Information Fusion 26 (2015) 36–48

Contents lists available at ScienceDirect

Information Fusion

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

A review of Nyström methods for large-scale machine learning

Shiliang Sun *, Jing Zhao, Jiang Zhu

Shanghai Key Laboratory of Multidimensional Information Processing, Department of Computer Science and Technology, East China Normal University, 500 Dongchuan Road, Shanghai 200241, PR China

ARTICLE INFO

Article history:
Received 29 December 2014
Received in revised form 4 March 2015
Accepted 8 March 2015
Available online 16 March 2015

Keywords:
Low-rank approximation
Nyström method
Sampling method
Machine learning

ABSTRACT

Generating a low-rank matrix approximation is very important in large-scale machine learning applications. The standard Nyström method is one of the state-of-the-art techniques to generate such an approximation. It has got rapid developments since being applied to Gaussian process regression. Several enhanced Nyström methods such as ensemble Nyström, modified Nyström and SS-Nyström have been proposed. In addition, many sampling methods have been developed. In this paper, we review the Nyström methods for large-scale machine learning. First, we introduce various Nyström methods. Second, we review different sampling methods for the Nyström methods and summarize them from the perspectives of both theoretical analysis and practical performance. Then, we list several typical machine learning applications that utilize the Nyström methods. Finally, we make our conclusions after discussing some open machine learning problems related to Nyström methods.

1. Introduction

Many large-scale machine learning problems involve generating a low-rank matrix approximation to reduce high time and space complexities. For example, let n be the number of data instances. The Gaussian process regression computes the inverse of an $n \times n$ matrix which takes time $O(n^3)$ and space $O(n^2)$. For the large-scale problems, n can be in the order of tens of thousands to millions, leading to difficulties in operating on, or even storing the matrix.

Various methods [1–3] have been utilized to generate low-rank matrix approximations. The standard Nyström method is one of the state-of-the-art methods. It selects a subset of columns of the original matrix to build an approximation. In general, the standard Nyström method is used to approximate symmetric positive semidefinite (SPSD) matrices, such as Gram or kernel matrices, or their eigenvalues/eigenvectors. For approximating matrix K, it consists of three steps. (1) Sampling step: it samples a subset of columns of K to form matrix C; (2) Pseudo-inverse step: it performs pseudo-inverse of the matrix W formed by the intersection between those sampled columns and the corresponding rows; (3) Multiplication step: it constructs a matrix by using the formulation $CW^{\dagger}C^{\top}$ to approximate the original matrix. For approximating eigenvalues/eigenvectors, it also consists of three steps. (1) Sampling step: it samples a subset of columns of K to form C; (2) Singular value decomposition (SVD) step: it performs SVD of the matrix W formed by the intersection between those sampled columns and the corresponding rows to get singular values and singular vectors, respectively; (3) Extension step: it uses the Nyström extension to get the approximate eigenvalues/eigenvectors of the original matrix.

The standard Nyström method was first introduced into Gaussian process regression for reducing the computational complexity [4]. By replacing the original matrix with a Nyström approximation and subsequently using the Woodbury formula, the matrix inversion can be easily solved with time complexity $O(\ell^2 n)$, where ℓ columns are sampled. After that, the standard Nyström method has got rapid developments. Several enhanced Nyström methods have been developed to provide more accurate matrix approximation or eigenvector approximation, e.g., density-weighted Nyström (DW-Nyström), ensemble Nyström, modified Nyström and modified Nyström method by spectral shifting (SS-Nyström). In addition, some techniques are developed to improve the inner procedures of the Nyström approximation. For the sampling step, various sampling methods [5–7] are utilized for the Nyström methods. For the SVD step, recently an approximate SVD [8] that utilizes randomized SVD algorithms [9] was proposed to accelerate the standard Nyström method for some extreme large-scale machine learning applications.

One key aspect of the Nyström methods is the sampling step. It influences the subsequent approximation accuracy and thus the performance of the learning methods [10]. Initially, uniform sampling is adopted when the standard Nyström method was applied

* Corresponding author. Tel.: +86 21 54345183; fax: +86 21 54345119.
E-mail addresses: shiliangsun@gmail.com, slsun@cs.ecnu.edu.cn (S. Sun).

http://dx.doi.org/10.1016/j.inffus.2015.03.001
1566-2535/

[4], and it is also the most widely applied sampling method due to its low time consumption. After that, various sampling methods [6,11–15] that focus on selecting the most informative columns are proposed. Thus, we call these sampling methods informative-column sampling. We classify these methods into two classes: (1) fixed sampling; (2) adaptive sampling. For fixed sampling, it means the matrix is sampled with a fixed, non-uniform distribution over the columns. The distribution can be defined by a function on the diagonal entities or the column entities of the original matrix [6,11]. For adaptive sampling, it means the matrix is sampled with adaptive techniques [12,13,16,17]. Different from fixed sampling, the sampling distribution will be modified at each iteration. Empirical results suggest that a tradeoff between efficiency and accuracy exists for uniform sampling and informative-column sampling as the latter spends more time in finding a concise subset of informative columns but can provide an improved approximation accuracy. The tradeoff also exists for fixed sampling and adaptive sampling. In real world applications, the tradeoff should be considered carefully before utilizing different sampling techniques.

