Fast and Accurate Refined Nyström Based Kernel SVM

Zhe Li,[1] **Tianbao Yang,**[1] **Lijun Zhang,**[2] and **Rong Jin**[3]
[1]Department of Computer Science, The University of Iowa, Iowa City, IA 52242, USA
[2]National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China
[3]Alibaba Group, Seattle, WA 98101, USA
{zhe-li-1,tianbao-yang}@uiowa.edu, zhanglj@lamda.nju.edu.cn, jinrong.jr@alibaba-inc.com

Abstract

In this paper, we focus on improving the performance of the Nyström based kernel SVM. Although the Nyström approximation has been studied extensively and its application to kernel classification has been exhibited in several studies, there still exists a potentially large gap between the performance of classifier learned with the Nyström approximation and that learned with the original kernel. In this work, we make novel contributions to bridge the gap without increasing the training costs too much by proposing a refined Nyström based kernel classifier. We adopt a two-step approach that in the first step we learn a sufficiently good dual solution and in the second step we use the obtained dual solution to construct a new set of bases for the Nyström approximation to re-train a refined classifier. Our approach towards learning a good dual solution is based on a sparse-regularized dual formulation with the Nyström approximation, which can be solved with the same time complexity as solving the standard formulation. We justify our approach by establishing a theoretical guarantee on the error of the learned dual solution in the first step with respect to the optimal dual solution under appropriate conditions. The experimental results demonstrate that (i) the obtained dual solution by our approach in the first step is closer to the optimal solution and yields improved prediction performance; and (ii) the second step using the obtained dual solution to re-train the model further improves the performance.

Kernel method (Scholkopf and Smola, 2001; Shawe-Taylor and Cristianini, 2004) (e.g., Support Vector Machine (SVM)) is one of the most effective learning methods widely used in classification and regression. Thanks to the kernel trick, low dimension features in the original space are mapped into high dimension features without explicitly computing inner product between high dimensional features. However, as the scale of data continues to grow, the kernel method suffers from the severe problem of computing and maintaining a tremendously large kernel matrix, rendering it prohibitive even impossible to learn a kernel classifier in real applications with big data. To speed up the training of a kernel classifier for big data, several fast kernel approximation methods have been developed, including the Nyström method (Williams and Seeger, 2001; Drineas and Mahoney, 2005) and the random Fourier features (Rahimi and Recht, 2007) among others (Le et al., 2013; Yang et al., 2015b). Recently, the authors in (Yang et al., 2012) studied the two approximation schemes in a unified framework and demonstrated that the Nyström method could achieve a better performance than random Fourier features in certain scenarios (e.g., when there is a large eigen-gap in the kernel matrix (Yang et al., 2012) or the eigen-values follow a power-law distribution (Jin et al., 2013)). In this work, we focus on further improving the performance of the Nyström based kernel classifier with the training time increased by a factor of two.

The Nyström method for approximating a kernel matrix works by constructing a set of bases referred to as landmark points and then constructing an approximation based on the kernel similarities between the landmark points and all data points (including the landmark points). It has been observed that the selection of the landmark points affects the performance of the Nyström method (Kumar et al., 2009a). Nevertheless, there still exists a gap between the performance of the Nyström based kernel classifier and that of the optimal kernel classifier. It remains an important problem to bridge the performance gap while maintaining the efficiency. Recently, there emerges a refined Nyström based kernel SVM that first learns an approximate dual solution close enough to the optimal dual solution to the original kernel SVM and then uses the approximate dual solution to construct a set of landmark points to improve the performance of the Nyström based kernel classifier (Hsieh et al., 2014b).

In this paper, we propose an improved method to obtain a good dual solution in the first step. Our approach is motivated by the fact that the original kernel classifier usually has a small number of support vectors, indicating the optimal dual solution is a sparse vector. However, when exploring a Nyström approximation, the number of support vectors could increase due to that some examples become difficult to be classified, leading to an increased number of support vectors, i.e., a denser dual solution. Therefore, in order to improve the quality of the dual solution, we introduce a sparsity-inducing regularizer into the dual formation defined with the Nyström approximation. We justify the proposed approach by a theoretical analysis of the error bound of the obtained dual solution under the incoherence and restricted eigen-value conditions. Empirically, we observe that the proposed approach achieves better perfor-

mance than the original Nyström based kernel classifier and the refined Nyström based kernel classifier using divide-and-conquer approach for obtaining an approximate dual solution.

The main contributions of the paper are summarized as: (i) we study a refined Nyström based kernel SVM and propose a new pipeline that first solves a sparse-regularized dual formulation with the approximated kernel and then utilizes the obtained dual solution to re-train a refined Nyström based kernel classifier; and (ii) we justify the proposed approach by a theoretical analysis and extensive empirical studies.

Related Work

In this section, we review some related work on approximate kernel methods, the Nyström method, sparse learning and randomized dimensionality reduction.

