ROW AND COLUMN GENERATION ALGORITHM FOR MINIMUM MARGIN MAXIMIZATION OF RANKING PROBLEMS

YOICHI IZUNAGA, KEISUKE SATO, KEIJI TATSUMI, AND YOSHITSUGU YAMAMOTO

ABSTRACT. We consider the ranking problem of learning a ranking function from the data set of objects each of which is endowed with an attribute vector and a ranking label chosen from the ordered set of labels. We propose a primal formulation with dual representation of normal vector, and then propose to apply the kernel technique to the formulation. We also propose algorithms based on the row and column generation in order to mitigate the computational burden due to the large number of objects.

1. INTRODUCTION

This paper is concerned with a multi-class classification problem of n objects, each of which is endowed with an m-dimensional *attribute vector* $\boldsymbol{x}^i = (x_1^i, x_2^i, \ldots, x_m^i)^\top \in \mathbb{R}^m$ and a *label* ℓ_i. The underlying statistical model assumes that object i receives label k, i.e., $\ell_i = k$, when the latent variable y_i determined by

$$y_i = \boldsymbol{w}^\top \boldsymbol{x}^i + \varepsilon^i = \sum_{j=1}^{m} w_j x_j^i + \varepsilon^i$$

falls between two thresholds p_k and p_{k+1}, where ε^i represents a random noise whose probabilistic property is not known. Namely, attribute vectors of objects are loosely separated by hyperplanes $H(\boldsymbol{w}, p_k) = \{\, \boldsymbol{x} \in \mathbb{R}^m \mid \boldsymbol{w}^\top \boldsymbol{x} = p_k \,\}$ for $k = 1, 2, \ldots, l$ which share a common normal vector $\boldsymbol{w}$, then each object is given a label according to the layer it is located in. Note that neither y_i's, w_j's nor p_k's are observable. Our problem is to find the normal vector $\boldsymbol{w} \in \mathbb{R}^m$ as well as the thresholds $p_1, p_2, \ldots, p_l$ that best fit the input data $\{\, (\boldsymbol{x}^i, \ell_i) \mid i \in N \,\}$.

This problem is known as the *ranking problem* and frequently arises in social sciences and operations research. See, for instance Crammer and Singer [2], Herbrich *et al.* [3], Liu [4], Shashua and Levin [6] and Chapter 8 of Shawe-Taylor and Cristianini [8]. It is a variation of the multi-class classification problem, for which several learning algorithms of the *support vector machine* (*SVM* for short) have been proposed. We refer the reader to Chapters 4.1.2 and 7.1.3 of Bishop [1], Chapter 10.10 of Vapnik [10] and Tatsumi *et al.* [9] and references therein. What distinguishes the problem from other multi-class classification problems is that the identical normal vector should be shared by all the separating hyperplanes. In this paper based on the formulation *fixed margin strategy* by Shashua and Levin [6], we propose a row and column generation algorithm to maximize the minimum margin for the ranking problems.

Key words and phrases. Ranking problem, Support vector machine, Multi-class classification, Kernel technique, Dual representation.

The last author is partly supported by The Okawa Foundation for Information and Telecommunications.

This paper is organized as follows. We give some definitions and notations in Section 2. In Section 3, we formulate the maximization of minimum margin with the hard margin constraints and apply the dual representation of the normal vector to the formulation. In Section 4, we propose a row and column generation algorithm and prove the validity of the algorithm. In Section 5, after reviewing the kernel technique, we apply the kernel technique to the hard margin problem with the dual representation. In Section 6, 7 and 8, we expand the discussions so far into the soft margin problem. In Section 9, we demonstrate an illustrative example for a small instance, then report computational experiments of our algorithm in Section 10. In Section 11, we discuss a desirable property of the separating curves.

2. Definitions and notation

Throughout the paper $N = \{1, 2, \ldots, i, \ldots, n\}$ denotes the set of n objects and $\boldsymbol{x}^i = (x^i_1, x^i_2, \ldots, x^i_m)^\top \in \mathbb{R}^m$ denotes the attribute vector of object i. The predetermined set of labels is $L = \{0, 1, \ldots, k, \ldots, l\}$ and the label assigned to object i is denoted by ℓ_i. Let $N(k) = \{\, i \in N \mid \ell_i = k \,\}$ be the set of objects with label $k \in L$, and for notational convenience we write $n(k) = |N(k)|$ for $k \in L$, and $N(k..k') = N(k) \cup N(k+1) \cup \cdots \cup N(k')$ for $k, k' \in L$ such that $k < k'$. For a succinct notation we define

$$\begin{aligned} X &= \begin{bmatrix} \cdots & \boldsymbol{x}^i & \cdots \end{bmatrix}_{i \in N} \in \mathbb{R}^{m \times n} \\ X_W &= \begin{bmatrix} \cdots & \boldsymbol{x}^i & \cdots \end{bmatrix}_{i \in W} \in \mathbb{R}^{m \times |W|} \end{aligned} \tag{2.1}$$

for $W \subseteq N$, and the corresponding Gram matrices

$$\begin{aligned} K &= X^\top X \in \mathbb{R}^{n \times n}, \\ K_W &= X_W^\top X_W \in \mathbb{R}^{|W| \times |W|}. \end{aligned}$$

We denote the k-dimensional zero vector and vector of 1's by $\boldsymbol{0}_k$ and $\boldsymbol{1}_k$, respectively. Given a subset $W \subseteq N$ and a vector $\boldsymbol{\alpha} = (\alpha_i)_{i \in W}$ we use the notation $(\boldsymbol{\alpha}_W, \boldsymbol{0}_{N \setminus W})$ to denote the n-dimensional vector $\bar{\boldsymbol{\alpha}}$ such that

$$\bar{\alpha}_i = \begin{cases} \alpha_i & \text{when } i \in W \\ 0 & \text{otherwise.} \end{cases}$$

3. Hard Margin Problems for Separable Case

3.1. Primal Hard Margin Problem. Henceforth we assume that $N(k) \neq \emptyset$ for all $k \in L$ for the sake of simplicity, and adopt the notational convention that $p_0 = -\infty$ and $p_{l+1} = +\infty$. We say that an instance $\{\, (\boldsymbol{x}^i, \ell_i) \mid i \in N \,\}$ is *separable* if there exist $\boldsymbol{w} \in \mathbb{R}^m$ and $\boldsymbol{p} = (p_1, p_2, \ldots, p_l)^\top \in \mathbb{R}^l$ such that

$$p_{\ell_i} < \boldsymbol{w}^\top \boldsymbol{x}^i < p_{\ell_i + 1} \quad \text{for } i \in N.$$

Clearly an instance is separable if and only if there are $\boldsymbol{w}$ and $\boldsymbol{p}$ such that

$$p_{\ell_i} + 1 \leq \boldsymbol{w}^\top \boldsymbol{x}^i \leq p_{\ell_i + 1} - 1 \quad \text{for } i \in N.$$

