Fast and Accurate Refined Nyström based Kernel SVM

Zhe Li*, Tianbao Yang*, Lijun Zhang♮, Rong Jin†

*The University of Iowa, ♮Nanjing University, †Alibaba Group

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

$$\min_{f \in \mathcal{H}_\kappa} \frac{1}{n}\sum_{i=1}^{n} \ell(f(\mathbf{x}_i), y_i) + \frac{\lambda}{2}\|f\|^2_{\mathcal{H}_\kappa}$$

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

$$\alpha_* = \arg\max_{\alpha \in \Omega^n} -\frac{1}{n}\sum_{i=1}^{n} \ell_i^*(\alpha_i) - \frac{1}{2\lambda n^2}\alpha^T K \alpha$$

- **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.

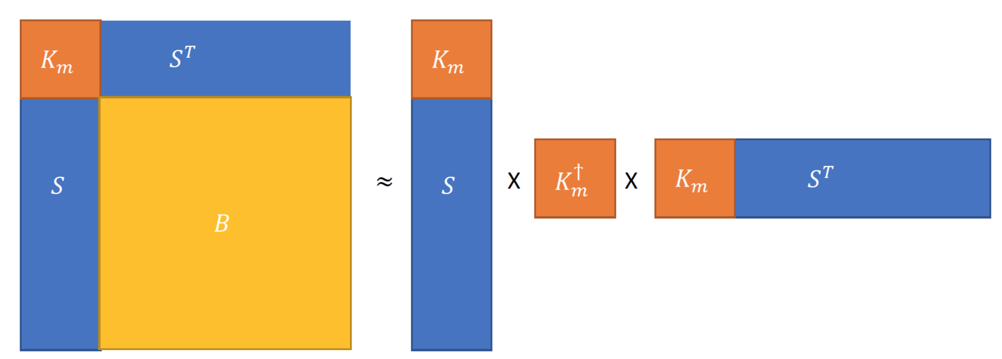
Three central questions

- 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 ℓ_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, $K_b \in \mathbb{R}^{n \times m}$ denote the sub-kernel matrix between n training examples and m samples and $\tilde{K}_m \in \mathbb{R}^{m \times m}$ denote the kernel matrix among m points and $\tilde{K}_m^\dagger$ is the pseudo-inverse of $\tilde{K}_m$, then the Nyström approximation of K is:

$$\hat{K} = K_b \tilde{K}_m^\dagger K_b^T$$



Refined Nyström based Kernel SVM — The first step

- Add ℓ_1 regularization term to reduce approximate error brought by Nyström approximation:

$$\max_{\alpha \in \mathbb{R}^n} -\frac{1}{n}\sum_{i=1}^{n} \ell_i^*(\alpha_i) - \frac{1}{2\lambda n^2}\alpha^T \hat{K} \alpha - \frac{\tau}{n}\|\alpha\|_1$$

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

$$\min_{\mathbf{w} \in \mathbb{R}^m} \frac{1}{n}\sum_{i=1}^{n} \max(0, (1-\tau) - y_i \mathbf{w}^\top \hat{\mathbf{x}}_i) + \frac{\lambda}{2}\|\mathbf{w}\|_2^2$$

- 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 $\lambda\mu + 2\gamma(16s) \geq (6 + \frac{64s}{m})\lambda_{k+1}$ and $m \geq 8k\tau_k(m, 16s)(16s\log d + \log\frac{k}{\delta})$, by setting $\tau \geq \frac{2}{\lambda n}\sum_{i=1}^{n}|[\alpha_i]|\|\hat{K}_{*i} - K_{*i}\|_\infty$, Then, with a probability $1-\delta$, we

$$\|\tilde{\alpha}_* - \alpha_*\| \leq \frac{1.5\lambda\sqrt{s}\tau}{\lambda\mu + 2\gamma(16s) - (6 + 64s/m)\lambda_{k+1}}$$

Refined Nyström based Kernel SVM–The second step

- Sample a new set of landmark points based on:

$$\Pr(\mathbf{x}_i \text{ is selected}) = \frac{|[\tilde{\alpha}_*]_i|}{\sum_{i=1}^{n}|[\tilde{\alpha}_*]_i|}$$