Due to the exceedingly high cost of computing and maintaining a big kernel matrix for large-scale data, several fast approximate kernel methods have been developed, including random Fourier features (Rahimi and Recht, 2007), Fastfood (Le et al., 2013) and the Nyström method (Drineas and Mahoney, 2005) as representatives. Yang et al. (2012) studied the random Fourier features and the Nyström method in a unified framework from the perspective of functional approximation. They demonstrated that the random Fourier features is equivalent to learning a predictive function using a set of basis functions that are generated independent of the data, while the Nyström method is equivalent to learning a predictive function using a set of data-dependent basis functions.

The Nyström method for approximating a positive semi-definite (PSD) matrix has been studied extensively in recent years (Drineas and Mahoney, 2005; Kumar et al., 2009b; Yang et al., 2012; Zhang et al., 2008; Gittens, 2011; Talwalkar and Rostamizadeh, 2010; Gittens and Mahoney, 2013; Jin et al., 2013). Nevertheless, when employed in kernel methods for classification and regression, there still exits a gap between the performance of the Nyström based kernel classifier and the optimal kernel classifier. Recently, (Hsieh et al., 2014b) proposed a refined Nyström based kernel classifier based on a two-step approach where in the first step an approximate dual solution is learned and in the second step a set of new landmark points are constructed using the approximate dual solution obtained in the first step. Our work differentiates from this work in how to learn an approximate dual solution as described in the introduction.

Sparse learning has been researched tremendously is machine learning and statistics. Almost all existing studies are centered around imposing a sparsity-induced regularizer (e.g., the $_1$ norm) on the model (i.e., the primal solution). In this work, we impose a $_1$ norm on the dual solution motivated by the fact that in kernel SVM many examples could be non-support vectors, indicating their corresponding dual variables are zeros. The most relevant work is presented in (Xu et al., 2015), which studied a sparse kernel regression with the Nyström approximation.

It was brought to our attention that the proposed approach for learning a good dual solution in the first step is similar to a recent work on the dual recovery analysis for randomized dimensionality reduction for solving high-dimensional learning problems (Yang et al., 2015a) , which employed the JL transform to reduce high-dimensional examples into a low-dimensional space, then proposed to solve a sparse-regularized dual formulation. Although the proposed approach share the same insight on the introduced sparsity-inducing regularizer on the dual variables, we emphasize that the present work makes non-trivial contributions in the analysis since the Nyström method is not a JL transform, therefore the analysis in (Yang et al., 2015a) based on the JL lemma can not carry over to the Nyström based kernel method.

The problem and Proposed Algorithm

Preliminaries and Motivation

Let $(\mathbf{x}_i, y_i), i = 1, \cdots, n$ denote a set of training examples, where $\mathbf{x}_i \in \mathbb{R}^d$ denotes the feature vector, and $y_i \in \{+1, -1\}$ denotes the class label. 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:

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

where $\ell(z, y) = \max(0, 1 - yz)^p$ ($p = 1$ or 2) is the hinge loss or the squared hinge loss. Using the convex conjugate function, the above optimization problem can be turned into a dual problem:

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

where Ω^n is the domain of the dual solution, $K \in \mathbb{R}^{n\times n}$ is the kernel matrix, and $\ell^*_i(\alpha_i)$ is the convex conjugate of $\ell(z, y_i)$ in terms of z. For example, if $\ell(z, y_i) = \max(0, 1 - y_i z)$, then $\ell_i(\alpha_i) = \alpha_i y_i$ and $\Omega^n = \{\alpha \in \mathbb{R}^n : -1 \leq \alpha \circ \mathbf{y} \leq 0\}$. When the number n of training examples is large it is prohibitive to compute and maintain the kernel matrix K. The Nyström method computes a low-rank approximation of K by sampling a small subset of columns of K or constructing a set of landmark points to address the computation and memory limitations. In particular, if we let $\mathcal{L}_m = \{\mathbf{c}_1, \cdots, \mathbf{c}_m\}$, where $\mathbf{c}_i \in \mathbb{R}^d$, denote a set of m landmark points, $\widetilde{K}_m \in \mathbb{R}^{m\times m}$ denote the sub-kernel matrix between the points in $\mathcal{L}_m$, and $K_b \in \mathbb{R}^{n\times m}$ denote the sub-kernel matrix between all examples and the landmark points, then the Nyström approximation of K is computed by

$$\widehat{K} = K_b \widetilde{K}^\dagger_m K_b^T \quad (3)$$

where $\widetilde{K}^\dagger_m$ denotes the pseudo-inverse of $\widetilde{K}_m$. When applying the Nyström approximation for solving the dual problem, we have the following optimization problem:

$$\widehat{\alpha}_* = \arg\max_{\alpha\in\Omega^n} -\frac{1}{n}\sum_{i=1}^{n} \ell^*_i(\alpha_i) - \frac{1}{2\lambda n^2}\alpha^T(K_b\widetilde{K}^\dagger K_b^T)\alpha \quad (4)$$

which is equivalent to the dual problem of using a short feature representation of training examples

$$\widehat{\mathbf{x}}_i = (\widetilde{K}_m^{\dagger})^{1/2}(\kappa(\mathbf{x}_i, \mathbf{c}_1), \cdots, \kappa(\mathbf{x}_i, \mathbf{c}_m))^T, i = 1, \ldots, n \tag{5}$$

