ATLANTIS PRESS

Atlantis Studies in Uncertainty Modelling, volume 1

11th Conference of the European Society for Fuzzy Logic and Technology (EUSFLAT 2019)

Scalarization for Approximate Multiobjective Multiclass Support Vector Machine Using the Large-k Norm

Yoshifumi Kusunoki and Keiji Tatsumi
Graduate School of Engineering, Osaka University,
2-1, Yamadaoka, Suita, Osaka 565-0871, Japan,
{kusunoki,tatsumi}@eei.eng.osaka-u.ac.jp

Abstract

We study a multiclass extension of support vector machine (SVM) based on geometric margin maximization. Since there are different margins for different class-pairs, the maximum-margin SVM can be formulated as a multiobjective optimization problem. This multiobjective multiclass SVM (MMSVM) is difficult to be solved because not only it is multiobjective but also it is nonconvex. In order to solve the MMSVM, we approximate it by a convex multiobjective problem, and furthermore scalarize its objective functions. In this paper, we propose a new scalarization for MMSVM using the large-k norm, which provides a spectrum between the ℓ_∞ and ℓ_1 norms.

Keywords: Support vector machine, Multiclass classification, Multiobjective optimization, Large-k norm

1 Introduction

The support vector machine (SVM) is a popular machine learning method. SVM was originally proposed for binary classification problems [3, 14]. It learns a linear classifier based on the principle of margin maximization, which stems from the geometric motivation that the hyperplane that widely separates instances of different classes attains good classification. On the other hand, in another viewpoint, the margin maximization is regarded as regularization in model selection.

There are several extensions [1, 4, 12] of SVM for multiclass classification problems, i.e., more than two classes are considered. A simple approach is one-against-all (OAA) or one-vs-all (OVA), in which a c-class problem is reduced to the c binary problems that one class is separated from the others Those problems are trained by the binary SVM. Another major approach is all-together (AT) or all-at-once (AO), in which all class-wise classifiers are learned by a single optimization problem. However, the aforementioned multiclass extensions do not exactly maximize geometric margins, which are associated with class-pairs. Especially, the existing ATSMVs are formulated as minimization of loss functions with regularization terms, instead of geometric interpretation. It was recently pointed out by Tatsumi and Tanino [13], and they have formulated a multiclass extension as a multiobjective optimization problem which simultaneously maximizes all class-pair margins. This model is called multiobjective multiclass SVM (MMSVM). One of the advantages of the multiobjective approach is to provide diversity of classifiers by Pareto optimal solutions, and systematic selection of them by scalarization techniques.

In order to obtain Pareto solutions, we consider scalarizations of MMSVM. However, due to nonconvexity of MMSVM, almost all conventional scalarization methods cannot be efficiently applied with the exception of the ε-constraint method. Tatsumi and Tanino [13] showed that classifiers obtained by the ε-constraint scalarization have better classification accuracy than those of ATSVM and OAASVM. However, the method needs high computational effort for training. Recently, a convex approximation of MMSVM have been proposed in [8]. Matsugi et al. [9] have proposed scalarizations of the approximate MMSVM using the reference point method.

In the reference point method, the augmented Tchebyshev scalarization function is used to aggregate multiple objective functions. It is regarded as a weighted sum of the Tchebyshev norm (ℓ_∞ norm) and the ℓ_1 norm. In this paper, we proposed a scalarization for MMSVM based on the large-k norm [7] instead of the augmented Tchebyshev function, which provides a spectrum between the ℓ_∞ and ℓ_1 norms. In a spacial

case, the proposed scalarization is reduced to the conventional ATSVM. It is formulated as a second-order cone programming and easily solvable by numerical solvers. By numerical experiments, we show that it outperforms ATSVM in classification capability by adjusting approximation and scalarization parameters.

This paper is organized as follows. In Section 2, after binary SVM is described as a maximum-margin classifier, MMSVM is introduced. Additionally, we explain how to extend MMSVM to a soft-margin and nonlinear model. In Section 3, we introduce approximate MMSVM and discuss properties of approximate solutions. And then, we propose the large-k norm scalarization for approximate MMSVM. In Section 4, results of numerical experiments are presented to examine performance of the proposed scalarization. Moreover, we investigate effect of parameters of the proposed method more deeply. Finally, in Section 5, concluding remarks are provided.

2 Multiobjective Multiclass Support Vector Machine

2.1 Multiclass Classification

We call n-dimensional real space $\mathbf{R}^n$ an input space, and $C = \{1, 2, \ldots, c\}$, $c \geq 2$ a set of class labels. A learning problem is to find a function $D : \mathbf{R}^n \to C$, called a classifier, using m labeled input vectors $(x_1, y_1), \ldots, (x_m, y_m) \in \mathbf{R}^n \times C$, which are called training instances. The objective of this problem is to find a classifier with high classification accuracy, i.e., it can correctly assign class labels to unseen instances as well as training ones. Let $M = \{1, \ldots, m\}$ be the index set of the training set. For $p \in C$, we define $M_p = \{i \in M \mid y_i = p\}$. Additionally, let $C^{\bar{2}} = \{pq \mid p, q \in C, p < q\}$ be the set of class label pairs. When $c = 2$, the problem is called binary classification. On the other hand, when $c \geq 3$, it is called multiclass classification. In this paper, we study multiclass classification problems.