For each $k \in L \setminus \{0\}$ we see that

$$\max_{i \in N(k-1)} \boldsymbol{w}^\top \boldsymbol{x}^i \le p_k - 1 < p_k < p_k + 1 \le \min_{j \in N(k)} \boldsymbol{w}^\top \boldsymbol{x}^j,$$

implying

$$\min_{j \in N(k)} \frac{\boldsymbol{w}^\top}{\|\boldsymbol{w}\|} \boldsymbol{x}^j - \max_{i \in N(k-1)} \frac{\boldsymbol{w}^\top}{\|\boldsymbol{w}\|} \boldsymbol{x}^i \ge \frac{2}{\|\boldsymbol{w}\|}.$$

Then the margin between $\{ \boldsymbol{x}^i \mid i \in N(k-1) \}$ and $\{ \boldsymbol{x}^j \mid j \in N(k) \}$ is at least $2/\|\boldsymbol{w}\|$. Hence the maximization of the minimum margin is formulated as the quadratic programming

$$(H) \quad \left| \begin{array}{ll} \text{minimize} & \|\boldsymbol{w}\|^2 \\ \text{subject to} & p_{\ell_i} + 1 \le (\boldsymbol{x}^i)^\top \boldsymbol{w} \le p_{\ell_i+1} - 1 \qquad \text{for } i \in N, \end{array} \right.$$

or more explicitly with the notation introduced in Section 2

$$(H) \quad \left| \begin{array}{lll} \text{minimize} & \|\boldsymbol{w}\|^2 & \\ \text{subject to} & 1 - (\boldsymbol{x}^i)^\top \boldsymbol{w} + p_{\ell_i} \le 0 & \text{for } i \in N(1..l) \\ & 1 + (\boldsymbol{x}^i)^\top \boldsymbol{w} - p_{\ell_i+1} \le 0 & \text{for } i \in N(0..l-1). \end{array} \right.$$

The constraints therein are called *hard margin* constraints, and we name this problem (H).

3.2. **Dual Representation.** A close look at the primal problem (H) shows that the following property holds for the optimum solution $\boldsymbol{w}^*$. See, for example Chapter 6 of Bishop [1], Shashua and Levin [6] and Theorem 1 in Schölkopf *et al.* [7].

Lemma 3.1. *Let $(\boldsymbol{w}^*, \boldsymbol{p}^*) \in \mathbb{R}^{m+l}$ be an optimum solution of (H). Then $\boldsymbol{w}^* \in \mathbb{R}^m$ lies in the range space of $X$, i.e., $\boldsymbol{w}^* = X\boldsymbol{\lambda}$ for some $\boldsymbol{\lambda} \in \mathbb{R}^n$.*

Proof. Let $\boldsymbol{w}_1$ be the orthogonal projection of $\boldsymbol{w}^*$ onto the range space of $X$ and let $\boldsymbol{w}_2 = \boldsymbol{w}^* - \boldsymbol{w}_1$. Then we obtain

$$(\boldsymbol{x}^i)^\top \boldsymbol{w}^* = (\boldsymbol{x}^i)^\top (\boldsymbol{w}_1 + \boldsymbol{w}_2) = (\boldsymbol{x}^i)^\top \boldsymbol{w}_1 \quad \text{for } i \in N,$$

meaning $(\boldsymbol{w}_1, \boldsymbol{p}^*)$ is feasible to (H), and

$$\|\boldsymbol{w}^*\|^2 = \|\boldsymbol{w}_1\|^2 + \|\boldsymbol{w}_2\|^2 \ge \|\boldsymbol{w}_1\|^2.$$

Hence by the optimality of $\boldsymbol{w}^*$ we conclude that $\boldsymbol{w}_2 = \boldsymbol{0}$. □

The representation $\boldsymbol{w} = X\boldsymbol{\lambda}$ is called the *dual representation*.

Substituting $X\boldsymbol{\lambda}$ for $\boldsymbol{w}$ yields another primal hard margin problem $(\bar{H})$:

$$(\bar{H}) \quad \left| \begin{array}{ll} \text{minimize} & \boldsymbol{\lambda}^\top K \boldsymbol{\lambda} \\ \text{subject to} & p_{\ell_i} + 1 \le (\boldsymbol{k}^i)^\top \boldsymbol{\lambda} \le p_{\ell_i+1} - 1 \qquad \text{for } i \in N, \end{array} \right.$$

where $(\boldsymbol{k}^i)^\top = ((\boldsymbol{x}^i)^\top \boldsymbol{x}^1, (\boldsymbol{x}^i)^\top \boldsymbol{x}^2, \ldots, (\boldsymbol{x}^i)^\top \boldsymbol{x}^n)$ is the ith row of the matrix $K$. Since $n$ is typically by far larger than $m$, problem $(\bar{H})$ might be less interesting than problem (H). However the fact that this formulation only requires the matrix $K$ will enable the application of kernel technique to the problem.

4. Algorithm for Hard Margin Problems

In this section, we consider the problem $(\bar{H})$ with dual representation of normal vector. The dimension m of the attribute vectors is usually much smaller than the number n of objects, hence we need a small number of attribute vectors for the dual representation. Furthermore it is likely that most of the constraints are redundant at the optimal solution. We introduce a subset W, called *working set*, of N and consider the following sub-problem.

$$(\bar{H}(W)) \quad \left| \begin{array}{ll} \text{minimize} & \boldsymbol{\lambda}_W^\top K_W \boldsymbol{\lambda}_W \\ \text{subject to} & p_{\ell_i} + 1 \leq (\boldsymbol{k}_W^i)^\top \boldsymbol{\lambda}_W \leq p_{\ell_i+1} - 1 \quad \text{for } i \in W, \end{array} \right.$$

where $(\boldsymbol{k}_W^i)^\top$ is the row vector consisting $(\boldsymbol{x}^i)^\top \boldsymbol{x}^j$ for $j \in W$. We propose to start the algorithm with a sub-problem which is much smaller than the problem $(\bar{H})$ in both variables and constraints, and increment W as the computation goes on. Note that the dimension of $\boldsymbol{\lambda}_W$ varies when the size of W changes as the computation goes on.

Algorithm RC$\bar{\text{H}}$ (Row and Column Generation Algorithm for $(\bar{H})$)

```
Step 1 : Let $W^0$ be an initial working set, and let $\nu = 0$.
Step 2 : Solve $(\bar{H}(W^\nu))$ to obtain $\boldsymbol{\lambda}_{W^\nu}$ and $\boldsymbol{p}^\nu$.
Step 3 : Let $\Delta = \{ i \in N \setminus W^\nu \mid (\boldsymbol{\lambda}_{W^\nu}, \boldsymbol{p}^\nu)$ violates $p_{\ell_i} + 1 \leq (\boldsymbol{k}_{W^\nu}^i)^\top \boldsymbol{\lambda}_W \leq p_{\ell_i+1} - 1 \}$.
Step 4 : If $\Delta = \emptyset$, terminate.
Step 5 : Otherwise choose $\Delta^\nu \subseteq \Delta$, let $W^{\nu+1} = W^\nu \cup \Delta^\nu$, increment $\nu$ by 1 and go to
         Step 2.
```

The following lemma shows that Algorithm RC$\bar{\text{H}}$ solves problem $(\bar{H})$ upon termination.

