Neurocomputing (xxxx) xxxx–xxxx

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Predictive Nyström method for kernel methods☆

Jiangang Wu, Lizhong Ding, Shizhong Liao*

School of Computer Science and Technology,Tianjin University,Tianjin 300350, China

ARTICLE INFO

Communicated by Zidong Wang

Keywords:
Nyström method
Matrix approximation
Kernel methods
Predictive sampling strategy

ABSTRACT

Nyström method is a widely used matrix approximation method for scaling up kernel methods, and existing sampling strategies for Nyström method are proposed to improve the matrix approximation accuracy, but leaving approximation independent of learning, which can result in poor predictive performance of kernel methods. In this paper, we propose a novel predictive sampling strategy (PRESS) for Nyström method that guarantees the predictive performance of kernel methods. PRESS adaptively updates the sampling distribution via the discrepancy between approximate and accurate solutions of kernel methods caused by kernel matrix approximation, and samples informative columns from the kernel matrix according to the sampling distribution to reduce the predictive performance loss of kernel methods. We prove upper error bounds on the approximate solutions of kernel methods produced by Nyström method with PRESS, whose convergence shows that approximate solutions of kernel methods are identical to accurate ones for large enough samples. Experimental results indicate that integrating learning into approximation is necessary for delivering better predictive performance, and PRESS significantly outperforms existing sampling strategies while preserving low computational cost.

1. Introduction

Kernel methods are a class of widely used learning algorithms for machine learning and data mining. However large-scale applications pose new challenges on kernel methods as it is hard to store and operate on large kernel matrices. To mitigate the burden on time and space requirement, many low-rank matrix approximation methods are developed [1–3].

Nyström method [4] is one of the most popular methods for approximate spectral decomposition of a large kernel matrix due to its simplicity and efficiency. It scales up kernel methods by reducing the decomposition of a large kernel matrix to that of a small intersection matrix via a sampling strategy. Initially Nyström method is used to speed up Gaussian process regression [4] and now it has been extended to various fields, such as kernel SVM [5,6], kernel discriminant analysis [7], multiple kernel learning [8], spectral clustering [9,10], manifold learning algorithms [11,12], signal processing [13,14] and so on.

Some work aims at refining the decomposition of intersection matrix. One-shot Nyström method [15] obtains an orthonormal set of approximate eigenvectors of Nyström method, which is widely used for kernel PCA [16]. Modified Nyström method [17] uses a new form of intersection matrix to minimize the matrix reconstruction error, but its computation is time-consuming. Ensemble Nyström method [18] consists of several Nyström approximations and delivers better approximation results empirically. Nyström method with randomized SVD [19,20] enjoys a lower time complexity by speeding up the decomposition of intersection matrix, but it needs to sample more columns than other methods due to the accuracy loss caused by approximate SVD.

In order to introduce the label information into the low-rank matrix decomposition, some supervised methods have been proposed. Cholesky with side information [21] exploits label information in the computation of low-rank decompositions for kernel matrices, which yields decompositions of significantly lower rank than incomplete Cholesky decomposition under the same prediction performance. Generalized Nyström method [22] requires that the reconstructed kernel matrix is close to the ideal kernel matrix defined on labeled samples besides the original kernel matrix, which shows better prediction performance than original Nyström method. The above methods both define new objective function to account for the information of kernel matrix and label, which are more effective than unsupervised matrix approximation methods in practice.

As the sampled columns from kernel matrix heavily affect the approximation accuracy, various sampling strategies for Nyström method are studied. Uniform sampling without replacement is the most basic sampling strategy [4] and its probabilistic error bounds for

☆ Funding: This work was supported in part by National Natural Foundation of China (No. 61673293).
* Corresponding author.
E-mail addresses: szliao@tju.edu.cn (S. Liao).

http://dx.doi.org/10.1016/j.neucom.2016.12.047
Received 13 January 2016; Received in revised form 15 December 2016; Accepted 17 December 2016

Nyström method are obtained [18,23]. Afterwards some non-uniform sampling strategies are proposed, such as diagonal sampling [24], column-norm sampling [25], leverage score sampling [26] and so on. Adaptive sampling [27] selects columns according to a iteratively updated sampling distribution. A variant of adaptive sampling [18] improves the time complexity by avoiding computing the entire kernel matrix.

Existing sampling strategies focus on the information of kernel matrix and most analyses about sampling strategies also only focus on the matrix approximation error of Nyström method, which are totally independent of kernel methods. Therefore such learning-irrelevant sampling strategies may deteriorate the predictive performance of kernel methods.

