_{pq} \geq 1, \quad q > p, \quad (p,q) \neq (r,s), \quad p,q \in P. \end{array}$$

Since (ε-MN2) and (ε-MN3) are equivalent, Theorems 2 - 5 imply that any Pareto optimal solutions of (MN2) by solving (ε-MN3) with an appropriate choice of ε_{-rs} instead of (ε-MN2). Then, we can obtain the discriminant function for an unseen sample $\bar{x}$ by using the optimal solution (z^*, b^*, σ^*) of (ε-MN3) as follows:

$$f(\bar{x}) = \arg\max_p \{z^{p*\top}(\bar{\Lambda}^{-\frac{1}{2}}T^{\top})\kappa(x,\bar{x}) + b^{p*}\}. \tag{13}$$

Finally, in the next section, we apply the proposed models to some classification problems.

5. Numerical Examples

In this section, we report the results of numerical experiments where existing model (NO) and the proposed model (MN2) were applied to a bench mark problem, Balance Scale dataset. We used (ε-MN3) in order to solve (MN2). Balance Scale dataset is a three-class classification problem involving 625 patterns with four features. We used the ten-fold cross-validation to estimate generalization abilities of two models; we randomly partitioned the dataset into ten subsets, and made ten different datasets consisting of a single test subset and nine training subsets, where each subset is used exactly once for the test one. Then, the average of test data accuracies was used as the generalization ability estimator of each classifier. We used optimization tools in MathWorks Matlab 7.0.1 and Mosek version 5.0 to solve two models. The kernel functions used in two models are polynomial kernels with varying $d \in \{2, 3, 4\}$ and RBF kernels with varying $\gamma \in \{0.1, 0.5\}$. The fixed pair (r, s) in (ε-MN3) was selected as all pairs of classes and $c_{rs} = 10$, and constant ε_{-rs} in (ε-MN3) is determined by the geometric margins of the optimal solution of the model (DNO).

Table 1 shows geometric margins obtained by two models with the polynomial kernel with $d = 3$ in the ten-fold cross-validation. We can observe that solutions obtained by the proposed

Table 1: Comparison of geometric margins for two models (Polynomial kernels [d = 3])

Model	d_{12}^g	d_{13}^g	d_{23}^g
(ON)	2.121	2.134	4.263
(ε-MN3)[(r,s)=(1,2)]	2.249	2.191	4.332
(ε-MN3)[(r,s)=(1,3)]	2.173	2.254	4.327
(ε-MN3)[(r,s)=(2,3)]	2.121	2.134	4.448

model dominate ones by the existing model. These results indicate that Pareto optimal solutions of (MN2) are obtained by (ε-MN3). Table 2 shows the generalization ability estimators for two models when kernel functions varies, which means that the generalization ability of the proposed

Table 2: Comparison of generalization ability estimator obtained by varying kernel functions for two models

Kernel functions	(ON)	(ε – MN3) $(r, s) = (1, 2)$	(ε – MN3) $(r, s) = (1, 3)$	(ε – MN3) $(r, s) = (2, 3)$
Polynomial (d = 2)	100.0	100.0	100.0	100.0
Polynomial (d = 3)	98.08	99.36	99.84	99.68
Polynomial (d = 4)	97.44	98.40	98.40	96.96
RBF (γ = 0.1)	97.44	97.60	98.24	97.28
RBF (γ = 0.5)	84.00	86.56	86.72	85.28

model are better than or equal to those of the existing model. Moreover, we can observe that (ε-MN3) can improve the generalization ability of classifiers obtained by (NO). At the same

time, we can obsene that obtained solutions by (ε-MN3) considerably depends on the selection of a pair (r, s) and the constant ε_{-rs}. Therefore, we can conclude that there exists a Pareto optimal solution of (MN2) having high generalization ability..

6. Conclusion

In this paper, we have focused on the nonlinear all together model of the support vector machine (SVM) for multiclass classification. In particular, we have discussed a multiobjective SVM model maximizing all geometric margins for multiclass classification, which was proposed for the high generalization. While the existing all together model can discriminate piecewise linearly inseparable data by using the kernel method, the multiobjective model is proposed as a linear one, and it is difficult to extend it into a nonlinear model to which the kernel method can be straightforwardly applied.

Therefore, we have proposed a nonlinear multiobjective SVM model by representing weight vectors of the discriminant functions as linear sums of the training data in the feature space, to which can be applied kernel method. Then, in order to obtain a Pareto optimal solution of the proposed nonlinear model, we have derived a single-objective optimization model based on ε-constraint method, and moreover, we have transformed the single-objective model into a standard SOCP problem by transforming its coordinate system. Furthermore, we have observed that the proposed model can maximize the geometric margins and obtain classifiers with the high generalization ability through some numerical experiments.

For further tasks, we should apply the proposed nonlinear multiobjective model to many kinds of classification problems with more than three classes in order to investigate its performance. Moreover, since we have focused on nonlinear hard-margin model in this paper, we will extend the proposed model into a soft-margin model, and apply it to classification problems including noisy data or outliers.

References

[1] F. Alizadeh and D. Goldfarb, "Second-order cone programming," *Mathematical Programming*, Ser. B, 95, 3-51, 2003.

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

[3] M. Ehrgott, *Multicriteria Optimization. 2nd ed.* , Springer, Berlin, 2005

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

[5] H.D. Mittelmann, "An independent benchmarking of SDP and SOCP solvers," *Mathematical Programming*, Ser. B, 95, 407-430, 2003.

[6] K. Tatsumi, R. Kawachi, K. Hayashida, and T. Tanino, "Multiobjective multiclass soft-margin support vector machine maximizing pair-wise interclass margins," *Proceedings of ICONIP08*, Part I, LNCS 5506, pp. 967-974, 2009.

[7] K. Tatsumi, R. Kawachi, K. Hayashida, and T. Tanino, "Multiobjective multiclass support vector machines maximizing geometric margins," to appear *in Pacific Journal of Optimization.*

[8] J. S. Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*, Cambridge University Press, 2004

[9] V. Vapnik, *Statistical Learning Theory*, A Wiley-Interscience Publication, 1998.

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