Lemma 4.1. *Let $(\hat{\boldsymbol{\lambda}}_W, \hat{\boldsymbol{p}}) \in \mathbb{R}^{|W|+l}$ be an optimum solution of $(\bar{H}(W))$. If*

$$\hat{p}_{\ell_i} + 1 \leq (\boldsymbol{k}_W^i)^\top \hat{\boldsymbol{\lambda}}_W \leq \hat{p}_{\ell_i+1} - 1 \quad \textit{for all } i \in N \setminus W, \tag{4.1}$$

then $(\hat{\boldsymbol{\lambda}}_W, \boldsymbol{0}_{N\setminus W}) \in \mathbb{R}^n$ together with $\hat{\boldsymbol{p}}$ forms an optimum solution of $(\bar{H})$.

Proof. Note that $((\hat{\boldsymbol{\lambda}}_W, \boldsymbol{0}_{N\setminus W}), \hat{\boldsymbol{p}})$ is a feasible solution of $(\bar{H})$ since $(\boldsymbol{k}^i)^\top \begin{pmatrix} \hat{\boldsymbol{\lambda}}_W \\ \boldsymbol{0}_{N\setminus W} \end{pmatrix} = (\boldsymbol{k}_W^i)^\top \hat{\boldsymbol{\lambda}}_W$, $(\hat{\boldsymbol{\lambda}}_W, \hat{\boldsymbol{p}})$ is feasible to $(\bar{H}(W))$ and satisfies (4.1).

For an optimum solution $(\boldsymbol{\lambda}^*, \boldsymbol{p}^*)$ of $(\bar{H})$, let $\boldsymbol{w}^* = X\boldsymbol{\lambda}^*$, $\boldsymbol{w}_1$ be its orthogonal projection onto the range space of X_W and $\boldsymbol{w}_2 = \boldsymbol{w}^* - \boldsymbol{w}_1$. Then $\boldsymbol{w}_1 = X_W \boldsymbol{\mu}_W^*$ for some $\boldsymbol{\mu}_W^* \in \mathbb{R}^{|W|}$ and

$$(\boldsymbol{\lambda}^*)^\top K \boldsymbol{\lambda}^* = \|\boldsymbol{w}^*\|^2 \geq \|\boldsymbol{w}_1\|^2 = (\boldsymbol{\mu}_W^*)^\top K_W \boldsymbol{\mu}_W^* \tag{4.2}$$

by the orthogonality between $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$. For $i \in N \cap W$ it holds that

$$\begin{aligned} (\boldsymbol{k}_W^i)^\top \boldsymbol{\mu}_W^* &= (\boldsymbol{x}^i)^\top X_W \boldsymbol{\mu}_W^* = (\boldsymbol{x}^i)^\top \boldsymbol{w}_1 = (\boldsymbol{x}^i)^\top (\boldsymbol{w}_1 + \boldsymbol{w}_2) \\ &= (\boldsymbol{x}^i)^\top \boldsymbol{w}^* = (\boldsymbol{x}^i)^\top X \boldsymbol{\lambda}^* = (\boldsymbol{k}^i)^\top \boldsymbol{\lambda}^*, \end{aligned}$$

which is between $p_{\ell_i}^* + 1$ and $p_{\ell_i+1}^* - 1$ since $(\boldsymbol{\lambda}^*, \boldsymbol{p}^*)$ is feasible to $(\bar{H})$. Then $(\boldsymbol{\mu}_W^*, \boldsymbol{p}^*)$ is feasible to $(\bar{H}(W))$. This and the optimality of $\hat{\boldsymbol{\lambda}}_W$ yield the inequality

$$(\boldsymbol{\mu}_W^*)^\top K_W \boldsymbol{\mu}_W^* \geq \hat{\boldsymbol{\lambda}}_W^\top K_W \hat{\boldsymbol{\lambda}}_W = \begin{pmatrix} \hat{\boldsymbol{\lambda}}_W \\ \boldsymbol{0}_{N\setminus W} \end{pmatrix}^\top K \begin{pmatrix} \hat{\boldsymbol{\lambda}}_W \\ \boldsymbol{0}_{N\setminus W} \end{pmatrix}. \tag{4.3}$$

The two inequalities (4.2) and (4.3) prove the optimality of $((\hat{\boldsymbol{\lambda}}_W, \mathbf{0}_{N\setminus W}), \hat{\boldsymbol{p}})$. □

Theorem 4.2. *The Algorithm RC$\bar{H}$ solves problem* $(\bar{H})$.

5. Kernel Technique for Hard Margin Problems

The matrix K in the primal hard margin problem $(\bar{H})$ with dual representation of normal vector is composed of the inner products $(\boldsymbol{x}^i)^\top \boldsymbol{x}^j$ for $i, j \in N$. This enables us to apply the *kernel technique* simply by replacing them by $\kappa(\boldsymbol{x}^i, \boldsymbol{x}^j)$ for some appropriate kernel function κ.

Let $\phi \colon \mathbb{R}^m \to \mathbb{F}$ be a function, possibly unknown, from $\mathbb{R}^m$ to some higher dimensional inner product space $\mathbb{F}$, so-called the *feature space* such that

$$\kappa(\boldsymbol{x}, \boldsymbol{y}) = \langle \phi(\boldsymbol{x}), \phi(\boldsymbol{y}) \rangle$$

holds for $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^m$, where $\langle \cdot, \cdot \rangle$ is the inner product defined on $\mathbb{F}$. In the sequel we denote $\tilde{\boldsymbol{x}} = \phi(\boldsymbol{x})$. The kernel technique considers the vectors $\tilde{\boldsymbol{x}}^i \in \mathbb{F}$ instead of $\boldsymbol{x}^i \in \mathbb{R}^m$, and finds the normal vector $\tilde{\boldsymbol{w}} \in \mathbb{F}$ and thresholds $p_1, \ldots, p_l$. Therefore the matrices X and K should be replaced by $\tilde{X}$ composed of vectors $\tilde{\boldsymbol{x}}^i$ and $\tilde{K} = \left[\langle \tilde{x}^i, \tilde{x}^j \rangle\right]_{i,j \in N}$, respectively. Note that the latter matrix is given as

$$\tilde{K} = \left[\kappa(\boldsymbol{x}^i, \boldsymbol{x}^j) \right]_{i,j \in N} \tag{5.1}$$

by the kernel function κ. The problem to solve is

$$(\tilde{H}) \quad \left| \begin{array}{ll} \text{minimize} & \boldsymbol{\lambda}^\top \tilde{K} \boldsymbol{\lambda} \\ \text{subject to} & p_{\ell_i} + 1 \le (\tilde{\boldsymbol{k}}^i)^\top \boldsymbol{\lambda} \le p_{\ell_i + 1} - 1 \qquad \text{for } i \in N. \end{array} \right.$$

Solving $(\tilde{H})$ to find $\boldsymbol{\lambda}^*$, the optimal normal vector $\tilde{\boldsymbol{w}}^* \in \mathbb{F}$ would be given as

$$\tilde{\boldsymbol{w}}^* = \sum_{i \in N} \lambda_i^* \tilde{\boldsymbol{x}}^i,$$

which is not available in general due to the absence of an explicit representation of $\tilde{\boldsymbol{x}}^i$'s. However, the value of $\langle \tilde{\boldsymbol{w}}^*, \tilde{\boldsymbol{x}} \rangle$ can be computed for a given attribute vector $\boldsymbol{x} \in \mathbb{R}^n$ in the following way:

$$\langle \tilde{\boldsymbol{w}}^*, \tilde{\boldsymbol{x}} \rangle = \langle \sum_{i \in N} \lambda_i^* \tilde{\boldsymbol{x}}^i, \tilde{\boldsymbol{x}} \rangle = \sum_{i \in N} \lambda_i^* \langle \tilde{\boldsymbol{x}}^i, \tilde{\boldsymbol{x}} \rangle = \sum_{i \in N} \lambda_i^* \langle \phi(\boldsymbol{x}^i), \phi(\boldsymbol{x}) \rangle = \sum_{i \in N} \lambda_i^* \kappa(\boldsymbol{x}^i, \boldsymbol{x}).$$

Then by locating the threshold interval determined by $\boldsymbol{p}^*$ into which this value falls, we can assign a label to the new object with $\boldsymbol{x}$.

