Training SVM with Indefinite Kernels

Jianhui Chen and Jieping Ye

Department of Computer Science and Engineering
Arizona State University

Introduction - I

- Support Vector Machines (SVMs) have been applied successfully in many domains.
  - Positive semidefinite (PSD) kernel matrix.
- The PSD property of the kernel matrix ensures the existence of Reproducing Kernel Hilbert Space (RHKS) and leads to the following convex quadratic program (QP):

$$\begin{aligned} \max_{\alpha} \quad & \alpha^T e - \frac{1}{2}\alpha^T YKY\alpha \\ \text{subject to} \quad & \alpha^T y = 0, \ 0 \leq \alpha \leq C. \end{aligned} \tag{1}$$

- The globally optimal solution can be obtained via the standard optimization techniques, such as primal-dual interior point methods.

Introduction - II

- Many applications may generate non-PSD similarity matrices (indefinite kernels):
  - Sequence similarity based on pairwise alignment.
- Common approaches for fitting indefinite kernels into SVM:
  - Modify the SVM formulation
    - The existence of RHKS or general representer theorem. (Lin & Lin, 2003; Haasdonk, 2005; Ong *et al.*, 2004)
  - Generate PSD kernels from indefinite kernels
    - Apply spectrum transformations (denoise, flip, diffusion, shift). (Wu *et al.*, 2005)
    - Result in the loss of valuable information.

Main Contributions

- Propose a semi-infinite quadratically constrained linear program (SIQLP) formulation for training SVM with indefinite kernels, in which the indefinite kernel matrix is treated as a noisy observation of some unknown positive semidefinite one (proxy kernel).
- Propose an iterative algorithm for solving the SIQLP formulation, and conduct the convergence analysis.
- Propose to employ an additional pruning strategy to improve the efficiency of the proposed iterative algorithm.
- Analyze the close relationship between the proposed SIQLP formulation and multiple kernel learning.

Problem Formulation - I

- Assume $K \in \mathbb{R}^{n \times n}$ is a valid kernel matrix. Let $y \in \mathbb{R}^n$ be the vector of class labels and $Y = \mathrm{diag}(y)$. The dual formulation of 1-norm soft margin SVM classification is given by

$$\begin{aligned} \max_{\alpha} \quad & \alpha^T e - \tfrac{1}{2}\alpha^T YKY\alpha \\ \text{subject to} \quad & \alpha^T y = 0,\ 0 \le \alpha \le C. \end{aligned} \tag{2}$$

- For a given indefinite kernel matrix K_0, we consider a regularized SVM formulation, in which the indefinite kernel is treated as a noisy observation of some unknown PSD kernel (proxy kernel) (Luss and d'Aspremont, 2007).

$$\begin{aligned} \max_{\alpha} \min_{K} \quad & \alpha^T e - \tfrac{1}{2}\alpha^T YKY\alpha + \rho\|K - K_0\|_F^2 \\ \text{subject to} \quad & \alpha^T y = 0,\ 0 \le \alpha \le C,\ K \succeq 0. \end{aligned} \tag{3}$$

Problem Formulation - II

$$\begin{aligned} \max_{\alpha} \min_{K} \quad & \alpha^T e - \tfrac{1}{2}\alpha^T YKY\alpha + \rho\|K - K_0\|_F^2 \\ \text{subject to} \quad & \alpha^T y = 0,\ 0 \le \alpha \le C,\ K \succeq 0. \end{aligned} \tag{4}$$

- Denote the objective function in Eq. (4) as

$$S(\alpha, K) = \alpha^T e - \frac{1}{2}\alpha^T YKY\alpha + \rho\|K - K_0\|_F^2. \tag{5}$$

- Reformulate Eq. (3) as a semi-infinite quadratically constrained linear program (SIQLP) as follows:

$$\begin{aligned} \max_{\alpha, t} \quad & t \\ \text{subject to} \quad & \alpha^T y = 0,\ 0 \le \alpha \le C \\ & t \le S(\alpha, K),\ \forall K \succeq 0. \end{aligned} \tag{6}$$

