Extracting Feature Subspace for Kernel Based Support Vector Machines

01703730 Tokyo Institute of Technology Yasutoshi YAJIMA
Tokyo Institute of Technology Hiroko OHI
01601360 Tokyo Institute of Technology Masao MORI

1 Introduction

We are going to propose linear programming formulations of support vector machines (SVMs) for generating kernel based nonlinear discriminate functions defined in a feature space $\mathcal{F}$ characterized by a kernel function. Unlike the standard SVMs using the quadratic programs, our approach explores a fairly small dimensional subspace of $\mathcal{F}$ to construct the nonlinear discriminator, which allows us to employ a small sized linear programming problem. We will demonstrate that an orthonormal basis of the subspace can be implicitly given by eigenvectors of the Gram matrix defined by the associated kernel function, and a linear programming formulation is successfully introduced. Moreover, when the number of given date points is extremely large, the subspace can be extracted by a number of the randomly selected data points. Numerical experiments are also included, which indicate that the subspace generated by less than 2% of the entire training data points achieves the reasonable performance for a huge datasets with 60000 data points.

2 Nonlinear Discrimination by Kernels

Let us first consider the standard Wolfe dual formulation for SVMs below. We assume that data points are represented by an $M \times N$ matrix A, $Q = AA^T$,and C_0 is a given positive parameter.

let $\mathcal{K}(x, x')$ denote the kernel function which gives inner product of $\phi(x)$ and $\phi(x')$ in $\mathcal{F}$. The following quadratic programming problem has been explored for obtaining nonlinear discriminate functions.

$$(2.1) \quad \left| \begin{array}{ll} \text{Maximize} & -\frac{1}{2}\alpha^T Y \mathcal{K} Y \alpha + e^T \alpha \\ \text{Subject to} & y^T \alpha = 0, \\ & 0 \le \alpha \le C_0 e. \end{array} \right.$$

$\beta = (\beta_1, \beta_2, \ldots, \beta_M)^T$ and let $\beta_j = y_j \alpha_j / C_0$, $j = 1, 2, \ldots, M$, then we write the problem (2.1) equivalently as follows:

In this paper, let us introduce a formulation for the kernel based nonlinear discriminate functions based on variants of the dual forms. let us consider the problem with the square of the 2-norm capacity constraint defined below:

$$(2.2) \quad \left| \begin{array}{ll} \text{Maximize} & y^T \beta \\ \text{Subject to} & \|A^T \beta\|_2^2 \le C, \\ & e^T \beta = 0, \\ & 0 \le Y\beta \le e. \end{array} \right.$$

Note that the quadratic constraint can be written as

$$(2.3) \quad \|A^T \beta\|_2^2 = \beta^T Q \beta \le C.$$

Then, replacing Q in (2.3) with the Gram matrix $\mathcal{K}$ defined by the inner product $\mathcal{K}(\cdot, \cdot)$, let us introduce the following problem:

$$(2.4) \quad \left| \begin{array}{ll} \text{Maximize} & y^T \beta \\ \text{Subject to} & \beta^T \mathcal{K} \beta \le C, \\ & e^T \beta = 0, \\ & 0 \le Y\beta \le e. \end{array} \right.$$

Lemma 2.1 *Under a suitable choice of the parameter C, the problem (2.4) generates any optimal solutions of the problem (2.1) with the parameter C_0.*

3 Linear Programming Formulations for Kernel SVM

Let $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_{M'} > 0$ be positive eigenvalues of the matrix $\mathcal{K}$ and $d'_1, d'_2, \ldots, d'_{M'} \in R^M$ be the associated eigenvectors normalized to unit length. Also, let us define $D = [d_1, d_2, \ldots, d_{M'}]$, where $d_i = \sqrt{\lambda_i} d'_i$, $i = 1, 2, \ldots, M'$.

$$\beta^T \mathcal{K} \beta = \left\|D^T \beta\right\|_2^2 \le C.$$

We will introduce an approximation for this quadratic constraint. To this end, we consider the largest $S \ll M$ positive eigenvalues $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_S$, and the associated column vectors of D.

$$\beta^T \mathcal{K} \beta \approx \beta^T D_S D_S^T \beta = \left\|D_S^T \beta\right\|_2^2.$$

the following formulations:

$$(3.5) \quad \left| \begin{array}{ll} \text{Maximize} & y^T \beta \\ \text{Subject to} & \left\|D_S^T \beta\right\|_\infty \le C, \\ & e^T \beta = 0, \\ & 0 \le Y\beta \le e, \end{array} \right.$$

by the linear programming problem.

The primal form of the linear programming formulation corresponding to the problem (3.5) can be explicitly described as follows:

$$(3.6)\quad \left|\begin{array}{ll} \text{Minimize} & C\,\|w_S\|_1 + e^T\xi \\ \text{Subject to} & Y\,(D_S w_S - b_S e) + \xi \geq e, \\ & \xi \geq 0, \end{array}\right.$$

where $w_S \in R^S$ and $b_S \in R^1$ are primal variables. Let us denote an optimal solution of (3.6) as $(w_S, b_S) = (w_S^*, b_S^*)$, which implies that we have obtained an optimal linear discriminate function as

$$f(x_S) = {w_S^*}^T x_S + b_S^*,$$

where x_S is an S dimensional variables.

3.1 Extracting the Subspace of $\mathcal{F}$

Let d_{jk} denote the $j-k$ elements of the matrix D_S. Also, associated with the k-th column vector $d_k = (\,d_{1k}\, d_{2k} \cdots d_{Mk}\,)^T$ of D_S, let us define vectors in $\mathcal{F}$ as follows:

$$(3.7)\quad \mathcal{V}_k = \frac{\sum_{j=1}^{M} d_{jk}\phi(A_j^T)}{\lambda_k}, \quad k = 1, 2, \ldots, S.$$

Note that each vector $\mathcal{V}_k$ can not be expressed since $\phi(A_j^T)$ is not described explicity.