We consider a linear classifier D given in the following form: for $x \in \mathbf{R}^n$,

$$D(x) = \operatorname*{argmax}_{p \in C} \{f_p(x) = w_p^\top x + b_p\}, \tag{1}$$

where $w_1, \ldots, w_c \in \mathbf{R}^n$ and $b_1, \ldots, b_c \in \mathbf{R}$. Each $f_p(x)$ is called a linear discriminant function of the class label p. If there is more than one label p whose value $f_p(x)$ is the maximum, we arbitrarily select one label among them. The parameters $(w_1, b_1), \ldots, (w_p, b_p)$ are trained using the training instances.

2.2 SVM and Geometric Margin Maximization

The original support vector machine (SVM) is a solution for the binary classification problems. For introduction to our proposed multiclass extension of SVM, we review the binary SVM. In binary classification, a linear classifier D is reduced to the following form:

$$D(x) = \operatorname{sgn}(f(x) = w^\top x + b) \tag{2}$$

Here, we suppose that the class labels are 1 and -1. w and b are parameters of the linear classifier. Instances x with $f(x) = 0$ are arbitrarily classified.

SVM selects a classifier whose boundary hyperplane has the largest margin. The margin $d(w, b)$ of a hyperplane $\{x \mid f(x) = 0\}$ is the distance between the hyperplane and the nearest training instance, namely,

$$d(w, b) = \frac{\min_{i \in M} |w^\top x_i + b|}{\|w\|}, \tag{3}$$

where $\|\cdot\|$ is the Euclidean norm. The largest-margin classifier is obtained by solving the following optimization problem.

$$\begin{aligned} \max_{w,b} \quad & \frac{r}{\|w\|} \\ \text{s.t.} \quad & y_i(w^\top x_i + b) \geq r > 0,\ i \in M \end{aligned} \tag{4}$$

The constraint ensures that the selected hyperplane $\{x \mid w^\top x + b = 0\}$ correctly classifies all training instances. We remark that the lower bound of r can be an arbitrary positive number. The objective function is invariant if (w, b) is multiplied by a positive value. Hence, without loss of generality, we can fix $r = 1$, and the above optimization problem is equivalent to the following.

$$\begin{aligned} \min_{w,b} \quad & \frac{1}{2}\|w\|^2 \\ \text{s.t.} \quad & y_i(w^\top x_i + b) \geq 1,\ i \in M. \end{aligned} \tag{5}$$

Here, we consider the problem minimizing the inverse-squared margin instead of maximizing the margin.

For $i \in M$, let α_i be the optimal dual variable with respect to the constraint $y_i(w^\top x_i + b) \geq 1$. A training instance x_i is called a support vector if $\alpha_i > 0$. The optimal hyperplane $\{x \mid w^\top x + b = 0\}$ depends on the set of support vectors only.

The learning model (5) has no feasible solution if positive ($y_i = 1$) and negative ($y_i = -1$) classes cannot be separated by any hyperplanes. Additionally, even if two classes are separable, a better hyperplane may be obtained by taking training instances near to the hyperplane as support vectors. In these cases, we take

errors of training instances into account.

$$\min_{w,b} \quad \frac{1}{2}\|w\|^2 + \frac{\mu^2}{2}\sum_{i\in M} l_i(w,b), \tag{6}$$

where $l_i(w,b) = (\max\{0, 1 - y_i(w^\top x_i + b)\})^2$ is the squared hinge loss function. This learning model, which tolerates errors, is called soft-margin SVM. On the other hand, the model (5) is called hard-margin SVM. μ is a hyperparameter to adjust the effect of the sum of losses. It is equivalent to,

$$\begin{aligned} \min_{w,b,\xi} \quad & \frac{1}{2}\|w\|^2 + \frac{\mu^2}{2}\sum_{i\in M}\xi_i^2 \\ \text{s.t.} \quad & y_i(w^\top x_i + b) + \xi_i \geq 1,\ i \in M, \end{aligned} \tag{7}$$

where $\xi = (\xi_1, \ldots, \xi_m) \in \mathbf{R}^m$ is the vector of additional decision variables.

2.3 Multiobjective Multiclass SVM

In this paper, we study multiobjective multiclass SVM (MMSVM) [13], which is an application of the margin maximization to multiclass classification. In Figure 1, we show an example of 3-class linear classification in a 2-dimensional input space. As shown in the figure, the margin of the boundary of class-pair $pq \in C^{\bar{2}}$ is defined by

$$d_{pq}(w,b) = \frac{\min_{i\in M_p\cup M_q} |(w_p - w_q)^\top x_i + b_p - b_q|}{\|w_p - w_q\|}. \tag{8}$$

In contrast to the binary case, there are more than two margins in the multiclass linear classifier. Hence, in MMSVM we define a learning model by the following multiobjective optimization problem.