Let $\widehat{X} = (\widehat{\mathbf{x}}_1, \cdots, \widehat{\mathbf{x}}_n) \in \mathbb{R}^{m \times n}$, it is straightforward to verify $\widehat{K} = \widehat{X}^T \widehat{X}$, and the problem (4) can be written as

$$\max_{\alpha \in \ ^n} -\frac{1}{n}\sum_{i=1}^{n} \ ^*_i(\alpha_i) - \frac{1}{2\lambda n^2}\alpha^T \widehat{X}^T \widehat{X}\alpha \tag{6}$$

which can be solved efficiently using stochastic optimization algorithms developed for linear methods (Shalev-Shwartz and Zhang, 2013; Johnson and Zhang, 2013; Lin et al., 2014). The overall running time of the Nyström based kernel classifier consists of the running time of computing the short feature representation of all training data, which is $O(m^2 n + m^3)$, and the running time of optimization. Hence, the Nyström based kernel classifier can be trained efficiently when m is relatively small. On the other hand, the generalization performance of the Nyström based kernel classifier is in the order of $O(1/\sqrt{m})$ for general data, though which can be improved to $O(1/m)$ for some special data (Yang et al., 2012). Therefore, with a small value of m, there still exists a potentially large gap between the performance of the Nyström based kernel classifier and the optimal kernel classifier. In this paper, we propose a refined Nyström based kernel SVM to bridge the gap between the Nyström based kernel classifier and the optimal kernel classifier. To motivate the proposed approach, we first note that given the optimal dual solution α_*, the optimal kernel classifier can be written as: $f_*(\cdot) = -\frac{1}{\lambda n}\sum_{i=1}^{n}[\alpha_*]_i \kappa(\mathbf{x}_i, \cdot)$. If we know that α_* is m-sparse with $m \ll n$ and choose the support vectors as landmark points, i.e., $\mathcal{L}_m^* = \{\mathbf{c}_1^*, \ldots, \mathbf{c}_m^*\} = \{\mathbf{x}_i : [\alpha_*]_i \neq 0\}$, then we can solve the following optimization problem

$$\min_{f \in \mathcal{H}_\kappa^m} \frac{1}{n}\sum_{i=1}^{n} \ (f(\mathbf{x}_i), y_i) + \frac{\lambda}{2}||f||^2_{\mathcal{H}_\kappa} \tag{7}$$

where $\mathcal{H}_\kappa^m = \{f : f = \sum_{i=1}^{m}\beta_i \kappa(\mathbf{c}_i^*, \cdot)\}$. As demonstrated in (Yang et al., 2012), the problem in (7) is equivalent to using the Nyström approximation constructed with the landmark points in $\mathcal{L}_m^*$. It is not difficult to show that under the discussed conditions the optimal solution to the above problem is also the optimal solution to (1). The details are shown in the supplement. From another perspective, following the Theorem 2 in (Hsieh et al., 2014b), the error of $\widehat{\alpha}_*$ is bounded by

$$\begin{aligned} &\|\widehat{\alpha}_* - \alpha_*\|_2 \leq \frac{1}{n\lambda_{nnz}}\|\widetilde{K}_m\|_2(1 + \|\widehat{K}\|_2)\Delta, \\ &\text{where } \Delta = \sum_{i=1}^{n}|[\alpha_*]_i|\|\widehat{K}_{*i} - K_{*i}\|_2 \end{aligned} \tag{8}$$

where λ_{nnz} is the minimum nonzero eigen-value of K/n and K_{*i} denotes the i-th column of K. It indicates that the quality of $\widehat{\alpha}_*$ is mostly affected by a small portion of columns of K with larger $|[\alpha_*]_i|$. The above argument suggests a two-step approach towards improving the performance of the Nyström based kernel classifier: in the first step we learn an approximate dual solution that is close to α_* and then in the second step we construct a set of landmark points aiming to minimize Δ using the approximate dual solution in place of α_*. (Hsieh et al., 2014b) also implements the two-step approach by learning an approximate dual solution using the divide-and-conquer approach (Hsieh et al., 2014a) that divides all examples into a number of groups and solves a small kernel SVM for each group to obtain an approximate dual solution. However, there is no guarantee on the quality of the obtained dual solution. Below, we propose a more solid approach to learn a refined Nyström based kernel SVM.

A Refined Nyström based kernel SVM

Our approach is inspired by the fact that in the optimal kernel classifier the number of support vectors is usually relatively small compared to the total number of examples, indicating the optimal dual solution is a sparse vector. However, when exploring a Nyström approximation, the number of support vectors could increase due to that some examples become difficult to be classified, leading to increased number of support vectors, i.e., a denser dual solution. Therefore, in order to improve the quality of the dual solution, we introduce a sparsity-inducing regularizer into the dual formation defined with the Nyström approximated kernel. In particular, we solve the following formulation to obtain an improved dual solution:

$$\widetilde{\alpha}_* = \arg\max_{\alpha \in \ ^n} -\frac{1}{n}\sum_{i=1}^{n} \ ^*_i(\alpha_i) - \frac{1}{2\lambda n^2}\alpha^T \widehat{K}\alpha - \frac{\tau}{n}||\alpha||_1 \tag{9}$$