In this paper, we propose a novel predictive sampling strategy (PRESS) for Nyström method, which integrates learning into the approximation of kernel matrix. PRESS adaptively defines the sampling distribution via the learning discrepancy of kernel methods, which is the discrepancy between approximate and accurate solutions of kernel methods caused by kernel matrix approximation, and then according to the sampling distribution PRESS samples informative columns that can reduce the predictive performance loss of kernel methods. We consider two learning algorithms in this paper: least squares support vector machine (LSSVM) and kernel ridge regression (KRR). We derive upper error bounds on the approximate solutions of LSSVM and KRR produced by Nyström method with PRESS, which tend to zero when the number of samples is large enough. This theoretical result guarantees that the approximate and accurate learning algorithms are consistent, i.e., approximate solutions of learning algorithms are identical to accurate ones for large enough samples. Experimental results demonstrate that PRESS delivers better predictive performance for kernel methods than existing sampling strategies while preserving competitive running time.

The rest of this paper is organized as follows. Section 2 briefly introduces LSSVM, KRR, Nyström method and its sampling strategies. Section 3 describes PRESS and gives upper error bounds on the approximate solutions of LSSVM and KRR. Experimental results are presented in Section 4. Finally we conclude in Section 5.

2. Preliminaries

In this section, we first present some notations and then introduce LSSVM, KRR, Nyström method and its sampling strategies.

2.1. Notations

Let $\mathcal{X}$ denote an input space and $\mathcal{Y}$ an output domain. We usually have $\mathcal{X} \subseteq \mathbb{R}^d$, $\mathcal{Y} = \{-1, 1\}$ for classification task and $\mathcal{Y} = \mathbb{R}$ for regression task. A training set is denoted by $\mathcal{S} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \in (\mathcal{X} \times \mathcal{Y})^n$. $\phi: \mathcal{X} \to \mathcal{F}$ denotes a feature mapping from an input space $\mathcal{X}$ to a feature space $\mathcal{F}$. $\mathbf{K} = [\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)]_{i,j=1}^n$ denotes a kernel matrix with a kernel function $\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)$. $\|\cdot\|_2$ and $\|\cdot\|_F$ denote spectral norm and Frobenius norm respectively.

For a matrix $\mathbf{M} \in \mathbb{R}^{m\times n}$, let $\rho = \mathrm{rank}(\mathbf{M})$. We write the compact singular value decomposition (SVD) of $\mathbf{M}$ as $\mathbf{U}_\mathbf{M}\mathbf{\Sigma}_\mathbf{M}\mathbf{V}_\mathbf{M}^\mathrm{T}$ where $\mathbf{\Sigma}_\mathbf{M}$ is a diagonal matrix containing ρ non-zero singular values $\sigma_1(\mathbf{M}), \sigma_2(\mathbf{M}), \ldots, \sigma_\rho(\mathbf{M})$ in decreasing order, $\mathbf{U}_\mathbf{M} \in \mathbb{R}^{m\times\rho}$ and $\mathbf{V}_\mathbf{M} \in \mathbb{R}^{n\times\rho}$ are the corresponding left and right singular vectors of $\mathbf{M}$. Let $\boldsymbol{u}_i(\mathbf{M})$ and $\boldsymbol{v}_i(\mathbf{M})$ denote the ith left and right singular vector, $\mathbf{M}_k = \sum_{i=1}^k \sigma_i(\mathbf{M})\boldsymbol{u}_i(\mathbf{M})\boldsymbol{v}_i(\mathbf{M})^\mathrm{T}$ the best rank-k approximation to $\mathbf{M}$, $\mathbf{M}^+ = \mathbf{V}_\mathbf{M}\mathbf{\Sigma}_\mathbf{M}^{-1}\mathbf{U}_\mathbf{M}^\mathrm{T}$ the Moore-Penrose pseudoinverse of $\mathbf{M}$. We use $\mathbf{M}_{\cdot i}$ and $\mathbf{M}_{i\cdot}$ to denote the ith column and row of $\mathbf{M}$ respectively. $\mathbf{I}_n$ denotes the $n \times n$ identity matrix and $\mathbf{1}_n$ the vector of n ones. $\bigcirc$ denotes the Hadamard product.