In the same way as for the hard margin problem $(\bar{H})$, we consider the sub-problem

$$(\tilde{H}(W)) \quad \left| \begin{array}{ll} \text{minimize} & \boldsymbol{\lambda}_W^\top \tilde{K}_W \boldsymbol{\lambda}_W \\ \text{subject to} & p_{\ell_i} + 1 \le (\tilde{\boldsymbol{k}}_W^i)^\top \boldsymbol{\lambda}_W \le p_{\ell_i + 1} - 1 \quad \text{for } i \in W, \end{array} \right.$$

where $\tilde{K}_W$ is the sub-matrix consisting of rows and columns of $\tilde{K}$ with indices in W, and $(\tilde{\boldsymbol{k}}_W^i)^\top$ is the row vector of $\kappa(\boldsymbol{x}^i, \boldsymbol{x}^j)$ for $j \in W$.

Algorithm RC$\tilde{\text{H}}$ (Row and Column Generation Algorithm for $(\tilde{H})$)

```
Step 1 : Let $W^0$ be an initial working set, and let $\nu = 0$.
Step 2 : Solve $(\tilde{H}(W^\nu))$ to obtain $\boldsymbol{\lambda}_{W^\nu}$ and $\boldsymbol{p}^\nu$.
Step 3 : Let $\Delta = \{\, i \in N \setminus W^\nu \mid (\boldsymbol{\lambda}_{W^\nu}, \boldsymbol{p}^\nu)$ violates $p_{\ell_i} + 1 \leq (\tilde{\boldsymbol{k}}^i_{W^\nu})^\top \boldsymbol{\lambda}_W \leq p_{\ell_i+1} - 1 \,\}$.
Step 4 : If $\Delta = \emptyset$, terminate.
Step 5 : Otherwise choose $\Delta^\nu \subseteq \Delta$, let $W^{\nu+1} = W^\nu \cup \Delta^\nu$, increment $\nu$ by 1 and go to
         Step 2.
```

The validity of the algorithm is straightforward from the following lemma, which is proved in exactly the same way as Lemma 4.1

Lemma 5.1. *Let $(\hat{\boldsymbol{\lambda}}_W, \hat{\boldsymbol{p}}) \in \mathbb{R}^{|W|+l}$ be an optimum solution of $(\tilde{H}(W))$. If*

$$\hat{p}_{\ell_i} + 1 \leq (\tilde{\boldsymbol{k}}^i_W)^\top \hat{\boldsymbol{\lambda}}_W \leq \hat{p}_{\ell_i+1} - 1 \quad \textit{for all } i \in N \setminus W,$$

then $(\hat{\boldsymbol{\lambda}}_W, \mathbf{0}_{N\setminus W}) \in \mathbb{R}^n$ together with $\hat{\boldsymbol{p}}$ forms an optimum solution of $(\tilde{H})$.

Theorem 5.2. *The Algorithm RC$\tilde{H}$ solves problem $(\tilde{H})$.*

6. Soft Margin Problems for Non-Separable Case

6.1. **Primal Soft Margin Problems.** Similarly to the binary SVM, introducing nonnegative slack variables ξ_{-i} and ξ_{+i} for $i \in N$ relaxes the hard margin constraints to *soft margin* constraints:

$$p_{\ell_i} + 1 - \xi_{-i} \leq \boldsymbol{w}^\top \boldsymbol{x}^i \leq p_{\ell_i+1} - 1 + \xi_{+i} \quad \text{for } i \in N.$$

Positive values of variables ξ_{-i} and ξ_{+i} mean misclassification, hence they should be as small as possible. If we penalize positive ξ_{-i} and ξ_{+i} by adding $\sum_{i\in N}(\xi_{-i} + \xi_{+i})$ to the objective function, we have the following *primal soft margin problem* with *1-norm penalty.*

$$(S_1) \quad \left| \begin{array}{ll} \text{minimize} & \|\boldsymbol{w}\|^2 + c\,\mathbf{1}_n^\top(\boldsymbol{\xi}_- + \boldsymbol{\xi}_+) \\ \text{subject to} & p_{\ell_i} + 1 - \xi_{-i} \leq (\boldsymbol{x}^i)^\top \boldsymbol{w} \leq p_{\ell_i+1} - 1 + \xi_{+i} \quad \text{for } i \in N \\ & \boldsymbol{\xi}_-, \boldsymbol{\xi}_+ \geq \mathbf{0}_n, \end{array} \right.$$

where $\boldsymbol{\xi}_- = (\xi_{-1}, \ldots, \xi_{-n}), \boldsymbol{\xi}_+ = (\xi_{+1}, \ldots, \xi_{+n})$ and c is a *penalty parameter.* When *2-norm penalty* is employed, we have

$$(S_2) \quad \left| \begin{array}{ll} \text{minimize} & \|\boldsymbol{w}\|^2 + c\,(\|\boldsymbol{\xi}_-\|^2 + \|\boldsymbol{\xi}_+\|^2) \\ \text{subject to} & p_{\ell_i} + 1 - \xi_{-i} \leq (\boldsymbol{x}^i)^\top \boldsymbol{w} \leq p_{\ell_i+1} - 1 + \xi_{+i} \quad \text{for } i \in N \\ & \boldsymbol{\xi}_-, \boldsymbol{\xi}_+ \geq \mathbf{0}_n. \end{array} \right.$$

Lemma 6.1. *The nonnegativity constraints on variables ξ_{-i} and ξ_{+i} of problem (S_2) are redundant.*

Proof. Let $(\boldsymbol{w}, \boldsymbol{\xi}_-, \boldsymbol{\xi}_+)$ be a feasible solution of (S_2) with the nonnegativity constraints removed. If $\xi_{-i} < 0$ for some $i \in N$, replacing it with zero will reduce the objective function value. Therefore $\boldsymbol{\xi}_-$ and $\boldsymbol{\xi}_+$ are nonnegative at any optimum solution of (S_2). □