It was shown in (Yang et al., 2015a) when the loss is the hinge loss or the squared hinge loss, adding the $_1$ norm on the dual variable α is equivalent to using a new loss function with a reduced margin $1 - \tau$ as compared with 1 used in the standard hinge loss. To see this, we can consider the hinge loss $(z, y) = \max(0, 1 - yz)$, then $^*_i(\alpha_i) = \alpha_i y_i$ and $\Omega^n = \{\alpha \in \mathbb{R}^n : -1 \leq \alpha \circ \mathbf{y} \leq 0\}$, and with a variable change the new problem in (9) can be reduced to

$$\max_{\beta \in [0,1]^n} \frac{1}{n}\sum_{i=1}^{n}\beta_i(1 - \tau) - \frac{1}{2\lambda n^2}(\beta \circ \mathbf{y})^T \widehat{K}(\beta \circ \mathbf{y})$$

which is the dual problem of the following problem

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

with the reduced margin $1 - \tau$ in the definition of the hinge loss. In the next subsection, we provide a theoretical analysis of the proposed sparse-regularized dual formation with the Nyström approximation by establishing an error bound of the obtained dual solution $\widetilde{\alpha}_*$. The above analysis also implies that the new formulation can be solved with the same time complexity as solving the original formulation in (6).

Next, we briefly discuss the second step that uses the obtained dual solution $\widetilde{\alpha}_*$ to re-train a refined Nyström based kernel classifier. The methodology is to select a new set of landmark points using the dual solution $\widetilde{\alpha}_*$ and then learn a Nyström based kernel classifier using the selected landmark points. In Hsieh et al. (2014b), the authors have suggested an approach based on weighed k-means clustering. This approach is grounded in that when the kernel function is stationary (i.e., $\kappa(\mathbf{x}_i, \mathbf{x}_j) = \kappa(\|\mathbf{x}_i - \mathbf{x}_j\|_2)$), the Δ in (8) is bounded by a quantity that is proportional to the square root of a weighted k-means objective defined with the weights given by square of the optimal dual solution, i.e., $\sum_{i=1}^{n}[\alpha_*]_i^2\|\mathbf{x}_i - \mathbf{c}_{\pi_i}\|_2$, where $\pi_i = \arg\min_j \|\mathbf{x}_i - \mathbf{c}_j\|_2$. Thus, one can perform a weighted k-means using the weights given by $[\widetilde{\alpha}_*]_i^2$, $i \in [n]$ and use the resulting cluster centers as landmark points to construct the Nyström approximation. However, this approach will introduce an additional cost of weighted k-means clustering to find the clusters and is restricted to stationary kernels. In this paper, we use a simple alternative based on a greedy approach. It is motivated by that if α_* is given we can select the examples that have largest $|[\alpha_*]_i|$ to minimize Δ. In practice, we only obtain an approximate dual solution $\widetilde{\alpha}_*$, hence we opt for a probabilistic sampling approach that selects examples based on the probability distribution $\Pr(\mathbf{x}_i \text{ is selected}) = \frac{|[\widetilde{\alpha}_*]_i|}{\sum_{i=1}^{n}|[\widetilde{\alpha}_*]_i|}$, which is observed to be more effective than a deterministic approach that simply selects examples that have largest $|[\widetilde{\alpha}_*]_i|$ and also competitive with the weighted k-means sampling approach.

A Theoretical Analysis

We provide a theoretical analysis of the error of $\widetilde{\alpha}_*$ below and finally present a theorem to summarize the main result. Let $\mathcal{S}$ be the support set of α_* and $s = |\mathcal{S}|$ be the number of non-zero entries in α_*. Denote by $\alpha_{\mathcal{S}}$ the vector that only contains elements of α in $\mathcal{S}$. We assume that $s \ll n$. Before presenting our analysis we need to define a few quantities regarding the kernel matrix as follows:

$$\widehat{\Delta} = \frac{1}{n}\sum_{i=1}^{n}|[\alpha_*]_i||\widehat{K}_{*,i} - K_{*,i}|_\infty,$$

$$\gamma(s) = \min_{1 \leq \|\alpha\|_0 \leq s} \frac{1}{n}\frac{\alpha^\top K \alpha}{\|\alpha\|_2} > 0$$

where $\gamma(s) > 0$ is known as restricted eigen-value condition of K/n in the literature (Bickel et al., 2009). We denote by $\lambda_i, i \in [n]$ the eigen-values of K/n ranked in the descending order. In addition, we introduce the following coherence measure to facilitate our analysis. For any real PSD matrix $A \in \mathbb{R}^{N\times N}$, let $\tau_k(A)$ denote the coherence of a dominant k-dimensional invariant subspace of A specified by $\tau_k(A) = \frac{N}{k}\max_i (P_{U_k})_{ii}$, where $P_{U_k} = U_k U_k^\top$ denotes the projection onto the dominant k invariant subspace of A (i.e., U_k contains the top-k eigen-vectors of A as its columns). The coherence measure has been used in matrix completion (Recht, 2011) and random matrix approximation (Gittens, 2011). To characterize the coherence measure of the kernel matrix K with respect to any subset Ω of cardinality $m + s$, we define

$$\tau_k(m, s) = \max_{\in [n], |\ | = m+s} \tau_k(K\ ,\ )$$

where K , is the submatrix of K with row and column indices in Ω. We first present the following lemma showing that $\widetilde{\alpha}_* - \alpha_*$ lies in the cones of dominant coordinates as in the definition of restricted eigen-value.

