Row and Column Generation Algorithm for Maximization of Minimum Margin for Ranking Problems

Izunaga, Yoichi
University of Tsukuba

Sato, Keisuke
Railway Technical Research Institute

Tatsumi, Keiji
Osaka University

Yamamoto, Yoshitsugu
University of Tsukuba

出版情報：Operations research proceedings. 2014, pp.249-255, 2016-02-21. Springer
バージョン：
権利関係：

Row and Column Generation Algorithm for Maximization of Minimum Margin for 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 two different formulations: primal problem, primal problem with dual representation of normal vector, and then propose to apply the kernel technique to the latter 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* $x^i = (x^i_1, x^i_2, \ldots, x^i_m)^\top \in \mathbb{R}^m$ 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 $y_i = w^\top x^i + \varepsilon_i = \sum_{j=1}^m w_j x^i_j + \varepsilon_i$ 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 $H(w, p_k) = \{ x \in \mathbb{R}^m \mid w^\top x = p_k \}$ for $k = 1, 2, \ldots, l$ which share a common normal vector 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 $w \in \mathbb{R}^m$ as well as the thresholds $p_1, p_2, \ldots, p_l$ that best fit the input data $\{ (x^i, \ell_i) \mid i = 1, 2, \ldots, n \}$.

Yoichi Izunaga, Yoshitsugu Yamamoto
University of Tsukuba, Ibaraki 305-8573, Japan, e-mail: s1130131@sk.tsukuba.ac.jp,
e-mail: yamamoto@sk.tsukuba.ac.jp

Keisuke Sato
Railway Technical Research Institute, Tokyo 185-8540, Japan e-mail: sato.keisuke.49@rtri.or.jp

Keiji Tatsumi
Osaka University, Osaka 565-0871, Japan, e-mail: tatsumi@eei.eng.osaka-u.ac.jp

This problem is known as the *ranking problem* and frequently arises in social sciences and operations research. See, for instance [2, 3, 4, 5, 7]. 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 [1, 8, 9]. 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 [5], we propose a row and column generation algorithm to maximize the minimum margin for the ranking problems.

Throughout the paper $N = \{1, 2, \ldots, i, \ldots, n\}$ denotes the set of n objects and $x^i = (x^i_1, x^i_2, \ldots, x^i_m)^\top \in \mathbb{R}^m$ 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 $k \in L$. For succinct notation we define $X = [x^i]_{i \in N} \in \mathbb{R}^{m \times n}$, $X_W = [x^i]_{i \in W} \in \mathbb{R}^{m \times |W|}$ for $W \subseteq N$, and the corresponding Gram matrices $K = X^\top X \in \mathbb{R}^{n \times n}$, $K_W = X_W^\top X_W \in \mathbb{R}^{|W| \times |W|}$. We denote the k-dimensional zero vector and vector of 1's by $\mathbf{0}_k$ and $\mathbf{1}_k$, respectively.

2 Hard Margin Problem for Separable Case

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 $\{\, (x^i, \ell_i) \mid i \in N \,\}$ is *separable* if there exist $w \in \mathbb{R}^m$ and $p = (p_1, p_2, \ldots, p_l)^\top \in \mathbb{R}^l$ such that $p_{\ell_i} < w^\top x^i < p_{\ell_i + 1}$ for any $i \in N$. Clearly an instance is separable if and only if there are w and p such that $p_{\ell_i} + 1 \leq w^\top x^i \leq p_{\ell_i + 1} - 1$ for any $i \in N$.

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

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

The constraints therein are called *hard margin* constraints.

A close look at the primal problem (H) shows that the following property holds for the optimum solution w^*. See, for example [1, 5, 6].

Lemma 1. *Let $(w^*, p^*) \in \mathbb{R}^{m+l}$ be an optimum solution of (H). Then $w^* \in \mathbb{R}^m$ lies in the range space of X, i.e., $w^* = X\lambda$ for some $\lambda \in \mathbb{R}^n$.*

The representation $w = X\lambda$ is called the *dual representation*. Substituting $X\lambda$ for w yields another primal hard margin problem $(\bar{H})$:

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

where $(k^i)^\top = ((x^i)^\top x^1, (x^i)^\top x^2, \ldots, (x^i)^\top x^n)$ is the ith row of the matrix K. Since n is typically by far more than m, problem $(\bar{H})$ might be less interesting than problem (H). However, the dimension m of the attribute vector is usually much smaller than the number of objects, hence we need a small number of attribute vectors for the dual representation, and it is likely that most of the constraints are redundant at the optimal solution. Then we propose to start the algorithm with a small number of attribute vectors as W and then increment it as the computation goes on. Moreover the fact that this formulation only requires the matrix K will enable the application of kernel technique to the problem. The sub-problem to solve is

$$(\bar{H}(W)) \quad \left| \begin{array}{l} \text{minimize } \lambda_W^\top K_W \lambda_W \\ \text{subject to } p_{\ell_i} + 1 \le (k_W^i)^\top \lambda_W \le p_{\ell_i+1} - 1 \text{ for } i \in W, \end{array} \right.$$

where $(k_W^i)^\top$ is the row vector consisting $(x^i)^\top x^j$ for $j \in W$. Note that the dimension of λ_W varies when the size of W changes as the computation goes on.

Algorithm RC$\bar{\mathbf{H}}$ (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 λ_W^ν and p^ν.
Step 3 : Let $\Delta^\nu = \{\, i \in N \setminus W^\nu \mid (\lambda_W^\nu, p^\nu)$ violates $p_{\ell_i} + 1 \le (k_W^i)^\top \lambda_W \le p_{\ell_i+1} - 1 \,\}$.
Step 4 : If $\Delta^\nu = \emptyset$, terminate.
Step 5 : Otherwise let $W^{\nu+1} = W^\nu \cup \Delta^\nu$, increment ν by 1 and go to Step 2.

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

Lemma 2. *Let $(\hat{\lambda}_W, \hat{p}) \in \mathbb{R}^{|W|+l}$ be an optimum solution of $(\bar{H}(W))$. If*

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

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

The validity of the algorithm follows from the above lemma.

Theorem 1. *The Algorithm RC$\bar{H}$ solves problem $(\bar{H})$.*

3 Kernel Technique for Hard Margin Problem

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

Let $\phi : \mathbb{R}^m \to \mathbb{F}$ 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 $\kappa(x,y) = \langle \phi(x), \phi(y) \rangle$ holds for $x, y \in \mathbb{R}^m$, where $\langle \cdot, \cdot \rangle$ is the inner product defined on $\mathbb{F}$. In the sequel we denote $\tilde{x} = \phi(x)$. The kernel technique considers the vectors $\tilde{x}^i \in \mathbb{F}$

instead of $x^i \in \mathbb{R}^m$, and finds the normal vector $\tilde{w} \in \mathbb{F}$ and thresholds $p_1, \ldots, p_l$. Therefore the matrices X and K should be replaced by $\tilde{X}$ composed of vectors $\tilde{x}^i$ 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 $\tilde{K} = [\kappa(x^i, x^j)]_{i,j \in N}$ by the kernel function κ. The problem to solve is

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

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

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

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

Algorithm RC$\tilde{\mathbf{H}}$ (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 λ_W^ν and p^ν.
Step 3 : Let $\Delta^\nu = \{\, i \in N \setminus W^\nu \mid (\lambda_W^\nu, p^\nu)$ violates $p_{\ell_i} + 1 \leq (\tilde{k}_W^i)^\top \lambda_W \leq p_{\ell_i+1} - 1 \,\}$.
Step 4 : If $\Delta^\nu = \emptyset$, terminate.
Step 5 : Otherwise let $W^{\nu+1} = W^\nu \cup \Delta^\nu$, increment ν by 1 and go to Step 2.

Theorem 2. *The Algorithm RC$\tilde{H}$ solves problem* $(\tilde{H})$.

4 Soft Margin Problems for Non-Separable Case

Introducing nonnegative slack variables ξ_{-i} and ξ_{+i} for $i \in N$ relaxes the hard margin constraints to *soft margin* constraints:

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

Positive values of variables ξ_{-i} and ξ_{+i} mean misclassification, hence they should be as small as possible. We penalize positive ξ_{-i} and ξ_{+i} by adding $\delta(\xi_-)$ and $\delta(\xi_+)$ to the objective function via some penalty function δ, where $\xi_- = (\xi_{-i})_{i \in N}$ and $\xi_+ = (\xi_{+i})_{i \in N}$. Then we have the following *primal soft margin problem*.

$$(S) \quad \left| \begin{array}{l} \text{minimize } \|w\|^2 + c\,(\delta(\xi_-) + \delta(\xi_+)) \\ \text{subject to } p_{\ell_i} + 1 - \xi_{-i} \leq (x^i)^\top w \leq p_{\ell_i+1} - 1 + \xi_{+i} \text{ for } i \in N \\ \qquad\qquad \xi_-, \xi_+ \geq 0_n, \end{array} \right.$$

where c is a penalty parameter. When 1-norm function (resp., 2-norm function) is employed as the function δ, we call the above problem *soft margin problem with*