Thus our problem with 2-norm penalty reduces to

$$(S_2)\qquad \left|\begin{array}{ll} \text{minimize} & \|\boldsymbol{w}\|^2 + c\,(\|\boldsymbol{\xi}_-\|^2 + \|\boldsymbol{\xi}_+\|^2) \\ \text{subject to} & p_{\ell_i} + 1 - \xi_{-i} \le (\boldsymbol{x}^i)^\top \boldsymbol{w} \le p_{\ell_i+1} - 1 + \xi_{+i} \quad \text{for } i \in N. \end{array}\right.$$

As proposed in Mangasarian and Musicant [5], the addition of a term $\|\boldsymbol{p}\|^2$ to the objective function yields the following two formulations (S_{12}) and (S_{22}).

$$(S_{12})\qquad \left|\begin{array}{ll} \text{minimize} & \|\boldsymbol{w}\|^2 + c\,\mathbf{1}_n^\top(\boldsymbol{\xi}_- + \boldsymbol{\xi}_+) + d\,\|\boldsymbol{p}\|^2 \\ \text{subject to} & p_{\ell_i} + 1 - \xi_{-i} \le (\boldsymbol{x}^i)^\top \boldsymbol{w} \le p_{\ell_i+1} - 1 + \xi_{+i} \quad \text{for } i \in N \\ & \boldsymbol{\xi}_-, \boldsymbol{\xi}_+ \ge \mathbf{0}_n, \end{array}\right.$$

and

$$(S_{22})\qquad \left|\begin{array}{ll} \text{minimize} & \|\boldsymbol{w}\|^2 + c\,(\|\boldsymbol{\xi}_-\|^2 + \|\boldsymbol{\xi}_+\|^2) + d\,\|\boldsymbol{p}\|^2 \\ \text{subject to} & p_{\ell_i} + 1 - \xi_{-i} \le (\boldsymbol{x}^i)^\top \boldsymbol{w} \le p_{\ell_i+1} - 1 + \xi_{+i} \quad \text{for } i \in N, \end{array}\right.$$

where d is a penalty parameter.

Naturally, we could add to each of the above formulations the constraints

$$p_{k'} + 1 - \xi_{-i} \le (\boldsymbol{x}^i)^\top \boldsymbol{w} \le p_{k''} - 1 + \xi_{+i} \quad \text{for } k', k'' \in L \text{ such that } k' \le \ell_i < k''.$$

It would, however, inflate the problem size and most of those constraints would be likely redundant. Therefore we will not discuss this formulation.

6.2. **Dual Representation.** Obviously we can replace $\|\boldsymbol{w}\|^2$ and $(\boldsymbol{x}^i)^\top \boldsymbol{w}$ in each of the primal problems given in the preceding subsection by $\boldsymbol{\lambda}^\top K \boldsymbol{\lambda}$ and $(\boldsymbol{k}^i)^\top \boldsymbol{\lambda}$, respectively, to obtain the primal problem with dual representation of normal vector. Take (S_1) for instance, and we have

$$(\bar{S}_1)\qquad \left|\begin{array}{ll} \text{minimize} & \boldsymbol{\lambda}^\top K \boldsymbol{\lambda} + c\,\mathbf{1}_n^\top(\boldsymbol{\xi}_- + \boldsymbol{\xi}_+) \\ \text{subject to} & p_{\ell_i} + 1 - \xi_{-i} \le (\boldsymbol{k}^i)^\top \boldsymbol{\lambda} \le p_{\ell_i+1} - 1 + \xi_{+i} \quad \text{for } i \in N \\ & \boldsymbol{\xi}_-, \boldsymbol{\xi}_+ \ge \mathbf{0}_n. \end{array}\right.$$

7. Algorithm for Soft Margin Problems

The algorithm for the soft margin problems may not differ substantially from that for the hard margin problems. Take the primal soft margin problem $(\bar{S}_1)$ with dual representation of normal vector. The sub-problem to solve is

$$(\bar{S}_1(W))\qquad \left|\begin{array}{ll} \text{minimize} & \boldsymbol{\lambda}_W^\top K_W \boldsymbol{\lambda}_W + c\,\mathbf{1}_{|W|}^\top(\boldsymbol{\xi}_{-W} + \boldsymbol{\xi}_{+W}) \\ \text{subject to} & p_{\ell_i} + 1 - \xi_{-i} \le (\boldsymbol{k}_W^i)^\top \boldsymbol{\lambda}_W \le p_{\ell_i+1} - 1 + \xi_{+i} \quad \text{for } i \in W \\ & \boldsymbol{\xi}_{-W},\, \boldsymbol{\xi}_{+W} \ge \mathbf{0}_{|W|}. \end{array}\right.$$

We propose the following algorithm for $(\bar{S}_1)$ in which we solve $(\bar{S}_1(W))$ repeatedly.

Algorithm RC$\bar{\mathbf{S}}_1$ (Row and Column Generation Algorithm for $(\bar{S}_1)$)

Step 1 : Let W^0 be an initial working set, and let $\nu = 0$.

Step 2 : Solve $(\bar{S}_1(W^\nu))$ to obtain $(\boldsymbol{\lambda}_{W^\nu}, \boldsymbol{p}^\nu, \boldsymbol{\xi}_{-W^\nu}, \boldsymbol{\xi}_{+W^\nu})$.
Step 3 : Let $\Delta = \{ i \in N \setminus W^\nu \mid (\boldsymbol{\lambda}_{W^\nu}, \boldsymbol{p}^\nu)$ violates $p_{\ell_i} + 1 \leq (\boldsymbol{k}^i_{W^\nu})^\top \boldsymbol{\lambda}_W \leq p_{\ell_i+1} - 1 \}$.
Step 4 : If $\Delta = \emptyset$, terminate.
Step 5 : Otherwise choose $\Delta^\nu \subseteq \Delta$, let $W^{\nu+1} = W^\nu \cup \Delta^\nu$, increment ν by 1 and go to Step 2.

Lemma 7.1. *Let $(\hat{\boldsymbol{\lambda}}_W, \hat{\boldsymbol{p}}, \hat{\boldsymbol{\xi}}_{-W}, \hat{\boldsymbol{\xi}}_{+W})$ be an optimum solution of $(\bar{S}_1(W))$. If*

$$\hat{p}_{\ell_i} + 1 \leq (\boldsymbol{k}^i_W)^\top \hat{\boldsymbol{\lambda}}_W \leq \hat{p}_{\ell_i+1} - 1 \quad \textit{for all } i \in N \setminus W,$$

then $((\hat{\boldsymbol{\lambda}}_W, \boldsymbol{0}_{N\setminus W}), \hat{\boldsymbol{p}}, (\boldsymbol{\xi}^\nu_{-W}, \boldsymbol{0}_{N\setminus W}), (\boldsymbol{\xi}^\nu_{+W}, \boldsymbol{0}_{N\setminus W}))$ is an optimum solution of $(\bar{S}_1)$.