$$\begin{aligned} \max_{w,b,r} \quad & g_{12}(w,r), \ldots, g_{1c}(w,r), \\ & \qquad g_{23}(w,r), \ldots, g_{(c-1)c}(w,r) \\ \text{s.t.} \quad & (w_p - w_q)^\top x_i + b_p - b_q \geq r_{pq}, i \in M_p, pq \in C^{\bar{2}}, \\ & (w_q - w_p)^\top x_i + b_q - b_p \geq r_{pq}, i \in M_q, pq \in C^{\bar{2}}, \\ & r_{pq} \geq 1,\ pq \in C^{\bar{2}}, \end{aligned} \tag{M}$$

where

$$g_{pq}(w,r) = \frac{r_{pq}}{\|w_p - w_q\|}, \tag{9}$$

and $r = (r_{12}, \ldots, r_{1c}, r_{23}, \ldots, r_{(c-1)c}) \in \mathbf{R}^{(c-1)c/2}$. It is shown that $d_{pq}(w^*, b^*)$ and $g_{pq}(w^*, r^*)$ coincide for all $pq \in C^{\bar{2}}$ at any Pareto optimal solution (w^*, b^*, s^*) [13] of (M). Note that we cannot fix all of r_{pq}, $pq \in C^{\bar{2}}$ to a constant.

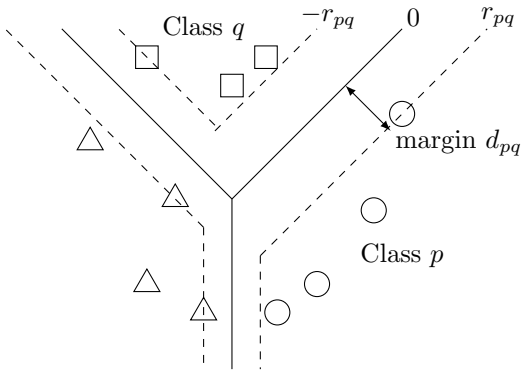


Figure 1: Linear discriminant for 3-class problem. The lines are the boundaries with respect to class-pairs, i.e., $\{x \in \mathbf{R}^n \mid (w_p - w_q)^\top x + (b_p - b_q) = 0\}$, and the two dashed lines parallel with each line indicate the margin region, i.e., $\{x \in \mathbf{R}^n \mid |(w_p - w_q)^\top x + (b_p - b_q)| \leq r_{pq}\}$.

Although the problem (M) is nonconvex because of the objective functions, one of the authors [13] have proposed efficient solution methods using the ϵ-constraint method, which is a popular scalarization method for multiobjective optimization [6]. Using this scalarization, it is reduced to a second-order cone programming (SOCP) and easily dealt with by several numerical solvers. Recently, the authors [8, 9] have proposed a convex approximation of (M), and solve it using the reference point method. The details of the latter method are shown in the next section.

Next, we introduce soft-margin model of MMSVM. In the above binary case, the error of (x_i, y_i) is given by the squared hinge loss function $l_i(w,b)$. In other words, it is the squared ratio of $\frac{1-y_i(w^\top x_i+b)}{\|w\|}$ to the margin $\frac{1}{\|w\|}$. We extend it to the multiclass case. Thus, we define the error of (x_i, y_i) with $y_i = p$ by the ratio of $\frac{r_{pq}-((w_p-w_q)^\top x_i+b_p-b_q)}{\|w_p-w_q\|}$ to $\frac{r_{pq}}{\|w_p-w_q\|}$, i.e.,

$$\begin{aligned} & l_i^{pq}(w,b,r) \\ & = \left(\max\left\{0, \frac{r_{pq} - ((w_p - w_q)^\top x_i + b_p - b_q)}{r_{pq}}\right\}\right)^2. \end{aligned} \tag{10}$$

Using this loss function, we define soft-margin MMSVM as the following multiobjective optimization problem.

$$\begin{aligned} \min_{w,b,r,\xi} \quad & h_{12}(w,r,\xi), \ldots, h_{(c-1)c}(w,r,\xi) \\ \text{s.t.} \quad & (w_p - w_q)^\top x_i + b_p - b_q + \xi_{qi} \\ & \qquad \geq r_{pq},\ i \in M_p,\ pq \in C^{\bar{2}}, \\ & (w_q - w_p)^\top x_i + b_q - b_p + \xi_{pi} \\ & \qquad \geq r_{pq},\ i \in M_q,\ pq \in C^{\bar{2}}, \\ & r_{pq} \geq 1,\ pq \in C^{\bar{2}}, \end{aligned} \tag{SM}$$

where

$$h_{pq}(w,r,\xi) = \frac{1}{2}\left(\frac{\|w_p - w_q\|}{r_{pq}}\right)^2 + \frac{\mu^2}{2}\left(\sum_{i\in M_p}\left(\frac{\xi_{qi}}{r_{pq}}\right)^2 + \sum_{i\in M_q}\left(\frac{\xi_{pi}}{r_{pq}}\right)^2\right), \quad (11)$$

and $\xi = ((\xi_{1i})_{i\notin M_1}, \ldots, (\xi_{ci})_{i\notin M_c}) \in \mathbf{R}^{(c-1)m}$. This soft-margin model minimizes the inverse-squared margins instead of maximizing the margins. It is parallel to the binary case of (7).