Proposed Algorithm

- We propose to solve Eq. (6) via an iterative algorithm, in which an additional constraint based on a kernel matrix K is added in each iteration.
  - It is guaranteed to converge to a globally optimal solution.
- We call the finite set of kernel matrices $\mathbf{K} = \{K_i\}_{i=1}^{p}$ as a localization set of the infinite many quadratic constraints in Eq. (6), and the suboptimal t and α to Eq. (6) based on $\mathbf{K}$ as the intermediate solution pair.
- The proposed iterative algorithm consists of two steps:
  - Step 1: Update the intermediate solution pair (t and α).
  - Step 2: Update the localization set $\mathbf{K}$.

Proposed Algorithm - Step 1

- Given an intermediate solution pair (t and α), the localization set can be updated by solving

$$\min_K S(\alpha, K) = \min_K \rho \|K - K_0\|_F^2 - \frac{1}{2}\alpha^T YKY\alpha. \tag{7}$$

  - The optimal K^* is given by (Luss and d'Aspremont, 2007)

$$K^* = \left(K_0 + Y\alpha\alpha^T Y/(4\rho)\right)_+, \tag{8}$$

    where X_+ denotes the positive part of a symmetric matrix X.
  - If the optimal K^* to Eq. (7) satisfies $t \leq S(\alpha, K^*)$, the current intermediate solution is globally optimal to the SIQLP problem in Eq. (6); otherwise the optimal K^* is added into the localization set $\mathbf{K}$.

Proposed Algorithm - Step 2

- Given any localization set $\mathbf{K} = \{K_i\}_{i=1}^{p}$, the intermediate solution pair (t and α) can be computed by solving

$$\begin{aligned} \max_{\alpha, t} \quad & t \\ \text{subject to} \quad & \alpha^T y = 0,\ 0 \leq \alpha \leq C \\ & t \leq S(\alpha, K_i),\ \ i = 1, \cdots, p. \end{aligned} \tag{9}$$

  - The problem in Eq. (9) is called the restricted master problem. It corresponds to a quadratically constrained linear program (QCLP), which can be solved via general purpose optimization solvers, such as MOSEK.
  - The complexity of QCLP depends on the number of quadratic constraints, i.e., the size of the localization set $\mathbf{K}$.

Convergence Analysis: Lower and Upper Bounds

- The proposed iterative algorithm alternates between updating the localization set $\mathbf{K}$ (step 1) and updating the intermediate solution pair (t and α) (step 2).
- At step 1, we compute the new kernel matrix K_i by

$$S(\alpha_{i-1}, K_i) = \min_{K \succeq 0} S(\alpha_{i-1}, K). \tag{10}$$

We denote the Lower Bound as

$$l_i^- = \max_j d_j = \max_j S(\alpha_{j-1}, K_j), \quad j = 1, \cdots, i. \tag{11}$$

- At step 2, we compute the Upper Bound as

$$u_i^+ = t_i = \max_{\alpha} \min_{K \in \mathbf{K}} S(\alpha, K) = \min_{K \in \mathbf{K}} S(\alpha_i, K), \tag{12}$$

where $\alpha_i = \arg\max_\alpha \left(\min_{k \in \mathbf{K}} S(\alpha, K)\right)$.

Global Convergence Property

Theorem

Let l_i^- and u_i^+ be defined in Eq. (11) and Eq. (12), respectively. Let (α^*, t^*) be the optimal solution pair to Eq. (6). Then

$$u_i^+ \geq t^* \geq l_i^-. \tag{13}$$

Moreover, the sequence $\{u_i^+\}$ is monotonically decreasing, and the sequences $\{l_i^-\}$ is monotonically non-decreasing.

- We can use the gap between u_i^+ and l_i^- to trace the global convergence of the proposed algorithm.
- When the gap is smaller than a pre-specified value, we stop the iterative algorithm.