Proof. First note that $((\hat{\boldsymbol{\lambda}}_W, \boldsymbol{0}_{N\setminus W}), \hat{\boldsymbol{p}}, (\hat{\boldsymbol{\xi}}_{-W}, \boldsymbol{0}_{N\setminus W}), (\hat{\boldsymbol{\xi}}_{+W}, \boldsymbol{0}_{N\setminus W}))$ is feasible to $(\bar{S}_1)$. Let $(\boldsymbol{\lambda}^*, \boldsymbol{p}^*, \boldsymbol{\xi}^*_{-W}, \boldsymbol{\xi}^*_{+W})$ be an optimum solution of $(\bar{S}_1)$, let $\boldsymbol{w}^* = X\boldsymbol{\lambda}^*$ and $\boldsymbol{w}_1$ be its orthogonal projection onto the range space of X_W. Then we see that the coefficient vector $\boldsymbol{\mu}^*_W$ such that $\boldsymbol{w}_1 = X_W \boldsymbol{\mu}^*_W$ together with $(\boldsymbol{p}^*, \boldsymbol{\xi}^*_{-W}, \boldsymbol{\xi}^*_{+W})$ forms a feasible solution of $(\bar{S}_1(W))$, and $\|\boldsymbol{w}^*\| \geq \|\boldsymbol{w}_1\|$. Therefore in the same manner as Lemma 4.1, we obtain the desired result. □

The validity of the algorithm directly follows the above lemma.

Theorem 7.2. *The Algorithm RC$\bar{S}_1$ solves problem $(\bar{S}_1)$.*

8. Kernel Technique for Soft Margin Problems

The kernel technique can apply to the soft margin problem in the same way as discussed in Section 5.

For the kernel version of soft margin problems with dual representation of normal vector, we have only to replace K by $\tilde{K}$ given by some kernel function κ. The kernel version of $(\bar{S}_1)$ for instance, is given as

$$(\tilde{S}_1) \quad \left| \begin{array}{lll} \text{minimize} & \boldsymbol{\lambda}^\top \tilde{K} \boldsymbol{\lambda} + c\boldsymbol{1}_n^\top(\boldsymbol{\xi}_- + \boldsymbol{\xi}_+) & \\ \text{subject to} & p_{\ell_i} + 1 - \xi_{-i} \leq (\tilde{\boldsymbol{k}}^i)^\top \boldsymbol{\lambda} \leq p_{\ell_i+1} - 1 + \xi_{+i} & \text{for } i \in N \\ & \boldsymbol{\xi}_-, \boldsymbol{\xi}_+ \geq \boldsymbol{0}_n. & \end{array} \right.$$

In the same way as in the previous section we consider the sub-problem of $(\tilde{S}_1)$, which is given as

$$(\tilde{S}_1(W)) \quad \left| \begin{array}{lll} \text{minimize} & \boldsymbol{\lambda}_W^\top \tilde{K}_W \boldsymbol{\lambda}_W + c\boldsymbol{1}_{|W|}^\top(\boldsymbol{\xi}_{-W} + \boldsymbol{\xi}_{+W}) & \\ \text{subject to} & p_{\ell_i} + 1 - \xi_{-i} \leq (\tilde{\boldsymbol{k}}^i_W)^\top \boldsymbol{\lambda}_W \leq p_{\ell_i+1} - 1 + \xi_{+i} & \text{for } i \in W \\ & \boldsymbol{\xi}_{-W}, \boldsymbol{\xi}_{+W} \geq \boldsymbol{0}_{|W|}. & \end{array} \right.$$

Algorithm RC$\tilde{\text{S}}_1$ (Row and Column Generation Algorithm for $(\tilde{S}_1)$)
Step 1 : Let W^0 be an initial working set, and let $\nu = 0$.
Step 2 : Solve $(\tilde{S}_1(W^\nu))$ to obtain $(\boldsymbol{\lambda}_{W^\nu}, \boldsymbol{p}^\nu, \boldsymbol{\xi}_{-W^\nu}, \boldsymbol{\xi}_{+W^\nu})$.
Step 3 : Let $\Delta = \{ i \in N \setminus W^\nu \mid (\boldsymbol{\lambda}_{W^\nu}, \boldsymbol{p}^\nu)$ violates $p_{\ell_i} + 1 \leq (\tilde{\boldsymbol{k}}^i_{W^\nu})^\top \boldsymbol{\lambda}_W \leq p_{\ell_i+1} - 1 \}$.
Step 4 : If $\Delta = \emptyset$, terminate.

Step 5 : Otherwise choose $\Delta^{\nu} \subseteq \Delta$, let $W^{\nu+1} = W^{\nu} \cup \Delta^{\nu}$, increment ν by 1 and go to Step 2.

We then obtain the following theorem.

Theorem 8.1. *The Algorithm $RC\tilde{S}_1$ solves problem $(\tilde{S}_1)$.*

9. Illustrative Example

We show with a small instance how different models result in different classifications. The instance is the grades in calculus of 44 undergraduates. Each student is given one of the four possible grades A, B, C, D according to his/her total score of mid-term exam, end-of-term exam and a number of in-class quizzes. We take the scores of mid-term and end-of-term exams to form an attribute vector, and the grade as a label.

Since the score of quizzes is not considered as an attribute, the instance is not separable. Hence the hard margin problem (H) is infeasible. The solution of the soft margin problem (S_1) with $c = 15$ is given in Fig. 1, where students of different grades are loosely separated by three straight lines.

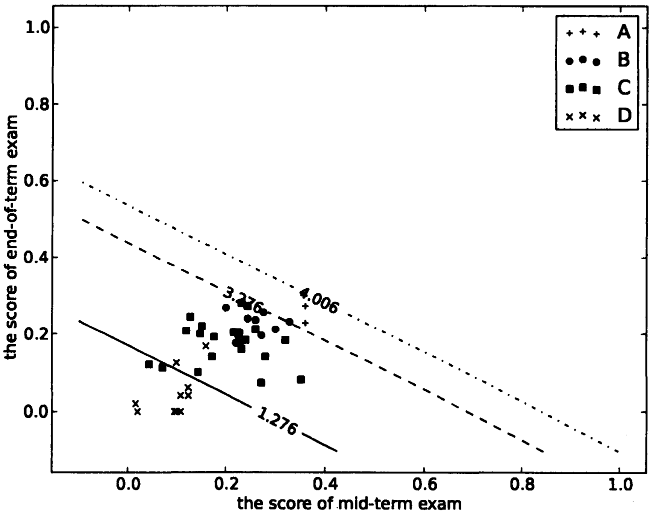

Figure 1. Classification by (S_1)

Using the following two kernel functions defined as

$$\kappa(\boldsymbol{x}, \boldsymbol{y}) = \exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{x} - \boldsymbol{y}\|^2\right) \quad \text{(Gaussian kernel)},$$

$$\kappa(\boldsymbol{x}, \boldsymbol{y}) = (1 + \boldsymbol{x}^{\mathrm{T}}\boldsymbol{y})^d \quad \text{(Polynomial kernel)}$$

with several different values of σ and d, we solved $(\tilde{S}_1)$. The result of the Gaussian kernel with $c = 10$ and $\sigma = 0.5$ is given in Fig. 2, where one can observe that the problem $(\tilde{S}_1)$ with the Gaussian kernel is exposed to the risk of over-fitting. The result of the polynomial kernel with $c = 15$ and $d = 4$ is given in Fig. 3. From Fig. 3, we observe that students of different grades are separated by three gentle curves.