2.4 Kernel Method

In the case that linear classifier (2) is not suitable to classification problems under consideration, we apply kernel trick to MMSVM. The key idea of the kernel trick is manipulating vectors in a high dimensional feature space using only a (positive definite) kernel function $\kappa : \mathbf{R}^n \times \mathbf{R}^n \to \mathbf{R}$. For any kernel κ, there is a mapping $\phi : \mathbf{R}^n \to \mathcal{H}_\kappa$ such that $\kappa(x,x') = \langle\phi(x), \phi(x')\rangle_{\mathcal{H}_\kappa}$ for all $x, x' \in \mathbf{R}^n$, where $\langle\cdot,\cdot\rangle_{\mathcal{H}_\kappa}$ is the inner product of reproducing kernel Hilbert space $\mathcal{H}_\kappa$ associated with κ [10]. The mapping can be given by $x \mapsto \phi(x) = \kappa(x,\cdot) \in \mathcal{H}_\kappa$. We reverse the above derivation. That is, we can compute the inner product $\langle\phi(x), \phi(x')\rangle$ using the kernel function $\kappa(x,x')$, even if the Hilbert space $\mathcal{H}_\kappa$ is infinite dimensional.

Consider the MMSVM problem (SM) for training set $(\phi(x_1), y_1), \ldots, (\phi(x_m), y_m)$ in the feature space. Without loss of generality, we can restrict variables $w_1, \ldots, w_c$ of (SM) to the finite subspace spanned by $\phi(x_1), \ldots, \phi(x_m)$. Then, the norm and the inner product appearing in the objective functions and constraints, respectively, are expressed by the kernel function. Moreover, we can obtain a coordinate system of the subspace by kernel principal component analysis (KPCA). By KPCA, a feature vector $\phi(x)$ for $x \in \mathbf{R}^n$ is expressed as a set of coordinates, in which each coordinate is obtained by the inner product of a principal axis and $\phi(x)$. Again, the inner production can be calculated using only the kernel function. Furthermore, we can reduce the computation of MMSVM by removing principal axes with small eigenvalues.

To sum up, even if MMSVM is performed in a feature space, the corresponding optimization problem can be reduced to the form of (SM).

In this paper, we use the radial basis function (RBF) kernel. For input vectors $x, x' \in \mathbf{R}^n$, the value $\kappa(x,x')$ of the RBF kernel is defined by

$$\kappa(x,x') = \exp\left(-\frac{\|x-x'\|^2}{2\sigma^2}\right), \quad (12)$$

where σ is a parameter for scaling input vectors. If σ is large, transformed input vectors are concentrated in similar feature vectors. On the other hand, if σ is small, transformed input vectors are orthogonal to each other in the feature space.

3 Approximate MMSVM and Scalarization Using Large-k Norm

3.1 Approximate MMSVM

First, we further reformulate (SM) by replacing r_{pq}^2 with s_{pq}.

$$\begin{aligned}
\min_{w,b,s,\xi}\ & h_{12}(w,s,\xi), \ldots, h_{(c-1)c}(w,s,\xi) \\
\text{s.t.}\ & (w_p - w_q)^\top x_i + b_p - b_q + \xi_{qi} \\
& \qquad \geq \sqrt{s_{pq}},\ i \in M_p,\ pq \in C^{\bar{2}}, \\
& (w_q - w_p)^\top x_i + b_q - b_p + \xi_{pi} \\
& \qquad \geq \sqrt{s_{pq}},\ i \in M_q,\ pq \in C^{\bar{2}}, \\
& s_{pq} \geq 1,\ pq \in C^{\bar{2}},
\end{aligned} \quad \text{(SM)}$$

and

$$h_{pq}(w,s,\xi) = \frac{1}{2}\frac{\|w_p - w_q\|^2}{s_{pq}} + \frac{\mu^2}{2}\frac{\sum_{i\in M_p}\xi_{qi}^2 + \sum_{i\in M_q}\xi_{pi}^2}{s_{pq}}, \quad (13)$$

where $s = (s_{12}, \ldots, s_{1c}, s_{23}, \ldots, s_{(c-1)c}) \in \mathbf{R}^{(c-1)c/2}$. This multiobjective optimization problem is still nonconvex, because of $\sqrt{s_{pq}}$ in the right hand sides of the first and second constraints. Hence, we replace $\sqrt{s_{pq}}$ with an affine function of s_{pq}, and make (SM) convex.