Pruning Strategy - I

- Limitation of the proposed iterative algorithm:

  1. The second step of the iterative algorithm is a QCLP problem.
  2. The complexity of solving a QCLP grows with the number of quadratic constraints.
  3. The number of quadratic constraints in QCLP increases by one at each iteration.

- Pruning strategy:

  We propose to prune the inactive constraints at each iteration.

Pruning Strategy - II

- Pruning Strategy:

  1. Let $\mathbf{K}^i = \{K_j\}_{j=1}^p$ be the localization set, and let (t_i, α_i) be the intermediate solution pair. We partition K^i as follows:

  $$\mathbf{K}^i = \mathbf{K}^i_{act} \cup \mathbf{K}^i_{int} \tag{14}$$

  where $\mathbf{K}^i_{act} = \{K | t_i = S(\alpha_i, K)\}$ and $\mathbf{K}^i_{int} = \{K | t_i < S(\alpha_i, K)\}$.

  2. Let K^* be the new kernel matrix. We propose to update the localization set as follows:

  $$\mathbf{K}^{i+1} = \mathbf{K}^i_{act} \cup K^*. \tag{15}$$

  Note that without pruning strategy, we apply $\mathbf{K}^{i+1} = \mathbf{K}^i \cup K^*$.

- The pruning strategy improves the computational efficiency while retaining the convergence property of the algorithm.

Relationship with Multiple Kernel Learning

- Lanckriet et al. (2004) propose to learn an optimal convex combination of a set of p pre-specified kernels $\{K_i\}_{i=1}^p$:

$$\min_{\{\theta_i\}} \max_{\alpha} \quad \alpha^T e - \frac{1}{2}\alpha^T Y \left(\sum_{i=1}^p \theta_i K_i\right) Y\alpha$$
$$\text{subject to} \quad \sum_{i=1}^p \theta_i \operatorname{tr}(K_i) = 1, \alpha^T y = 0, 0 \le \alpha \le C. \quad (16)$$

- Let K_0 be an indefinite kernel. Denote $u_i = \|K_i - K_0\|_F^2$, where u_i measures the distance between K_i and K_0 $(i = 1, \cdots, p)$.
- We consider a regularized version of Eq. (16) as follows:

$$\min_{\{\theta_i\}} \max_{\alpha} \quad \alpha^T e - \frac{1}{2}\alpha^T Y \left(\sum_{i=1}^p \theta_i \, K_i\right) Y\alpha + \rho \sum_{i=1}^p \theta_i \mu_i$$
$$\text{subject to} \quad \sum_{i=1}^p \theta_i = 1, \quad \alpha^T y = 0, \quad 0 \le \alpha \le C, \quad (17)$$

- It is equivalent to the proposed SIQLP formulation in Eq. (6).

Experimental Setup

- We evaluate the convergence property of the proposed algorithms, and compare them with other representative ones.
- We construct indefinite kernels through perturbation:
  1. Generate a set of Gaussian kernels and choose the best one via cross-validation in terms of classification accuracy.
  2. Generate a (perturbed) matrix E with zero mean and identity covariance matrix, and apply $\xi(E+E^T)/2$ as the perturbation.
- The parameters C and ρ in SIQLP are determined via cross-validation.
- The reported classification accuracy is averaged over 10 random partitions of the data into a training set and a test set with a ratio $4:1$.

Global Convergence

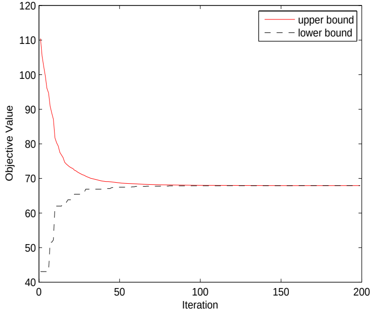

Figure: Convergence of the algorithm without pruning.