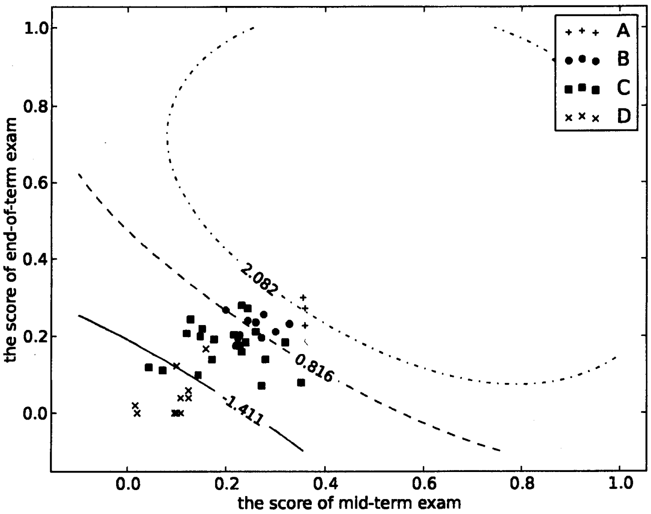

FIGURE 2. Classification by $(\tilde{S}_1)$ with the Gaussian kernel

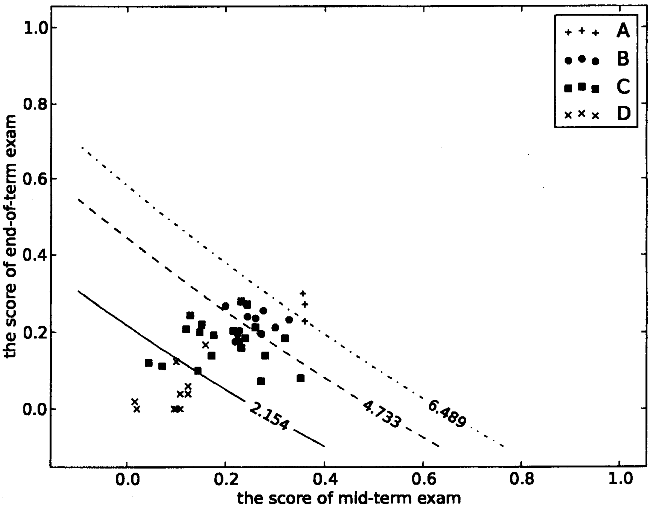

FIGURE 3. Classification by $(\tilde{S}_1)$ with the polynomial kernel

10. COMPUTATIONAL EXPERIMENTS

We report the computational experiments with our proposed algorithms. The experiment was performed on a PC with Intel Core i7, 3.07 GHz processor and 12.0 GB of memory. We implemented the algorithms in Python, and used Gurobi 5.6.3 as the QP solver.

We used several randomly generated instances, which are divided into two types. The first type consists of separable instances and the second type consists of non-separable instances. First we generated n attribute vectors $\boldsymbol{x}^i = (x_1^i, x_2^i)$ at random from the unit square $[0,1] \times [0,1]$. Each object i was assigned a label $\ell_i \in L = \{0, 1, \ldots, 3\}$ according to the following rule:

$$\ell_i = \max_{\ell \in L}\{\ell \in L \mid x_1^i + x_2^i > p_\ell\}$$

where $(p_0, p_1, p_2, p_3) = (-\infty, 0.5, 1.0, 1.5)$. Then this instance is separable. Perturbing each element of the attribute vector, we have a perturbed attribute vector $(x_1^i + \varepsilon_1^i, x_2^i + \varepsilon_2^i)$, where ε_1^i as well as ε_2^i is a random noise following a normal distribution with a zero mean and a standard deviation of 0.03. Then we give a label ℓ_i to an object i according to the sum of elements of the perturbed vector instead of the value $x_1^i + x_2^i$. Due to the presence of the random noise, the instance is not necessarily separable. We generate five datasets for each instance with n objects since the results may change owing to the random variables used in the above instance generation method. We name the separable type datasets (the non-separable datasets) "S.n.q" ("NS.n.q"), where q is a dataset ID.

We solved the separable instances by the algorithm $RC\tilde{H}$ and the non-separable instances by the algorithm $RC\tilde{S}_1$ with $c = 10$. In all experiments we used the polynomial kernel with a parameter $d = 4$. To make the initial working set W^0, we first calculate the value of $x_1^i + x_2^i$ for each object i, and choose three objects whose values are the highest, the lowest and the median among objects assigned the same label. At Step 5 in the algorithms, we add at most two objects corresponding to the most violated constraints at $(\boldsymbol{\lambda}_{W^\nu}, \boldsymbol{p}^\nu)$ to the current working set W^ν at a time, more precisely, we add the objects i and $j \in N \setminus W^\nu$ such that

$$i = \operatorname{argmax}\left\{1 - (\tilde{\boldsymbol{k}}_{W^\nu}^i)^\top \boldsymbol{\lambda}_{W^\nu} + p_{\ell_i}^\nu > 0 \mid i \in N \setminus W^\nu\right\},$$

$$j = \operatorname{argmax}\left\{1 + (\tilde{\boldsymbol{k}}_{W^\nu}^i)^\top \boldsymbol{\lambda}_{W^\nu} - p_{\ell_i+1}^\nu > 0 \mid i \in N \setminus W^\nu\right\}.$$

Table 1 shows the results of $RC\tilde{H}$ and $RC\tilde{S}_1$ for each instance, where the columns "# iter.", "# added obj." and "time" represent the number of sub-problems solved, the number of added objects and the computation time in seconds, respectively. The entries "ave." and "st.dev." show the average and the standard deviation across the five results of a corresponding column.

TABLE 1. Results of $RC\tilde{H}$ and $RC\tilde{S}_1$

instance	n	# iter.		# added obj.		time (s)	
		ave.	st.dev.	ave.	st.dev.	ave.	st.dev.
S.100.1 – S.100.5	100	5.80	1.79	8.60	3.05	0.32	0.18
S.500.1 – S.500.5	500	8.40	1.34	13.80	2.39	1.36	0.32
S.1000.1 – S.1000.5	1000	11.80	0.84	20.40	0.89	4.73	0.46
NS.100.1 – NS.100.5	100	14.40	3.71	23.40	4.16	2.06	0.83
NS.500.1 – NS.500.5	500	65.00	30.91	108.80	23.54	108.23	103.02
NS.1000.1 – NS.1000.5	1000	111.80	43.50	198.60	41.43	580.94	524.09

From Table 1, we observe that the number of added objects is much smaller than that of the original problem in the separable case. In the non-separable case, we observe that the number of iterations as well as the number of added objects is larger than the separable case. Nevertheless the number of added objects was approximately 20% of the whole.

11. Monotonicity issue

In some situations it would be desirable that the separating curves have some monotonicity property, namely an object with attribute vector $\boldsymbol{x}$ be ranked higher than an object with $\boldsymbol{y}$ such that $\boldsymbol{y} \le \boldsymbol{x}$. Then we discuss the monotonicity of the separating curves in this section.

