Main contribution

- Proposed a refined Nyström based kernel SVM
- Developed a two-step pipeline that firstly solves a sparse-regularized dual formulation with the approximated kernel and then utilizes the obtained dual solution to retrain a refined Nyström based kernel classifier.
- Justified the proposed approach by a theoretical analysis and extensive empirical studies.

Problem

- Let \( (x_i, y_i), i = 1, 2, \ldots, n \) denote a set of training examples, \( x_i \in \mathbb{R}^d \) and \( y_i \in \{+1, -1\} \)
- Let \( \kappa(\cdot, \cdot) \) denote a valid kernel function and \( \mathcal{H}_k \) denote a Reproducing Kernel Hilbert Space endowed with \( \kappa(\cdot, \cdot) \)
- The kernel SVM is to solve the following optimization problem:

\[
\min_{f \in \mathcal{H}_k} \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i) + \frac{\lambda}{2} ||f||^2_{\mathcal{H}_k}
\]

- Using conjugate function, the above optimization problem can be turned into a dual problem:

\[
\alpha_* = \arg \max_{\alpha \in \mathbb{R}^n} \left( \sum_{i=1}^{n} \langle \alpha, \kappa(x_i, \cdot) \rangle - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j \kappa(x_i, x_j) \right)
\]

- **Key Problem:** when \( n \) is very large, it is prohibitive to compute or maintain kernel matrix \( K \)
- **The goal:** to achieve classification performance as high as Kernel SVM but with computational cost as low as Linear SVM.

Three central questions

- **Q:** How to approximate that large kernel matrix \( K \)?
  - A: Using the Nyström approximation.

- **Q:** How to improve the performance of the learned classifier suffered from the Nyström approximation error?
  - A: Adding \( \ell_1 \) regularization term.

- **Q:** How to further improve the performance of the learned classifier?
  - A: Sampling a new set of landmark points based on that good dual solution for Nyström approximation to re-train a refined classifier.

Nyström approximation

- Let \( K \) be the original kernel matrix, sample \( m \) points from \( n \) training examples, \( K_{mn} \in \mathbb{R}^{m \times m} \) denote the sub-kernel matrix between \( n \) training examples and \( m \) samples and \( K_{nm} \in \mathbb{R}^{mn \times m} \) denote the kernel matrix among \( m \) points and \( K_{nm} \) is the pseudo-inverse of \( K_{nm} \), then the Nyström approximation of \( K \) is:

\[
K \approx K_{nm} K_{mn}^T
\]

Refined Nyström based Kernel SVM — The first step

- Add \( \ell_1 \) regularization term to reduce approximate error brought by Nyström approximation:

\[
m_{ge} \alpha_T - \frac{1}{n} \sum_{i=1}^{n} \ell_T(\alpha_i) - \frac{1}{2m} \alpha^T K \alpha - \frac{\rho}{n} \alpha_i
\]

- Equivalently, the primal form of the above optimization problem is (using hinge loss as an example):

\[
\min_{\alpha \in \mathbb{R}^m} \frac{1}{m} \sum_{i=1}^{m} \max(0, 1 - \langle x, y_i w \rangle) + \frac{\lambda}{2} ||w||^2
\]

- Intuitively, the margin is reduced to \((1 - \tau)\) for hinge loss. Theoretically, we can prove the following theorem:

**Theoretical guarantee for learning a good dual solution**

**Theorem 1:** Assume for some \( k \) and \( \delta \in (0, 1) \) and the following condition hold \( \lambda k + 2\gamma (16s) \geq (6 + \frac{64\delta}{m}) \lambda k + 1 \) and \( m \geq 8k \tau (m, 16s) (16s \log d + \log \frac{1}{\delta}) \), by setting \( \tau \geq \frac{1}{2m} \sum_{i=1}^{m} ||\alpha_i||K_{nm} - K_{nm} \), then, with a probability \( 1 - \delta \), we

\[
||\tilde{\alpha}_i - \alpha_i|| \leq \frac{1.5\sqrt{\delta}}{\lambda k + 2\gamma (16s) - (6 + 64s/m) \lambda k + 1}
\]

Refined Nyström based Kernel SVM—The second step

- Sample a new set of landmark points based on:

\[
P_T (x_i) \text{ is selected} = \frac{||\tilde{\alpha}_i||}{\sum_{i=1}^{m} ||\tilde{\alpha}_i||}
\]

- Construct the Nyström approximation based on this new set of landmark points and re-train the classifier.

Experimental Results

- Figure 1: Test Accuracy for linear SVM, RBF SVM and Nyström based kernel classifier with different number of samples
- Figure 2: Test accuracy of the sparse-regularized Nyström based kernel classifier
- Figure 3: Test accuracy of the refined Nyström based kernel classifier (sp-pro-nys) \( m = 1024 \)
- Figure 4: Training time of linear SVM, Kernel SVM, the standard Nyström based classifier and the refined Nyström based classifier for \( m = 1024 \)