Let ρ be a positive value. We replace $\sqrt{s_{pq}}$ with $\frac{s_{pq}+\rho}{1+\rho}$, and put additional constraints $s_{pq} \leq \rho^2$ for $pq \in C^{\bar{2}}$. Then, we obtain the following problem.

$$\begin{aligned}
\min_{w,b,s,\xi}\ & h_{12}(w,s,\xi), \ldots, h_{(c-1)c}(w,s,\xi) \\
\text{s.t.}\ & (w_p - w_q)^\top x_i + b_p - b_q \\
& \qquad \geq \frac{s_{pq}+\rho}{1+\rho},\ i \in M_p, pq \in C^{\bar{2}}, \\
& (w_q - w_p)^\top x_i + b_q - b_p \\
& \qquad \geq \frac{s_{pq}+\rho}{1+\rho},\ i \in M_q, pq \in C^{\bar{2}}, \\
& 1 \leq s_{pq} \leq \rho^2,\ pq \in C^{\bar{2}}.
\end{aligned} \quad \text{(ASM)}$$

Figure 2 shows the relation between $\sqrt{s}$ and $(s+\rho)/(1+\rho)$. In the section of $1 \leq s \leq \rho^2$, it holds that $\sqrt{s} \geq (s+\rho)/(1+\rho)$. Roughly speaking, the Pareto optimal set of (ASM) is a lower bound of that

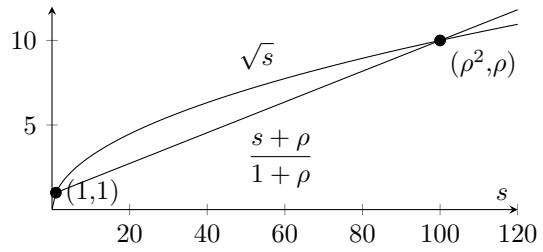


Figure 2: Approximation of $\sqrt{s}$ by $(s+\rho)/(1+\rho)$ (when $\rho = 10$).

of (SM), when ρ is sufficiently large. We explain it more precisely. Let $z^\star = (w^\star, b^\star, s^\star, \xi^\star)$ be a Pareto optimal solution of (SM). Let $s^\star_{\min} = \min_{pq\in C^{\bar{2}}} s^\star_{pq}$. Assume $s^\star_{pq}/s^\star_{\min} \le \rho^2$ for all $pq \in C^{\bar{2}}$. Then, we define another solution $z' = (w', b', s', \xi')$ of (ASM) by $w' = w^\star/\sqrt{s^\star_{\min}}$, $b' = b^\star/\sqrt{s^\star_{\min}}$, $\xi' = \xi^\star/\sqrt{s^\star_{\min}}$ and $s'_{pq} = (1+\rho)\sqrt{s^\star_{pq}/s^\star_{\min}} - \rho$ for $pq \in C^{\bar{2}}$. By the assumptions and the definition, z' is feasible for (ASM) and dominating $z^\star$, i.e., $h_{pq}(w', s', \xi') \le h_{pq}(w^\star, s^\star, \xi^\star)$ for all $pq \in C^{\bar{2}}$. Hence, there is a Pareto optimal solution of (ASM) which dominates $z^\star$. In other words, for any Pareto optimal solution $z^\star$ of (SM), there is an Pareto optimal solution of (ASM) which dominates it, unless $s^\star_{pq}/s^\star_{\min} > \rho^2$ for some $pq \in C^{\bar{2}}$.

On the other hand, from a Pareto optimal solution $z^* = (w^*, b^*, s^*, \xi^*)$ of (ASM), we can construct a feasible solution of (SM) as follows. For each $pq \in C^{\bar{2}}$, we define

$$s'_{pq} = \left(\frac{s^*_{pq}+\rho}{1+\rho}\right)^2. \tag{14}$$

Then, (w^*, b^*, s', ξ^*) is feasible for (SM). Moreover, for each $pq \in C^{\bar{2}}$, we have the following relation between $h_{pq}(w^*, s^*, \xi^*)$ and $h_{pq}(w^*, s', \xi)$:

$$\frac{h_{pq}(w^*, s', \xi^*)}{h_{pq}(w^*, s^*, \xi^*)} \in [1, \theta(\rho)], \tag{15}$$

where

$$\theta(\rho) = \frac{(1+\rho)^2}{4\rho}. \tag{16}$$

The θ monotonically increases with respect to ρ. $\theta(\rho)$ is approximated by $\rho/4 + 1/2$, i.e., the upper bound of the ratio of approximation deteriorates linearly with respect to ρ. On the other hand, the range of s_{pq} in (ASM) increases quadratically.

3.2 Large-k Norm Scalarization

A popular approach to solve multiobjective optimization problems is scalarization, in which the objective functions are reduced to a single objective function. In [13], the ϵ-constraint scalarization method is applied to MMSVM. On the other hand, In [9], the reference point scalarization method based on the augmented Tchebyshev function is applied to approximated MMSVM.

In this paper, we propose scalarization using the large-k norm (or D-norm) [2, 7]. Let $k \in [1, l]$ be an integer. The large-k norm of $t \in \mathbf{R}^l$, denoted by $|||t|||_k$, is defined as follows:

$$|||t|||_k = \sum_{i=1}^{k} |t_{(i)}|, \tag{17}$$

where $|t_{(1)}| \ge |t_{(2)}| \ge \cdots \ge |t_{(l)}|$. Remark that when $k = 1$ it is the Tchebyshev norm (ℓ_∞-norm) and when $k = l$ it is the ℓ_1-norm.