Lemma 1. $\mathcal{S}$ α_* $\mathcal{S}^c$ $\tau \geq \frac{2}{\lambda n}\sum_{i=1}^{n}|[\alpha_*]_i|\|\widehat{K}_{*i} - K_{*i}\|_\infty$ $\|[\widetilde{\alpha}_* - \alpha_*]_{\mathcal{S}^c}\|_1 \leq 3\|[\widetilde{\alpha}_* - \alpha_*]_{\mathcal{S}}\|_1$.

Due to limit of space, we put all proofs in the supplement. We assume $^*_i(\alpha)$ is μ-strongly convex, where $\mu \geq 0$ [1]. Following the optimality condition of $\widetilde{\alpha}_*$ to (9) and the optimality condition of α_* to (2), there exists $g_* \in \partial\|\widetilde{\alpha}_*\|_1$ such that

$$(\widetilde{\alpha}_* - \alpha_*)^\top(\nabla f(\widetilde{\alpha}_*) + \frac{1}{\lambda n^2}\widehat{K}\widetilde{\alpha}_* + \frac{\tau}{n}g_*) \leq 0,$$
$$(\widetilde{\alpha}_* - \alpha_*)^\top(\nabla f(\alpha_*) + \frac{1}{\lambda n^2}K\alpha_*) \geq 0$$

where $f(\alpha) = \frac{1}{n}\sum_{i=1}^{n}\ ^*_i(\alpha_i)$. Using the Cauchy-Schwarz inequality, we have

$$(\widetilde{\alpha}_* - \alpha_*)^\top(\nabla f(\widetilde{\alpha}_*) + \frac{1}{\lambda n^2}\widehat{K}\widetilde{\alpha}_*) + \frac{\tau}{n}\|[\widetilde{\alpha}_*]_{S^c}\|_1$$
$$\leq \frac{\tau}{n}\|[\widetilde{\alpha}_* - \alpha_*]_S\|_1$$

Thus, we have

$$\begin{aligned}
&\frac{\tau}{n}\|[\widetilde{\alpha}_* - \alpha_*]_S\|_1 - \frac{\tau}{n}\|[\widetilde{\alpha}_*]_{S^c}\|_1\\
&\geq (\widetilde{\alpha}_* - \alpha_*)^\top(\nabla f(\widetilde{\alpha}_*) + \frac{1}{\lambda n^2}\widehat{K}\widetilde{\alpha}_*)\\
&= (\widetilde{\alpha}_* - \alpha_*)^\top(\nabla f(\alpha_*) + \frac{1}{\lambda n^2}K\alpha_*)\\
&+ (\widetilde{\alpha}_* - \alpha_*)^\top(\nabla f(\widetilde{\alpha}_*) - \nabla f(\alpha_*))\\
&+ \frac{1}{\lambda n^2}(\widetilde{\alpha}_* - \alpha_*)^\top(\widehat{K}\widetilde{\alpha}_* - K\alpha_*)\\
&\geq 0 + \frac{\mu}{n}\|\widetilde{\alpha}_* - \alpha_*\|_2^2 + \frac{1}{\lambda n}\underbrace{\frac{1}{n}(\widetilde{\alpha}_* - \alpha_*)^\top(\widehat{K} - K)\alpha_*}_{A}\\
&+ \frac{1}{\lambda n}\underbrace{\frac{1}{n}(\widetilde{\alpha}_* - \alpha_*)^\top\widehat{K}(\widetilde{\alpha}_* - \alpha_*)}_{B}
\end{aligned}$$

where the second inequality uses that $f(\alpha)$ is (μ/n)-strongly. Next, we bound the two terms A, and B. For bounding A, we have $A \geq -\|\widetilde{\alpha}_* - \alpha_*\|_1\widehat{\Delta}$. For bounding B, we prove the following lemma in the supplement.

Lemma 2. $m \geq 8k\tau_k(m, 16s)\left(16s\log d + \log\frac{k}{\delta}\right)$ $\tau \geq \frac{2}{\lambda n}\sum_{i=1}^{n}|[\alpha_*]_i|\|\widehat{K}_{*i} - K_{*i}\|_\infty$

$$B \geq 2\left(\gamma(16s) - \left(3 + \frac{32s}{m}\right)\lambda_{k+1}\right)\|\widetilde{\alpha}_* - \alpha_*\|_2^2.$$

[1] When $\mu = 0$, it is a convex function. The squared hinge loss is $1/2$-strongly convex.