Let P be a hyperplane in $\mathbb{F}$ defined by

$$P = \{\, \tilde{\boldsymbol{x}} \in \mathbb{F} \mid \langle \tilde{\boldsymbol{w}}^*, \tilde{\boldsymbol{x}} \rangle = b \,\}$$

for some constant $b \in \mathbb{R}$ and let C denote its inverse image under the unknown function ϕ, i.e.,

$$C = \{\, \boldsymbol{x} \in \mathbb{R}^m \mid \phi(\boldsymbol{x}) \in P \,\}.$$

Then $\boldsymbol{x} \in C$ if and only if $\langle \tilde{\boldsymbol{w}}^*, \phi(\boldsymbol{x}) \rangle = b$. Since $\tilde{\boldsymbol{w}}^* = \sum_{i \in N} \lambda_i^* \tilde{\boldsymbol{x}}^i = \sum_{i \in N} \lambda_i^* \phi(\boldsymbol{x}^i)$, we obtain

$$\langle \sum_{i \in N} \lambda_i^* \phi(\boldsymbol{x}^i), \phi(\boldsymbol{x}) \rangle = b.$$

Due to the bi-linearlity of the inner product $\langle \cdot, \cdot \rangle$, we have

$$\langle \sum_{i \in N} \lambda_i^* \phi(\boldsymbol{x}^i), \phi(\boldsymbol{x}) \rangle = \sum_{i \in N} \lambda_i^* \langle \phi(\boldsymbol{x}^i), \phi(\boldsymbol{x}) \rangle = \sum_{i \in N} \lambda_i^* \kappa(\boldsymbol{x}^i, \boldsymbol{x}),$$

and then an expression of the inverse image

$$C = \{\, \boldsymbol{x} \in \mathbb{R}^m \mid \sum_{i \in N} \lambda_i^* \kappa(\boldsymbol{x}^i, \boldsymbol{x}) = b \,\}$$

by the kernel function κ.

Suppose that the kernel function $\kappa(\boldsymbol{x}^i, \cdot)$ is nondecreasing for $i \in N$, in the sense that

$$\boldsymbol{x} \le \boldsymbol{x}' \Rightarrow \kappa(\boldsymbol{x}^i, \boldsymbol{x}) \le \kappa(\boldsymbol{x}^i, \boldsymbol{x}'),$$

and $\lambda_i^* \ge 0$ for $i \in N$. Then $\sum_{i \in N} \lambda_i^* \kappa(\boldsymbol{x}^i, \boldsymbol{x})$ is nondecreasing as a whole.

Lemma 11.1. *The kernel function $\kappa(\boldsymbol{x}^i, \cdot)$ is nondecreasing and $\lambda_i^* \ge 0$ for $i \in N$. Then the contours are nondecreasing.*

The polynomial kernel

$$\kappa(\boldsymbol{x}^i, \boldsymbol{x}) = (1 + (\boldsymbol{x}^i)^\top \boldsymbol{x})^d$$

is nondecreasing with respect to $\boldsymbol{x}$ if $\boldsymbol{x}^i \ge \boldsymbol{0}$ for $i \in N$. Therefore it would be appropriate to use the polynomial kernel when all the attribute vectors are nonnegative including those of potential objects, and the monotonicity is desirable. In this case the kernel hard margin problem $(\tilde{H})$ should be added non-negativity constraints of variables λ_i's:

$$(\tilde{H}_+) \quad \left| \begin{array}{lll} \text{minimize} & \boldsymbol{\lambda}^\top \tilde{K} \boldsymbol{\lambda} & \\ \text{subject to} & p_{\ell_i} + 1 \le (\tilde{\boldsymbol{k}}^i)^\top \boldsymbol{\lambda} \le p_{\ell_i + 1} - 1 & \text{for } i \in N \\ & \boldsymbol{\lambda} \ge \boldsymbol{0}. & \end{array} \right.$$

The problem remains an ordinary convex quadratic optimization, and the additional non-negativity constraints do not make it more difficult to solve.

12. Conclusions

In this paper, we proposed to apply the dual representation of the normal vector to the formulation based on the fixed margin strategy by Shashua and Levin [6] for the ranking problem. The obtained problem has the drawback that it has n of variables as well as n of constraints. However the fact that it enables the application of kernel technique outweighs the drawback. Then we proposed a row and column generation algorithm. Namely, we start the algorithm with a sub-problem which is much smaller than the master problem in both variables and constraints, and increment both of them as the computation goes on. Furthermore we proved the validity of the algorithm. Through some preliminary experiments, our algorithm performed fairly well. However it should need further research such as a setting of the initial working set W^0 and a choice of Δ^ν since a clever choice of these may enhance the efficiency of the algorithm.

References

[1] C.M. Bishop, *Pattern Recognition and Machine Learning*, Springer, New York, 2006.

[2] K. Crammer and Y. Singer, "Pranking with ranking," in: T.G. Dietterich, S. Becker and Z. Ghahramani eds., *Advances in Neural Information Processing Systems 14*, MIT Press, Cambridge, 2002, pp.641–647.

[3] R. Herbrich, T. Graepel and K. Obermayer, "Large margin rank boundaries for ordinal regression," in: A.J. Smola, P. Bartlette, B. Schölkopt and D. Schuurmans eds., *Advances in Large Margin Classifiers*, MIT Press, Cambridge, 2000, pp.115–132.

[4] T.-Y. Liu, *Learning to Rank for Information Retrieval*, Springer-Verlag, Heidelberg, 2011.

[5] O.L. Mangasarian and D.R. Musicant, "Successive overrelaxation for support vector machines," *IEEE Transactions on Neural Networks* **10** (1999) 5, 1032–1037.

[6] A. Shashua and A. Levin, "Ranking with large margin principles: two approaches," in: *Advances in Neural Information Processing Systems 15 (NIPS 2002)*,2003, pp.937–944.

[7] B. Schölkopf, R. Herbrich and A.J. Smola, "A generalized representer theorem," in : D. Helmbold and B. Williamson eds., *Computational Learning Theory*, Lecture Notes in Computer Science Vol. 2111 (2001) pp. 416–426.

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

[9] K. Tatsumi, K. Hayashida, R. Kawachi and T. Tanino, "Multiobjective multiclass support vector machines maximizing geometric margins," *Pacific Journal of Optimization* **6** (2010) 115–140.

[10] V.N. Vapnik, *Statistical Learning Theory*, John-Wiley & Sons, New York, 1998.

(Y. Izunaga) Graduate School of Systems and Information Engineering, University of Tsukuba, Tsukuba, Ibaraki 305-8573, Japan
E-mail address: s1130131@sk.tsukuba.ac.jp

(K. Sato) Signalling and Transport Information Technology Division, Railway Technical Research Institute, 2-8-38 Hikari-cho, Kokubunji, Tokyo 185-8540, Japan
E-mail address: sato.keisuke.49@rtri.or.jp

(K. Tatsumi) Electrical, Electronic and Information Engineering, Graduate School of Engineering, Osaka University, Suita, Osaka 565-0871, Japan
E-mail address: tatsumi@eei.eng.osaka-u.ac.jp

(Y. Yamamoto) Faculty of Engineering, Information and Systems, University of Tsukuba, Tsukuba, Ibaraki 305-8573, Japan
E-mail address: yamamoto@sk.tsukuba.ac.jp