The large-k norm is expressed as follows.

$$|||t|||_k = \sup_s \left\{ \sum_{i=1}^{l} |t_i| s_i \;\middle|\; \sum_{i=1}^{l} s_i = k,\ s \in [0,1]^l \right\}. \tag{18}$$

Considering its dual problem, we have

$$|||t|||_k = \inf_u \left\{ ku + \sum_{i=1}^{l} \max\{0, |t_i| - u\} \right\}. \tag{19}$$

We apply the large-k norm scalarization to approximate MMSVM (ASM). Let $k = 1, \ldots, (c-1)c/2$. Let $h(w, s, \xi) = (h_{12}(w, s, \xi), \ldots, h_{(c-1)c}(w, s, \xi)) \in \mathbf{R}^{(c-1)c/2}$. The objective function of the proposed scalarization is $|||h(w, s, \xi)|||_k$. By the expression of (19), it is obtained as follows.

$$\begin{aligned}
\min_{w,b,\xi,s,u}\quad & ku + \sum_{pq\in C^{\bar{2}}} \max\{0, h_{pq}(w, s, \xi) - u\} \\
\text{s.t.}\quad & (w^p - w^q)^\top x^i + b^p - b^q + \xi_{qi} \\
& \qquad \ge \frac{s_{pq}+\rho}{1+\rho},\ pq \in C^{\bar{2}},\ i \in M^p, \\
& (w^q - w^p)^\top x^i + b^q - b^p + \xi_{pi} \\
& \qquad \ge \frac{s_{pq}+\rho}{1+\rho},\ pq \in C^{\bar{2}},\ i \in M^q, \\
& 1 \le s_{pq} \le \rho^2,\ pq \in C^{\bar{2}}.
\end{aligned} \tag{ASMkN}$$

Remark that $h_{pq}(w, s, \xi) \ge 0$ necessarily holds. Replacing $h_{pq}(w, s, \xi)$ with (13), you can see that this problem is an SOCP. Hence, it is easily solved by numerical solvers.

When $k = (c-1)c/2$, the problem (ASMkN) is reduced to minimization of the sum of $h_{12}(w, s, \xi), \ldots, h_{(c-1)c}(w, s, \xi)$, which has been

studied in [8]. Moreover, when $k = (c-1)c/2$ and $\rho = 1$, it is the conventional ATSVM [4, 15]. In other words, the proposed model is a generalization of those multiclass SVMs.

We shortly discuss a relation between optimal solutions of (ASMkN) and Pareto optimal solutions of (ASM). The large-k norms is strictly increasing, namely for $t_{pq} < t'_{pq}$ for all $pq \in C^{\bar{2}}$ we have $|||t|||_k < |||t'|||_k$. Therefore, as shown in [6], if $z^* = (w^*, b^*, \xi^*, s^*, u^*)$ is optimal for (ASMkN) then z^* is weakly efficient for (ASM), i.e., there is no feasible solution $z = (w, b, \xi, s)$ of (ASM) such that $h_{pq}(z) < h_{pq}(z^*)$ for all $pq \in C^{\bar{2}}$.

4 Numerical Experiments

To examine performance of the scalarization of approximate MMSVM based on the large-k norm, we execute numerical experiments using 8 benchmark data sets in UCI Machine Learning Repository [5]. We compared classification ability of classifiers obtained by (ASMkN) in different values of parameters ρ and k. To solve the optimization problem, we used software package MOSEK [11]. Classification errors of classifiers was measured by 10 times 10-fold cross-validation with balancing class distribution.

In order to obtain nonlinear classification, we performed MMSVM in the feature space associated with the RBF kernel (12). Furthermore, feature vectors were projected to a 200-dimensional space by KPCA. Remember that the RBF kernel includes parameter σ.

The penalty parameter μ was fixed to 1000. The approximation parameter ρ was varied in $1, 1 + 2^{-2}, \ldots, 1+2^6$. For each data set, the kernel parameter σ and norm parameter k were varied as shown in Table 1. The ranges of parameters σ was selected so that ATSVM ($\rho = 1$ and $k = \frac{(c-1)c}{2}$) can attain the best classification performance. We also show the numbers of class labels and their pairs for each benchmark data set in Table 1.

Table 2 shows classification errors measured in the numerical experiments. Each row shows the result of the data set whose name is described in the first column. The columns "err" show the mean of errors measured by 10 times cross-validation, followed by the standard deviation. For each data set and each method, the error is the result of the best set of parameters (σ for ATSVM, (σ, ρ, k) for MMSVM), which is also shown in the table. For all data sets, the proposed MMSVM outperforms the conventional ATSVM. Moreover, for all data sets, the best ρ is more than 1 and the best k is less than $\frac{(c-1)c}{2}$. It indicates that both of maximum-margin and multiobjective approaches are meaningful for multiclass SVM learning.

data	c	$\frac{(c-1)c}{2}$	$\log_2(\sigma)$			k
iris	3	3	8	9	10	$1, 2, 3$
wine	3	3	10	11	12	$1, 2, 3$
zoo	7	21	9	10	11	$12, \ldots, 21$
balance	3	3	3	4	5	$1, 2, 3$
forest	4	6	11	12	13	$1, \ldots, 6$
car	4	6	1	2	3	$1, \ldots, 6$
dermatology	6	15	10	11	12	$6, \ldots, 15$
vehicle	4	6	4	5	6	$1, \ldots, 6$

Table 1: Numbers of classes (c) and class-pairs ($\frac{(c-1)c}{2}$). Ranges of kernel parameter (σ) and scalarization parameter (k).