Table 1: Statistics of datasets

Name	usps	letter	ijcnn1	webspam	cod-rna	covtype
#Training	7,291	12,000	91,701	280,000	271,617	464,810
#Testing	2,007	6000	49,990	70,000	59,535	116,202
#Features	256	16	22	254	8	54

Given the above bounds for A and B, we have

$$\|\widetilde{\alpha}_* - \alpha_*\|_2^2(\lambda\mu + 2\gamma(16s) - (6 + 64s/m)\lambda_{k+1}) + (\lambda\tau - \widehat{\Delta})|\hat{\alpha}_{S^c}|_1 \leq (\lambda\tau + \widehat{\Delta})\|[\widetilde{\alpha}_* - \alpha_*]_S\|_1$$

If we assume that $\tau \geq \frac{2\widehat{\ }}{\lambda}$, then it is not difficult to prove the error bound stated in the following theorem.

Theorem 1. , , , , , k , $\delta \in (0, 1)$, , , , , , , , , , , , , ,

$$\lambda\mu + 2\gamma(16s) \geq \left(6 + \frac{64s}{m}\right)\lambda_{k+1},$$

$$m \geq 8k\tau_k(m, 16s)\left(16s\log d + \log\frac{k}{\delta}\right)$$

, , , , $\tau \geq \frac{2}{\lambda n}\sum_{i=1}^{n}|[\alpha_*]_i|\|\widehat{K}_{*i} - K_{*i}\|_\infty$, , , , , , , , $1 - \delta$, ,

$$\|\widetilde{\alpha}_* - \alpha_*\|_2 \leq \frac{1.5\lambda\sqrt{s}\tau}{\lambda\mu + 2\gamma(16s) - (6 + 64s/m)\lambda_{k+1}}$$

Remark: It is interesting to compare our error bound of $\widetilde{\alpha}_*$ with the bound of the original Nyström based formulation in terms of $\widehat{\alpha}_*$ as in (8) derived by (Hsieh et al., 2014b) and the error bound of dual solution obtained by the divide-and-conquer approach (Hsieh et al., 2014a). Considering $\tau = \Theta(\frac{1}{\lambda n}\sum_{i=1}^{n}|[\alpha_*]_i|\|\widehat{K}_{*i} - K_{*i}\|_\infty)$, compared with (8), our error bound is proportional to $\sum_{i=1}^{n}|[\alpha_*]_i|\|\widehat{K}_{*i} - K_{*i}\|_\infty$ which is smaller than Δ as in the error bound of $\widehat{\alpha}_*$. The error bound of $\widehat{\alpha}_*$ has an inverse dependence on the minimum non-zero eigen-value of K/n, which in practice could be very close to zero, leading to potentially a large error in $\widehat{\alpha}_*$. In contrast, our error bound is inversely proportional to $\lambda\mu + 2\gamma(16s) - (6 + 64s/m)\lambda_{k+1}$, depending on the minimum restricted eigen-value. In addition, the error bound in (8) depends on $\|\widetilde{K}_m\|_2$ and $\|\widehat{K}\|_2$, while our error bound only depends on $\sqrt{s}$, making the proposed refined Nyström based kernel classifier attractive when the number of support vectors is relatively small. Compared with the error bound of the approximate solution obtained by the divide-and-conquer approach (Theorem 1 (Hsieh et al., 2014a)), which depends on how well the data is clustered and is inversely proportional to the minimum eigen-value of the kernel matrix, the bound in Theorem 1 is better.

Experiments

Implementation

In our experiments, we implement both the feature construction by the Nyström method and the optimization of linear SVM in a cluster environment. The training data is randomly partitioned over 5 nodes. Given the landmark points, we construct the short feature representation as in Eqn. (5) for all training examples by running the code parallel on 5 nodes. To solve the linear SVM problem in a distributed fashion, we use the recently proposed distributed stochastic dual coordinate ascent algorithm (Yang, 2013; Ma et al., 2015).

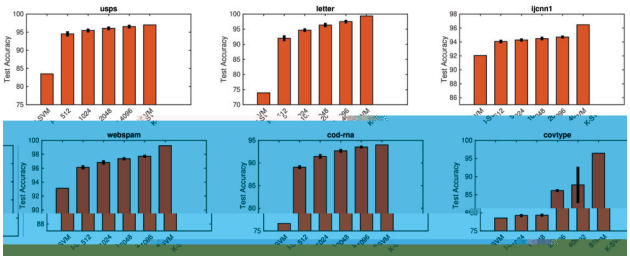

Figure 1: Test Accuracy for linear SVM, RBF SVM and Nyström based kernel classifier with different number of samples on the six datasets.

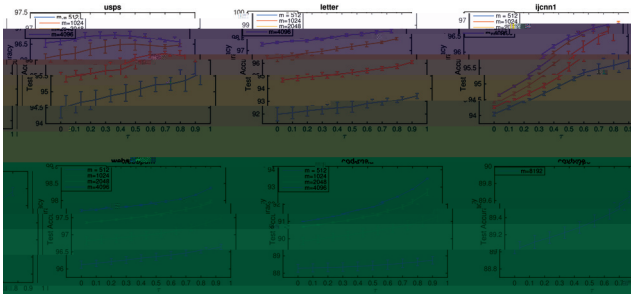

Figure 2: Test Accuracy of the spare-regularized Nyström based kernel classifier.