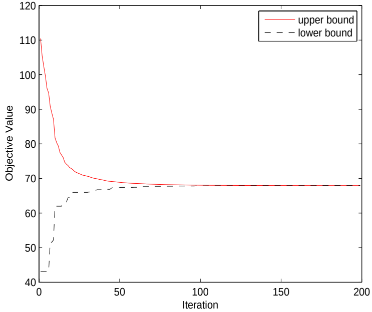

Figure: Convergence of the algorithm with pruning.

- Upper bound (u_i^+) monotonically decreases while lower bound (l_i^-) monotonically increases, and they approach each other gradually.
- The proposed algorithms with or without pruning strategy result in a similar convergent rate.

Size of the Localization Set K

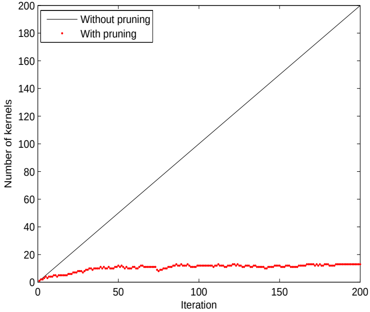

Figure: Number of kernels involved in the algorithms with and without pruning.

- With pruning strategy (red), the number of kernels involved in the iterative algorithm stabilizes around a smaller number.
- Without pruning strategy (black), the number of kernels increases gradually.
- For both cases, it takes around 150 iterations for convergence.

Classification Performance

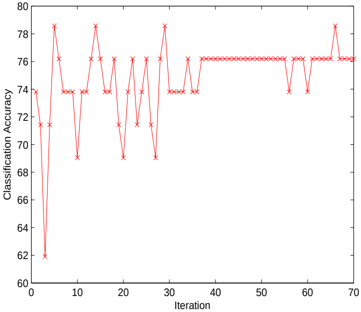

Figure: Classification performance of the proposed algorithm with pruning.

- There is a large variation of classification accuracy at the first few iterations.
- The accuracy becomes stable after 40 iterations (around 76%).
- We may apply early-stopping strategy.

Comparative Study

Table: Comparison of the proposed algorithm with other representative ones in terms of classification accuracy.

Data Set	Size	λ^- num.	λ^+ num.	λ_{min}	λ_{max}	$\lvert\frac{\lambda_{max}}{\lambda_{min}}\rvert$	Denoise	Flip	Shift	SVM	Indef. SVM
Sonar	208	57.41	150.62	-1.36	18.42	13.55	78.57	79.52	78.10	72.86	80.95
Ionosphere	351	169.62	181.45	-25.50	94.49	3.71	75.57	71.43	71.41	68.00	77.43
B. Cancer	683	323.21	359.82	-3.51	390.52	111.26	95.38	95.62	95.38	89.54	95.36
Heart	270	125.57	144.71	-10.96	42.93	3.92	71.02	67.28	65.42	65.43	72.22
USPS-3-5	1200	520.12	680.31	-3.54	81.99	23.16	96.25	96.88	95.63	96.11	96.81
Diabetes	768	381.22	385.18	-3.93	8.13	2.06	68.83	64.28	62.98	66.23	70.08

- Indefinite SVM (Indef. SVM) is competitive with all other algorithms (Denoise, Flip, Shift, and SVM using indefinite kernel).
- Indef. SVM outperforms other algorithms, when the perturbed kernel matrix has a relative small $\lvert\frac{\lambda_{max}}{\lambda_{min}}\rvert$, where the indefinite kernel matrices are highly non-PSD.

Conclusion and Future Work

Conclusion:

- Propose an SIQLP formulation for training SVM with indefinite kernels.
- Propose an iterative algorithm for solving the SIQLP formulation and conduct a convergence analysis.
- Propose a pruning strategy for improving the efficiency while retaining the convergence property.
- Analyze the close relationship between the proposed formulation and multiple kernel learning.

Future Work:

- Employ alternative optimization techniques.
- Apply the algorithm to real-world applications involving indefinite similarity matrices.

Acknowledgements

This research is in part supported by:

- Arizona State University
- National Science Foundation Grant IIS-0612069

Qustions?

Thank You!