In order to see the effect of the parameters σ, ρ and k, we show a 3-dimensional array of errors with respect to the parameters for Wine Data Set in Figure 3. In the array of the largest σ (Figure 3(c)), better results are obtained in larger ρ. It indicates that the maximum-margin approach is effective in lower dimensional feature spaces. We obtain better results at $k = 1$, which is regarded as maximizing of minimum class-pair margin, for the largest σ and larger ρ.

Finally, we show actual classification situation for Wine Data Set in Figure 4. In these figures, all instances were used for training. Parameter σ was set to 2^{12}. The values of "obj. value" in the figure are the class-pair objective values of feasible solutions of (SM) constructed from optimal solutions of (ASMkN). When $\rho = 1 + 2^6$ (Figures 4(c) and 4(d)), instances near the boundary of class-pair 1 and 2 are successfully classified. On the other hand, when $\rho = 1$ (Figures 4(a) and 4(b)), two or three of them are missed. The objective value of class-pair 1 and 2 for $k = 1$ and $\rho = 1 + 2^6$ is smaller than that for $k = 3$ and $\rho = 1 + 2^6$. It may be a reason that we obtain better results at $k = 1$ as shown in Figure 3(c). In Figure 4(a), a classifier of large margin is obtained[1]. It may be because the objective value of class-pair 1 and 3 were preferentially optimized.

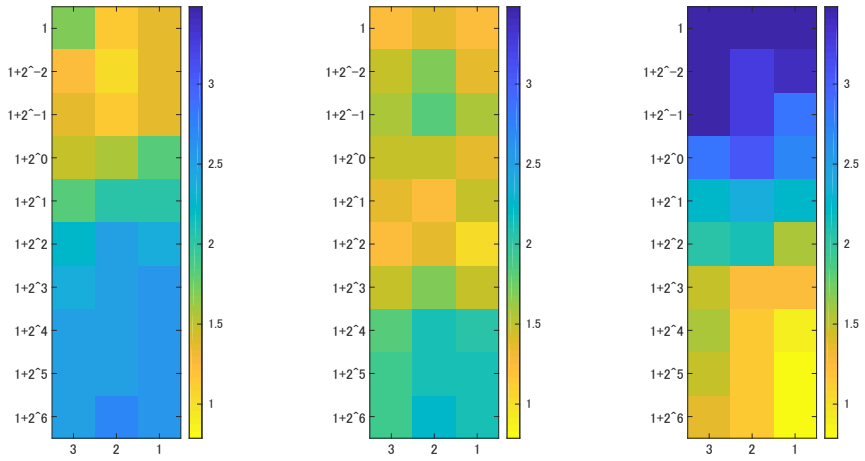
5 Concluding Remarks

In this paper, we have proposed the large-k norm scalarization for the multiobjective multiclass support vector machine (MMSVM). The large-k norm provides another spectrum between the Tchebyshev (ℓ_∞) norm and the ℓ_1 norm instead of the augmented Tchebyshev norm. In numerical experiments, the large-k-norm scalarization of approximate MMSVM outper-

[1] We remark that for (a) and (b) the optimality was not ensured by MOSEK.

data	ATSVM		MMSVM			
	err (%)	σ	err (%)	σ	ρ	k
iris	3.07 ± 0.37	2^8	2.40 ± 0.89	2^{10}	$1+2^6$	1
wine	1.69 ± 0.40	2^{10}	0.79 ± 0.31	2^{12}	$1+2^6$	1
zoo	3.56 ± 0.54	2^{10}	2.97 ± 0.00	2^{11}	$1+2^1$	12
balance	0.42 ± 0.14	2^3	0.06 ± 0.14	2^3	$1+2^6$	2
forest	8.64 ± 0.68	2^{11}	8.03 ± 0.38	2^{13}	$1+2^5$	5
car	0.52 ± 0.17	2^1	0.26 ± 0.13	2^2	$1+2^1$	5
dermatology	2.13 ± 0.12	2^{10}	1.91 ± 0.00	2^{12}	$1+2^6$	6
vehicle	14.78 ± 0.33	2^4	14.47 ± 0.32	2^4	$1+2^{-1}$	3

Table 2: Numerical Experiments



(a) $\sigma = 2^{10}$ (b) $\sigma = 2^{11}$ (c) $\sigma = 2^{12}$

Figure 3: 3-dimensional array of classification errors with respect to σ (in different figures), ρ (rows) and k (columns) for Wine Data Set.

forms the conventional ATSVM in classification capability for all benchmark data sets, if we can select the best set of parameters. In the future work, we will examine that the proposed scalarization achieves good classification capability in automatic parameter selection.

References

[1] S. Abe, Support Vector Machines for Pattern Classification (Advances in Pattern Recognition), Springer-Verlag, Berlin, Heidelberg, 2005.

[2] D. Bertsimas, D. Pachamanova, M. Sim, Robust linear optimization under general norms, Operations Research Letters 32 (6) (2004) 510–516.