2.2. LSSVM and KRR

LSSVM is a learning algorithm for classification task. It seeks a linear classifier $f(\boldsymbol{x}) = \langle \boldsymbol{w}, \phi(\boldsymbol{x})\rangle + b$ in a feature space $\mathcal{F}$, and the parameters $(\boldsymbol{w}, b)$ are given by the minimizer of a regularized least squares training function [28]:

$$L = \frac{1}{2}\|\boldsymbol{w}\|^2 + \frac{1}{2\mu n}\sum_{i=1}^n [y_i - \langle \boldsymbol{w}, \phi(\boldsymbol{x}_i)\rangle - b]^2,$$

where $\mu > 0$ is called the regularization parameter. Further, we can obtain the dual form as follows:

$$\begin{bmatrix} \mathbf{K}_{\mu,n} & \mathbf{1}_n \\ \mathbf{1}_n^\mathrm{T} & 0 \end{bmatrix}\begin{bmatrix} \boldsymbol{a} \\ b \end{bmatrix} = \begin{bmatrix} \boldsymbol{y} \\ 0 \end{bmatrix}, \tag{1}$$

where $\boldsymbol{a} = (a_1, a_2, \ldots, a_n)^\mathrm{T} \in \mathbb{R}^n$ is a vector of Lagrange multipliers, $\mathbf{K}_{\mu,n}$ denotes $\mathbf{K} + \mu n\mathbf{I}_n$ and $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^\mathrm{T}$. Unfortunately the matrix on the left-hand side of (1) is not positive definite, so it cannot be solved directly. In order to solve it, we introduce two variables $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ and solve

$$\mathbf{K}_{\mu,n}\boldsymbol{\alpha} = \boldsymbol{y} \quad \text{and} \quad \mathbf{K}_{\mu,n}\boldsymbol{\beta} = \mathbf{1}_n, \tag{2}$$

then the solutions $(\boldsymbol{a}, b)$ are derived by

$$b = \frac{\mathbf{1}_n^\mathrm{T}\boldsymbol{\alpha}}{\mathbf{1}_n^\mathrm{T}\boldsymbol{\beta}} \quad \text{and} \quad \boldsymbol{a} = \boldsymbol{\alpha} - \boldsymbol{\beta}b.$$

The decision function of LSSVM can be written as

$$f(\boldsymbol{x}) = \sum_{i=1}^n a_i\kappa(\boldsymbol{x}_i, \boldsymbol{x}) + b.$$

KRR is a learning algorithm for regression task. Its primal optimization problem is as follows [29]:

$$\min_{\boldsymbol{w}}\left\{\mu\|\boldsymbol{w}\|^2 + \frac{1}{n}\sum_{i=1}^n [y_i - \langle \boldsymbol{w}, \phi(\boldsymbol{x}_i)\rangle]^2\right\}.$$

The following is the dual optimization problem of KRR [30]:

$$\max_{\boldsymbol{\alpha}}\{2\boldsymbol{y}^\mathrm{T}\boldsymbol{\alpha} - \boldsymbol{\alpha}^\mathrm{T}(\mathbf{K} + \mu n\mathbf{I}_n)\boldsymbol{\alpha}\}, \tag{3}$$

where μ is the ridge parameter. By setting the gradient of (3) to zero, we can obtain the optimal solution by solving

$$\mathbf{K}_{\mu,n}\boldsymbol{\alpha} = \boldsymbol{y}. \tag{4}$$

The decision function of KRR can be written as

$$f(\boldsymbol{x}) = \sum_{i=1}^n \alpha_i\kappa(\boldsymbol{x}_i, \boldsymbol{x}).$$

2.3. Nyström method

Suppose we have obtained l columns from a kernel matrix $\mathbf{K} \in \mathbb{R}^{n\times n}$ using certain sampling strategy with $l \ll n$. Let $\mathbf{C}$ denote the $n \times l$ matrix formed by the l columns and $\mathbf{W}$ the $l \times l$ intersection matrix formed by the intersection entries of the l columns with the corresponding l rows. Without loss of generality, after rearranging columns and rows, $\mathbf{K}$ and $\mathbf{C}$ can be written as:

$$\mathbf{K} = \begin{bmatrix} \mathbf{W} & \mathbf{K}_{21}^\mathrm{T} \\ \mathbf{K}_{21} & \mathbf{K}_{22} \end{bmatrix} \quad \text{and} \quad \mathbf{C} = \begin{bmatrix} \mathbf{W} \\ \mathbf{K}_{21} \end{bmatrix}.$$

Nyström method constructs a rank-k approximate matrix as follows:

$$\widetilde{\mathbf{K}} = \mathbf{C}\mathbf{W}_k^+\mathbf{C}^\mathrm{T},$$

where $\mathbf{W}_k^+$ is the Moore-Penrose pseudoinverse of the best rank-k approximation to $\mathbf{W}$. The time complexity of Nyström method is