The standard Nyström method has been applied to many machine learning applications, e.g., manifold learning [18–20], spectral clustering [21–24], kernel-based methods such as kernel support vector machine (SVM) [25,13,26,27] and kernel ridge regression [10,27], signal processing [28,29] and statistical learning [30,31]. Walkar et al. [18] proved that the standard Nyström method combined with Isomap [32] is an efficient tool to extract the low-dimensional manifold structure given millions of high-dimensional face images, and the approximate Isomap tends to perform better than Laplacian Eigenmaps [33] and is tied to the original Isomap on both clustering and classification with the labeled CMU-PIE [34] data set. In the work of Fowlkes et al. [22], spectral clustering with the standard Nyström method outperforms the traditional Lanczos method [35] where the clustering result is measured by normalized cut [36]. In the work of Williams et al. [26], kernel-based Gaussian processes have been accelerated using the Nyström approximation to the kernel matrix, with the time complexity scaled down from $O(n^3)$ to $O(\ell^2 n)$.

Nyström methods can be seen as a kind of information fusion methods. They use the partial data many times to approximate the values that we are interested, such as the eigenvalues/eigenvectors of a matrix or the inverse of a matrix. When applied to machine learning problems, Nyström methods will bring improvements in efficiency on the premise of not reducing performance much.

The remainder of the paper is organized as follows. Section 2 introduces some preliminary knowledge. In Section 3, we introduce the various Nyström methods. In Section 4, some related low-rank matrix approximation methods are presented to differentiate from the Nyström methods. In Section 5, several sampling methods for the Nyström methods including uniform sampling and informative-column sampling are listed, and we subsequently give some comparisons between uniform sampling and informative-column sampling from the perspectives of both theoretical analysis and practical performance. In Section 6, we provide a summary of typical large-scale machine learning applications of the Nyström methods. We make our conclusions in Section 8 after discussing several open machine learning problems related to Nyström methods in Section 7.

2. Preliminary knowledge

2.1. Notations

For an $n \times n$ SPSD matrix $K = [K_{ij}]$, we define $K^{(j)}, \quad j = 1, \ldots, n$, as the jth column vector of K, $K_{(i)}, i = 1, \ldots, n$, as the ith row vector of K. For a vector $\mathbf{x} \in R^n$, let $\|\mathbf{x}\|_\xi, \xi = 1, 2, \infty$, denote the 1-norm, Euclidean norm, and ∞-norm, respectively. Let Diag(K) denote the vector consisting of the diagonal entries of the matrix K and Σ denote the matrix containing the eigenvalues of K. Then, $\|K\|_2 = \|Diag(\Sigma)\|_\infty$ denotes the spectral norm of K; $\|K\|_F = \|Diag(\Sigma)\|_2$ denotes the Frobenius norm of K; and $\|K\|_\otimes = \|Diag(\Sigma)\|_1$ denotes the trace norm (or nuclear norm) of K. Clearly,

$$\|K\|_2 \leqslant \|K\|_F \leqslant \|K\|_\otimes \leqslant \sqrt{n}\|K\|_F \leqslant n\|K\|_2. \tag{1}$$

2.2. Best rank-k approximation

Given a matrix with rank(K) = p, the SVD of K can be written as $K = U\Sigma V$, where Σ is diagonal and contains the singular values $(\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_n)$ of K, and U and V have orthogonal columns and contain the left and right singular vectors of K. Let U_k and V_k be the first k $(k < p)$ columns of U and V, respectively, and Σ_k be the $k \times k$ top sub-block of Σ. Then, the $n \times n$ matrix $K_k = U_k\Sigma_k V_k$ is the best rank-k approximation to K, when measured with respect to any unitarily-invariant matrix norm, e.g., the spectral, Frobenius, or trace norm [37]. We have

$$\|K - K_k\|_2 = \lambda_{k+1}, \tag{2}$$

$$\|K - K_k\|_F = \left(\sum_{i=k+1}^{n} \lambda_i^2\right)^{1/2}, \tag{3}$$

$$\|K - K_k\|_\otimes = \sum_{i=k+1}^{n} \lambda_i. \tag{4}$$

2.3. Pseudo-inverse of a matrix

Another kind of useful Nyström related knowledge is the pseudo-inverse (Moore–Penrose inverse). The pseudo-inverse of an $m \times n$ matrix A can be expressed from the SVD of A as follows. Let the SVD of A be

$$A = U\begin{pmatrix} S & 0 \\ 0 & 0 \end{pmatrix}V^\top, \tag{5}$$

where U, V are both orthogonal matrices, and S is a diagonal matrix containing the (non-zero) singular values of A on its diagonal. Then the pseudo-inverse of A is an $n \times m$ matrix defined as

$$A^\dagger = V\begin{pmatrix} S^{-1} & 0 \\ 0 & 0 \end{pmatrix}U^\top. \tag{6}$$

Note that $A^\dagger$ has the same dimension as the transpose of A. If A is square, invertible, then its pseudo-inverse is the true inverse, that is, $A^\dagger = A^{-1}$.

2.4. Orthogonal projection

In linear algebra and functional analysis, a projection is a linear transformation $\mathcal{P}$ from a vector space to itself such that $\mathcal{P}^2 = \mathcal{P}$. A projection is orthogonal if and only if it is self-adjoint. One way to construct the projection operator on the range space of A is that

$$\mathcal{P}_A = A(A^\top A)^\dagger A^\top = AA^\dagger = HH^\top = UU^\top, \tag{7}$$

where H represents the orthogonal basis on the range space of A and U represents the left singular vectors of A corresponding to the non-zero singular values. Given an $n \times \ell$ matrix C, the projection of K onto the column space of C is defined as $\mathcal{P}_C K = CC^\dagger K$.