[3] C. Cortes, V. N. Vapnik, Support-vector networks, Machine Learning 20 (3) (1995) 273–297.

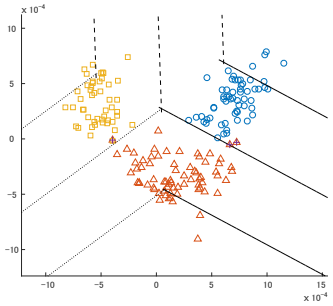
[4] Ü. Doğan, T. Glasmachers, C. Igel, A unified view on multi-class support vector classification, Journal of Machine Learning Research 17 (45) (2016) 1–32.

[5] D. Dua, C. Graff, UCI machine learning repository (2017).
URL `http://archive.ics.uci.edu/ml`

[6] M. Ehrgott, Multicriteria Optimization, Springer-Verlag, Berlin, Heidelberg, 2005.

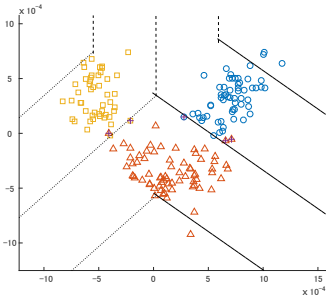
[7] J.-Y. Gotoh, S. Uryasev, Two pairs of families of polyhedral norms versus ℓ_p-norms: Proximity and applications in optimization, Math. Program. 156 (1-2) (2016) 391–431.
URL `http://dx.doi.org/10.1007/s10107-015-0899-9`

[8] Y. Kusunoki, K. Tatsumi, A multi-class support vector machine based on geometric margin maximization, in: V.-N. Huynh, M. Inuiguchi, D. H. Tran, T. Denoeux (Eds.), Integrated Uncertainty in Knowledge Modelling and Decision Making,



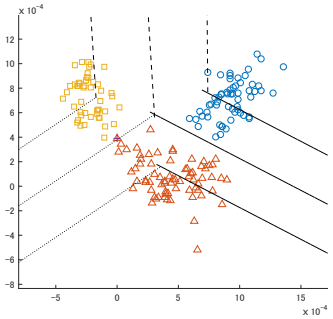
class-pair	12	13	23
margin $\times 10^{-4}$	6.343	5.786	6.047
obj. value$\times 10^{+6}$	14.295	2.788	9.795

(a) $k = 3$, $\rho = 1$, $\sigma = 2^{12}$ (ATSVM)



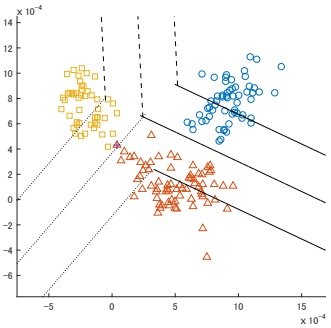
class-pair	12	13	23
margin $\times 10^{-4}$	7.448	5.721	6.803
obj. value $\times 10^{+6}$	17.305	2.866	11.471

(b) $k = 1$, $\rho = 1$, $\sigma = 2^{12}$



class-pair	12	13	23
margin $\times 10^{-4}$	3.547	4.805	3.818
obj. value $\times 10^{+6}$	8.363	2.687	7.268

(c) $k = 3$, $\rho = 1 + 2^6$, $\sigma = 2^{12}$



class-pair	12	13	23
margin $\times 10^{-4}$	3.399	2.880	3.256
obj. value $\times 10^{+6}$	8.287	6.101	7.487

(d) $k = 1$, $\rho = 1 + 2^6$, $\sigma = 2^{12}$

Figure 4: Separating lines obtained by MMSVM for Wine Data Set. There are three classes — class 1: blue circles, class 2: orange triangles, class 3: yellow squares. The solid lines are the separating line with the margins of class-pair 12. The broken lines are of class-pair 13. The dotted lines are of class-pair 23. The data are plotted in the 2-dimensional affine subspace passing through 3 normal vectors of 3 classes.

Springer International Publishing, Cham, 2018, pp. 101–113.

[9] Y. Matsugi, T. Sugimoto, Y. Qi, Y. Kusunoki, K. Tatsumi, Approximate multiobjective multi-class svm by using the reference point method, in: 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2018, pp. 3535–3540.

[10] M. Mohri, A. Rostamizadeh, A. Talwalkar, Foundations of Machine Learning, 2nd Edition, The MIT Press, 2018.

[11] MOSEK ApS, The MOSEK optimization toolbox for MATLAB manual. Version 8.1. (2017). URL `http://docs.mosek.com/8.1/toolbox/index.html`

[12] R. Rifkin, A. Klautau, In defense of one-vs-all classification, J. Mach. Learn. Res. 5 (2004) 101–141.

[13] K. Tatsumi, T. Tanino, Support vector machines maximizing geometric margins for multi-class classification, TOP 22 (3) (2014) 815–840.

[14] V. N. Vapnik, Statistical Learning Theory, A Wiley-Interscience Publication, New York, 1998.

[15] J. Weston, C. Watkins, Support vector machines for multi-class pattern recognition, in: ESANN, 1999, pp. 219–224.