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

Experimental Results

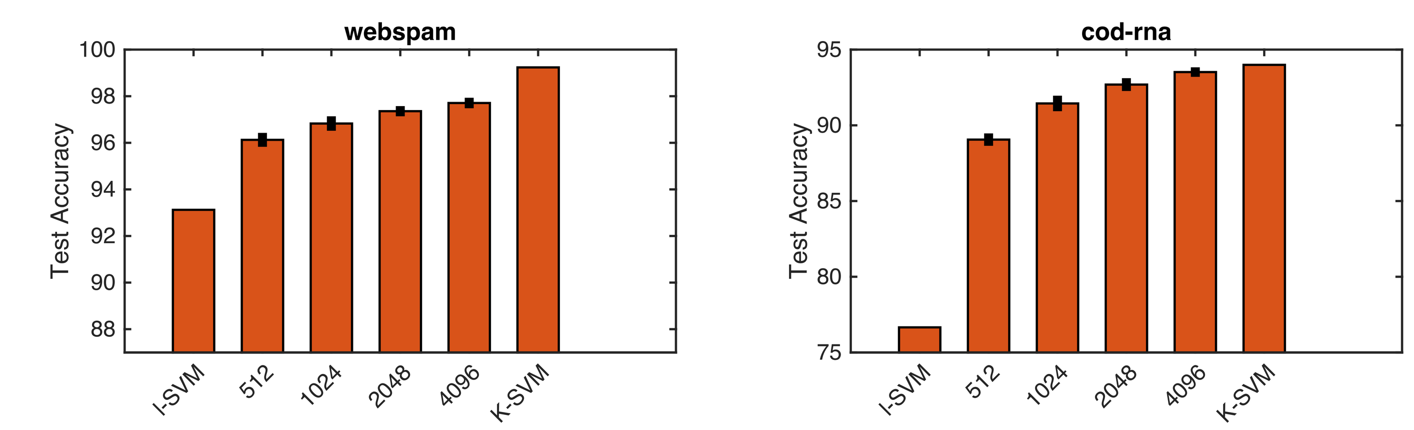


Figure 1: Test Accuracy for linear SVM, RBF SVM and Nyström based kernel classifier with different number of samples

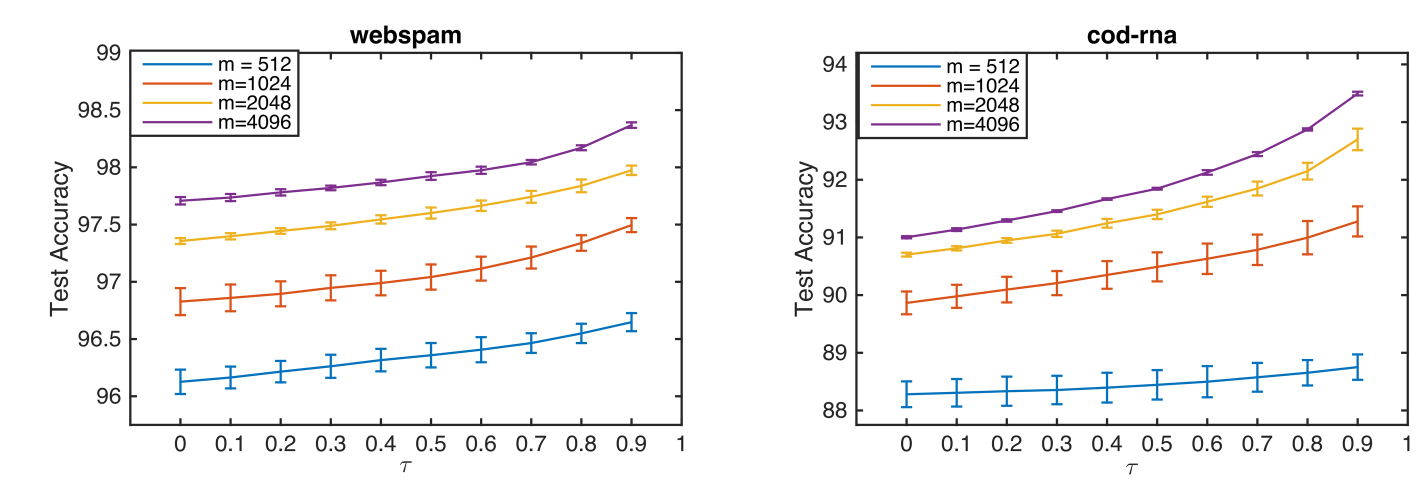


Figure 2: Test accuracy of the sparse-regularized Nyström based kernel classifier