Experimental Results

In this section, we present empirical evaluations of the proposed refined Nyström based kernel classifier on six real-world datasets, namely usps, letter, ijcnn1, webspam, cod-rna and covtype, of which we use the version available on LIBSVM website [2]. Table 1 summarizes the statistics of these datasets. We run linear SVM and kernel SVM using LIBLINEAR and LIBSVM, respectively. The kernel used in the experiments is the RBF kernel and the loss function is the hinge loss. Through cross-validation, we choose the best parameter C from $2^{[-6:1:6]}$ and the best parameter γ for the RBF kernel from $2^{[-6:2:6]}$. For the methods that involves randomness, the results are averaged over five random trials of sampling.

We first compare the original Nyström based kernel classifier with different number of samples to linear SVM and kernel SVM, with results shown in Figure 1. For the original Nyström approximation, we use uniform sampling to select examples from the training data. We can see that as the number of samples m for the Nyström approximation increases, the test accuracy is monotonically increas-

[2] http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/

ing. However, there still exits a potentially large gap in the performance as compared with the optimal kernel classifier. For example, the optimal kernel classifier outperforms the Nyström based kernel classifier with $m = 4096$ by 3 percent on webspam dataset and by 4 percent on the ijcnn1 dataset.

Next, we verify the proposed sparse-regularized dual formulation with the Nyström approximation. The test accuracy is evaluated on the returned model constructed from the obtained dual variables. Comparing this with the the standard Nyström based kernel classifier allows us to see the effect of the added sparsity-inducing regularizer. The results are shown in Figure 2 with the value of τ varying from 0 to 0.9 [3]. When the value of τ is 0, it reduces to the standard Nyström based kernel classifier. From the results, we can see that adding the sparsity-inducing regularizer to the dual formulation with the Nyström approximation can boost the performance. For example, when $m = 4096$ the performance of the Nyström based kernel classifier is improved by 2 percent on the ijcnn1 dataset.

Next, we examine the performance of the refined Nyström based kernel SVM. We use the obtained dual solution to the sparse-regularized dual formulation to construct a new set of landmark points to re-train a Nyström based kernel classifier. For each value of τ, we optimize the sparse-regularized dual formulation and obtain a dual solution, then we use the dual solution to construct the same number of landmark points to compute a new Nyström approximation for learning a new classifier. We report the results of the probabilistic approach (referred to as **sp-pro-nys**) for constructing the landmark points, which is described on page 4. For baselines, we include the results of the model directly constructed from the obtained dual solution in the first step (referred to as **sp**) and the approach that uses the divide-and-conquer approach (Hsieh et al., 2014a) to obtain a dual solution to re-train a Nyström based kernel classifier using the weighted k-means to find the centers as the landmark points as suggested in (Hsieh et al., 2014b). This approach is referred to as **dc-wkm-nys**. Note that the divide-and-conquer approach requires a clustering on the training data in order to obtain a partition of the training data. We follow the idea in (Hsieh et al., 2014b) and use the standard k-means clustering to partition the data instead of the expensive kernel k-means clustering as suggested in (Hsieh et al., 2014a). The results are shown in Figure 3. From the results we can see that (i) the second step that re-trains a new Nyström based kernel classifier using the obtained dual solution in the first step can further improve the performance; (ii) the proposed new pipeline outperforms the divide-and-conquer approach followed by the weighted k-means sampling approach for constructing a new Nyström approximation.

Finally, we compare the training time of linear SVM, kernel SVM, the standard Nyström based kernel classifier and the refined Nyström based kernel classifier. We report the results on two datasets webspam and cod-rna with $m = 1024$ in the Figure 4. It shows that the training time of Nyström on both datasets is much less than kernel SVM. On the other hand, the training time of the refined Nyström based classifier is also comparable to training time of the standard Nyström method.

[3] When $\tau > 1$, it will yield trivial solution with the optimal model being zero. To avoid clutter, we only show one curve on the covtype dataset.

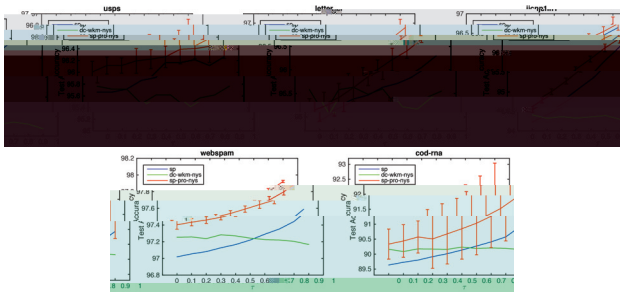

Figure 3: Test Accuracy of the refined Nyström based kernel classifier (sp-pro-nys). The value of m is set to 1024 in the proposed algorithm. For the divide-and-conquer approach, the x-axis denotes $1-nc/100$, where nc denotes the number of clusters used in divide-and-conquer approach. We did not report the result on the covtype due to that the divide-and-conquer approach needs to solve multiple kernel SVM on each partition. When the number of partitions is small, each kernel SVM is still expensive.