The following lemma holds.

Lemma 3.2 *The set of vectors* $\{\,\mathcal{V}_1, \mathcal{V}_2, \ldots, \mathcal{V}_S\,\}$ *satisfies*

$$\langle\, \mathcal{V}_k\,,\, \mathcal{V}_{k'}\,\rangle = \begin{cases} 0 & \textit{if } k \neq k', \\ 1 & \textit{o.w}, \end{cases}$$

where $\langle\,\cdot\,,\,\cdot\,\rangle$ *denotes the inner product defined in* $\mathcal{F}$.

Therefore, it follows from Lemma 3.2 that the set of vectors

$$\{\mathcal{V}_1, \mathcal{V}_2, \ldots, \mathcal{V}_S\}$$

constitutes an orthonormal basis of the S dimensional subspace of $\mathcal{F}$ which will be denoted by $\mathcal{F}_S$.

For any point $x \in R^N$, let us denote the S dimensional coordinate vector of the projection of $\phi(x)$ onto $\mathcal{F}_S$ with respect to the basis $\mathcal{V} = \{\mathcal{V}_1, \mathcal{V}_2, \ldots, \mathcal{V}_S\}$ as $[x]_{\mathcal{V}}$, i.e.,

$$[x]_{\mathcal{V}} = \begin{pmatrix} \langle\, \phi(x)\,,\, \mathcal{V}_1\,\rangle \\ \langle\, \phi(x)\,,\, \mathcal{V}_2\,\rangle \\ \cdots \\ \langle\, \phi(x)\,,\, \mathcal{V}_S\,\rangle \end{pmatrix} \in R^S,$$

which can be explicity described.

Lemma 3.3 *Let* $\mathcal{V} = \{\mathcal{V}_1,\ \mathcal{V}_2,\ \ldots,\ \mathcal{V}_S\}$ *be an orthonormal basis defined in (3.7), then*

$$D_{S_j}^T = [A_j^T]_{\mathcal{V}}, \quad j = 1, 2, \ldots, M,$$

where D_{S_j} *denotes the* j*-th row vector of* D_S.

Recall that the optimal solution (w_S^*, b_S^*) of the problem (3.6) generates the discriminate function.

$$f(x_S) = {w_S^*}^T x + b_S^*.$$

Note that this linear function is defined in the S dimensional space. Let us now consider classifying an arbitrary N dimensional data point by this function. To this end, we need to calculate the coordinate vector $[x]_{\mathcal{V}}$. As we have seen above, the k-th element of the vector $[x]_{\mathcal{V}}$ is given by the projection onto $\mathcal{V}_k$, that is $\langle\, x\,,\, \mathcal{V}_k\,\rangle$. Substituting (3.7), we have

$$\langle\, x\,,\, \mathcal{V}_k\,\rangle = \frac{\sum_{j=1}^{M} d_{jk}\mathcal{K}\,(\,x\,,\, A_j^T\,)}{\lambda_k}, \quad k = 1, 2, \ldots, S.$$

Here, it should be emphasized that each element of the vector $[x]_{\mathcal{V}}$ is explicitly calculated without knowing the vectors $\mathcal{V}_k$, $k = 1, 2, \ldots, S$. Then, one can classify the point $x \in R^N$ according to the sign of

$$[x]_{\mathcal{V}}^T w_S^* + b_S^*,$$

which is also calculated, explicitly.

3.2 Sampling Procedures

Furthermore, when the number of points, M, is extremely huge, the considerable amount of computational work would be required for obtaining the largest S eigenvectors of the $M \times M$ matrix $\mathcal{K}$. To avoid this computational difficulties one can choose L sample points, where $L \ll M$ and extract an orthonormal basis of the subspace of $\mathcal{F}$. Let us assume that, for simplicity, the L sample points correspond to the first L rows of the matrix A, and that, associated with the sample points, the matrix A and the Gram matrix $\mathcal{K}$ are partitioned as follows:

$$A = \begin{bmatrix} A^L \\ A' \end{bmatrix}, \quad \text{and} \quad \mathcal{K} = \begin{bmatrix} \mathcal{K}^L & {\mathcal{K}'}^T \\ \mathcal{K}' & \mathcal{K}'' \end{bmatrix},$$

where $A^L \in R^{L\times N}$, $\mathcal{K}^L \in R^{L\times L}$, and $\mathcal{K}' \in R^{(M-L)\times L}$. Thus, the inequalities corresponding to the norm constraints in (3.5) and (??) should be

$$(3.8)\quad \left\| \begin{bmatrix} D_S^L \\ \mathcal{K}' D_S^L \Lambda^{L^{-1}} \end{bmatrix}^T \beta \right\|_p \leq Ce,$$

where p is ∞ and 1, respectively.

It is worth noting that in our sample scheme, L sample points are used only for extracting the basis of the S dimensional subspace $\mathcal{F}_S$, and that the all M samples are involved in the linear programs.

References

[1] O. L. Mangasarian, *Nonlinear Programming*, vol. 10 of Classics in Applied Mathematics, SIAM, Philadelphia, PA, 1994.

[2] B. Schölkopf, A. J. Smola, and K. Müeller, *Kernel principal component analysis*, in Advances in Kernel Methods, B. Schölkopf, C. Burges, and A. Smola, eds., The MIT Press, 1999, pp. 327–352.