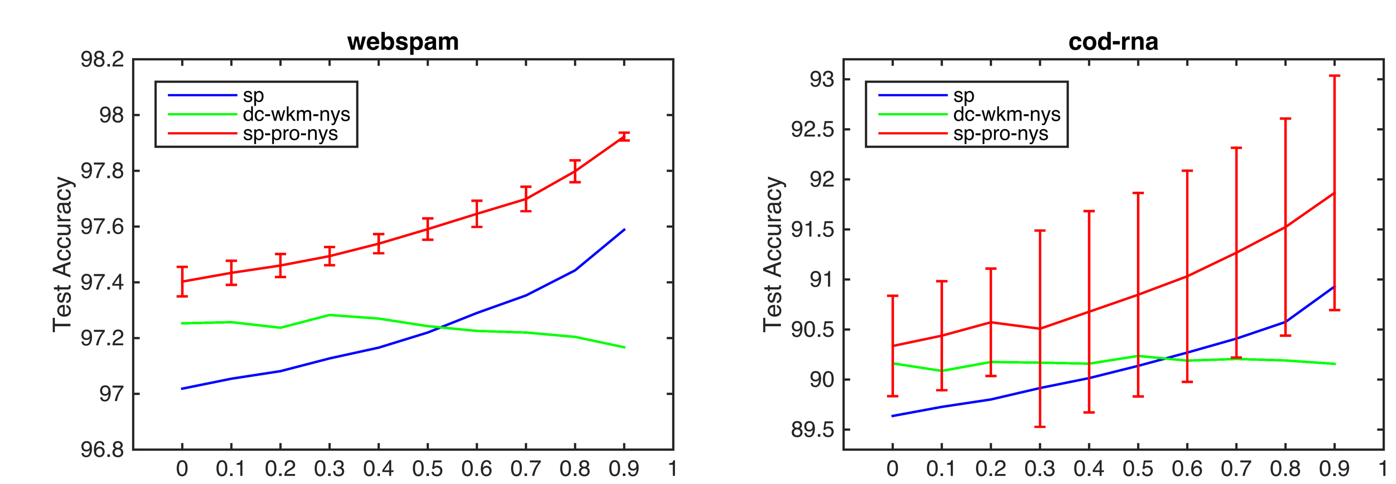


Figure 3: Test accuracy of the refined Nyström based kernel classifier(sp-pro-nys), $m = 1024$

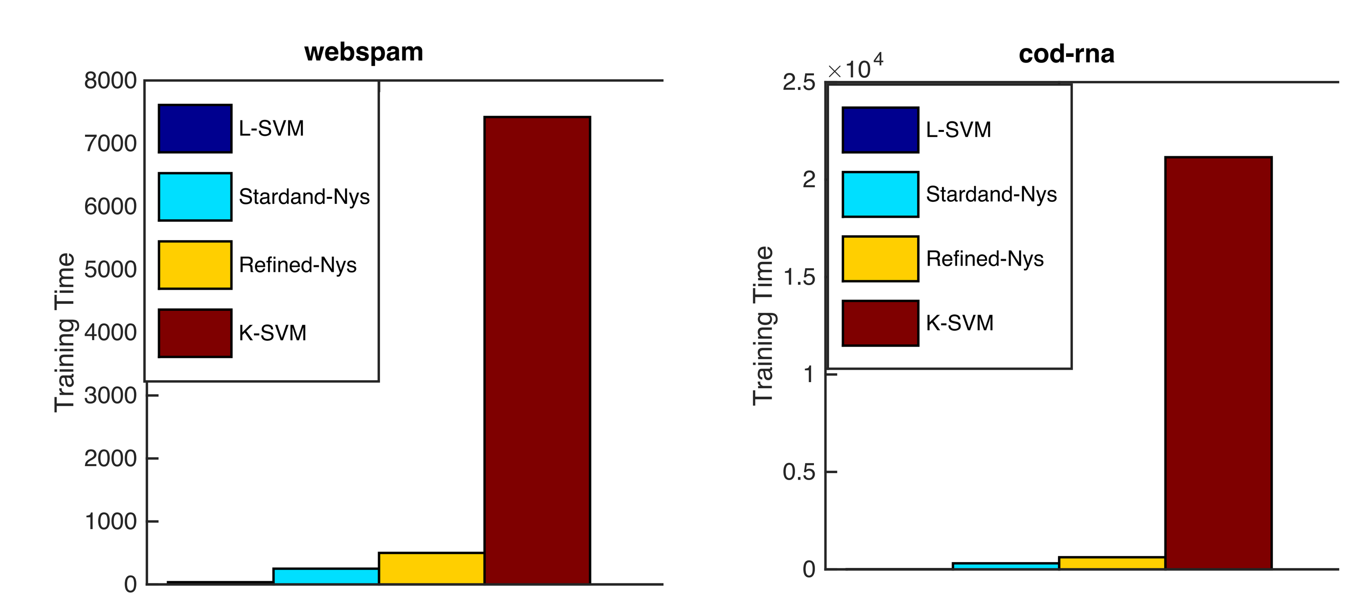


Figure 4: Training time of linear SVM, Kernel SVM, the standard Nyström based classifier and the refined Nyström based classifier for $m = 1024$