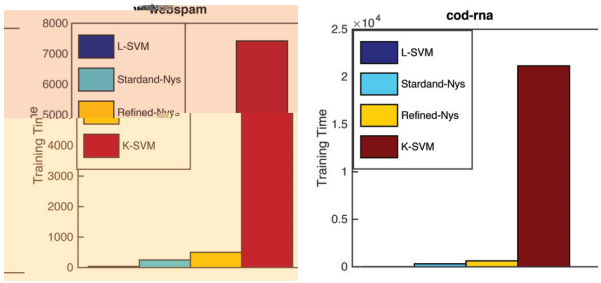

Figure 4: Training time of linear SVM, kernel SVM, the standard Nyström based classifier, and the refined Nyström based classifier for $m = 1024$ on two datasets webspam and cod-rna.

Conclusions

In this paper, we have considered improving the performance of the Nyström based kernel SVM. We proposed a fast and accurate refined Nyström based kernel classifier that consists of two steps, where in the first step we learn an accurate dual solution based on a sparse-regularized dual formulation with the Nyström approximation and in the second step we use the obtained dual solution to re-train a Nyström based kernel classifier. We established an error bound of the obtained dual solution in the first step, which is better than previous theoretical results. The empirical evaluations on various datasets further demonstrate the effectiveness of the proposed algorithm.

Acknowlegements

This work was partially supported by NSF IIS-1463988 and NSF IIS-1545995.

References

Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov. Simultaneous analysis of lasso and dantzig selector. , 37(4), 2009.

Petros Drineas and Michael W. Mahoney. On the nyström method for approximating a gram matrix for improved kernel-based learning. , pages 2153–2175, 2005.

Alex Gittens. The spectral norm error of the naive nystrom extension. , 2011.

Alex Gittens and Michael W. Mahoney. Revisiting the nystrom method for improved large-scale machine learning. , abs/1303.1849, 2013.

Cho-Jui Hsieh, Si Si, and Inderjit S. Dhillon. A divide-and-conquer solver for kernel support vector machines. In , pages 566–574, 2014a.

Cho-Jui Hsieh, Si Si, and Inderjit S. Dhillon. Fast prediction for large-scale kernel machines. In , pages 3689–3697, 2014b.

Rong Jin, Tianbao Yang, Mehrdad Mahdavi, Yu-Feng Li, and Zhi-Hua Zhou. Improved bounds for the nyström method with application to kernel classification. , 59(10): 6939–6949, 2013.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In , pages 315–323, 2013.

Sanjiv Kumar, Mehryar Mohri, and Ameet Talwalkar. On sampling-based approximate spectral decomposition. In , pages 553–560, 2009a.

Sanjiv Kumar, Mehryar Mohri, and Ameet Talwalkar. Sampling techniques for the nystrom method. In , pages 304 – 311, 2009b.

Quoc Le, Tamás Sarlós, and Alex Smola. Fastfood–approximating kernel expansions in loglinear time. In , 2013.

Qihang Lin, Zhaosong Lu, and Lin Xiao. An accelerated proximal coordinate gradient method. In , pages 3059–3067, 2014.

Chenxin Ma, Virginia Smith, Martin Jaggi, Michael I. Jordan, Peter Richtárik, and Martin Takác. Adding vs. averaging in distributed primal-dual optimization. In , pages 1973–1982, 2015.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In , pages 1177–1184, 2007.

Benjamin Recht. A simpler approach to matrix completion. , 12:3413–3430, 2011.

Bernhard Scholkopf and Alexander J Smola. . MIT press, 2001.

Shai Shalev-Shwartz and Tong Zhang. Stochastic Dual Coordinate Ascent Methods for Regularized Loss Minimization. , pages 567–599, 2013.

John Shawe-Taylor and Nello Cristianini. . Cambridge university press, 2004.

Ameet Talwalkar and Afshin Rostamizadeh. Matrix coherence and the nystrom method. In , 2010.

Christopher Williams and Matthias Seeger. Using the nyström method to speed up kernel machines. In , pages 682–688, 2001.

Zenglin Xu, Rong Jin, Bin Shen, and Shenghuo Zhu. Nystrom approximation for sparse kernel methods: Theoretical analysis and empirical evaluation. In , pages 3115–3121, 2015.

Tianbao Yang. Trading computation for communication: Distributed stochastic dual coordinate ascent. In , pages 629–637, 2013.

Tianbao Yang, Yu-Feng Li, Mehrdad Mahdavi, Rong Jin, and Zhi-Hua Zhou. Nystrom method vs random fourier features: A theoretical and empirical comparison. In , pages 485–493, 2012.

Tianbao Yang, Lijun Zhang, Rong Jin, and Shenghuo Zhu. Theory of dual-sparse regularized randomized reduction. , 2015a.

Zichao Yang, Andrew Gordon Wilson, Alexander J. Smola, and Le Song. A la carte - learning fast kernels. In , 2015b.

Kai Zhang, Ivor W. Tsang, and James T. Kwok. Improved nystrom low-rank approximation and error analysis. In , 2008.