1-norm penalty (resp., 2-norm penalty). As we discussed in the previous section, we can replace $\|w\|^2$ and $(x^i)^\top w$ in the problem (S) by $\lambda^\top K\lambda$ and $(k^i)^\top\lambda$ to obtain the primal problem with dual representation of normal vector. Then we have

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

The sub-problem $(\bar{S}(W))$ for the working set W will be

$$(\bar{S}(W)) \quad \left| \begin{array}{l} \text{minimize } \lambda_W^\top K_W\lambda_W + c\,(\delta(\xi_{-W})+\delta(\xi_{+W})) \\ \text{subject to } p_{\ell_i}+1-\xi_{-i} \le (k_W^i)^\top\lambda \le p_{\ell_i+1}-1+\xi_{+i} \text{ for } i\in W \\ \qquad\qquad \xi_{-W},\xi_{+W} \ge 0_{|W|}, \end{array} \right.$$

where $\xi_{-W}=(\xi_{-i})_{i\in W}$ and $\xi_{+W}=(\xi_{+i})_{i\in W}$.

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

Step 1 : Let W^0 be an initial working set, and let $\nu=0$.
Step 2 : Solve $(\bar{S}(W^\nu))$ to obtain $(\lambda_W^\nu, p^\nu, \xi_{-W}^\nu, \xi_{+W}^\nu)$.
Step 3 : Let $\Delta^\nu = \{\, i\in N\setminus W^\nu \mid (\lambda_W^\nu, p^\nu)$ violates $p_{\ell_i}+1 \le (k_W^i)^\top\lambda_W \le p_{\ell_i+1}-1\,\}$.
Step 4 : If $\Delta^\nu=\emptyset$, terminate.
Step 5 : Otherwise let $W^{\nu+1}=W^\nu\cup\Delta^\nu$, increment ν by 1 and go to Step 2.

Lemma 3. *Let $(\hat{\lambda}_W, \hat{p}, \hat{\xi}_{-W}, \hat{\xi}_{+W})$ be an optimum solution of $(\bar{S}(W))$. If*

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

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

Theorem 3. *The Algorithm RC$\bar{S}$ solves problem $(\bar{S})$.*

Since the kernel technique can apply to the soft margin problem in the same way as discussed in Section 3, we omit the kernel version of soft margin problem.

5 Illustrative Example and Conclusion

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) with 1-norm penalty is given in Fig. 1. We set the parameter c to 15.

Using the Gaussian kernel defined as $\kappa(x,y) = \exp(-\|x-y\|^2/2\sigma^2)$, we solved $(\tilde{H})$. The result with $\sigma = 7$ is given in Fig. 2, where one can observe that the problem $(\tilde{H})$ is exposed to the risk of over-fitting. Other kernel functions with a combination of various parameter values should be tested.

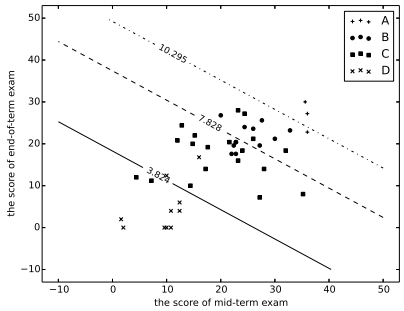

Fig. 1 Classification by (S) with 1-norm penalty

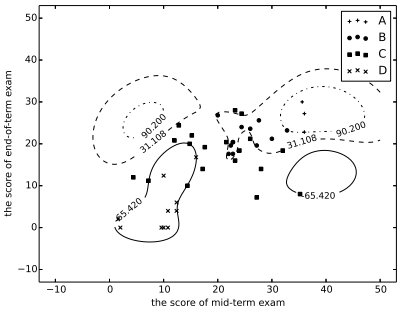

Fig. 2 Classification by $(\tilde{H})$

In this paper, we considered the ranking problem and proposed a row and column generation algorithm to alleviate the computational burden. Furthermore we proved the validity 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. 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.
6. 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.
7. J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*, Cambridge University Press, Cambridge, 2004.
8. 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.
9. V.N. Vapnik, *Statistical Learning Theory*, John-Wiley & Sons, New York, 1998.