ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8. Available from http://www.i6doc.com/en/.

Efficient low rank approximation via alternating least squares for scalable kernel learning

Piyush Bhardwaj[1] and Harish Karnick[2]

1- Microsoft India Development Center
9, Lavelle Road, Bangalore, 560001 - India

2- Indian Institute of Technology Kanpur
Kanpur, 208016 - India

Abstract. Kernel approximation is an effective way of dealing with the scalability challenges of computing, storing and learning with kernel matrix. In this work, we propose an $O(|\Omega|r^2)$ time algorithm for rank r approximation of the kernel matrix by computing $|\Omega|$ entries. The proposed algorithm solves a non-convex optimization problem by random sampling of the entries of the kernel matrix followed by a matrix completion step using alternating least squares (ALS). Empirically, our method shows better performance than other baseline and state-of-the-art kernel approximation methods on several standard real life datasets. Theoretically, we extend the current guarantees of ALS for kernel approximation.

1 Introduction

Kernel methods have shown good performance on several machine learning tasks while providing a firm theoretical understanding. However, kernel methods are restricted in their use due to the scalability issues in using kernel matrix for learning. Challenges in kernel based learning arise from the cost of computing, storing and performing multiplication operations with kernel matrix.

Let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a positive semi-definite kernel function, where $\mathcal{X}$ is the set of possible inputs. In machine learning tasks, often some input data is available in form of a learning set, $\mathcal{L}$. For $\mathcal{L} = \{x_1, \ldots, x_n\}$, the kernel matrix K is defined as the $n \times n$ matrix of kernel function evaluation at observed points i.e. $K_{ij} = K(x_i, x_j)$. The cost of computation and storage of kernel matrix for $\mathcal{X} \subseteq \mathbb{R}^d$ is $O(n^2 d)$ and $O(n^2)$, respectively. Learning using kernels often requires multiple matrix-vector multiplication. This computation often takes $O(n^3)$ time using exact kernel matrix.

Low rank matrix approximations of the form $K \approx UU^T$ where $U \in \mathbb{R}^{n \times r}$ and $r \ll n$ provide a way to overcome the scalability issues. These rank r approximations require nr parameters for storage and the matrix vector multiplications can be performed in $O(nr)$ time. However, computation of a good low rank approximation is challenging. The provably optimal rank r approximation is given by the top r singular vectors computed via SVD which takes $O(n^2 r)$ time and requires computation of the entire kernel matrix [1].

In this paper we extend a general matrix approximation algorithm, Low rank matrix completion (LRMC) by Jain et al. [2] and propose an $O(|\Omega|r^2)$ time algorithm for rank r approximation where $|\Omega|$ is the number of entries of the kernel

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8. Available from http://www.i6doc.com/en/.

matrix that need to be computed. For commonly used kernel functions, we observe that $|\Omega| = O(nr \log n)$ is sufficient for good approximation. Although kernel matrix approximation is a well studied problem, to our knowledge, general matrix approximation schemes haven't been efficiently extended for kernel matrices because these schemes don't respect the symmetry and positive definite nature of kernel matrices.

1.1 Literature Review

Nystrom approximations [3] are among the most popular kernel matrix approximation schemes. These methods attempt to recover the kernel matrix by observing some m columns and corresponding rows of the matrix. Nystrom approximation is given by $K \approx KS(S^TKS)^{\dagger}_r S^TK = UU^T$ where S is a column sampling matrix, $\dagger$ denotes the pseudo-inverse and subscript r denotes the optimal rank r approximation. Extensions of Nystrom method using non-uniform sampling schemes to select rows and columns have been analysed [4]. Sampling of columns based on k-means clustering of input data has also been proposed as K-means Nystrom[5]. This method shows improvement over the standard Nystrom and non-uniform sampling schemes in [4].

Recently Si et al.proposed memory efficient kernel approximation (MEKA) algorithm [6]. This approach works by finding blocks in the kernel matrix followed by approximation of each block. The blocks are composed by clustering the data in the input space. The intracluster kernel evaluation form the diagonal blocks of the kernel matrix and are approximated using standard Nystrom. The off-diagonal blocks are approximated by sampling few elements and solving a least squares problem. MEKA is space efficient; in almost the same space as the rank r approximation of the Nystrom method, MEKA can perform a rank cr approximation, where c is the number of clusters.

Algorithm	Space complexity	Rank	Time Complexity
Nystrom	$O(nr)$	r	$O(m^3 + nmr) \approx O(nr^2)$
KALS	$O(nr)$	r	$O(\|\Omega\|r^2 + nr^3) \approx O(n \log nr^3)$
MEKA	$O(nr + c^2r^2)$	cr	$O(nr^2 + cm^3) + T_L + T_C$

Table 1: Comparison of time and space complexities of different algorithms for kernel approximation. KALS is the approach presented in this paper. T_C and T_L denote the time required for clustering and solving least square problem in MEKA. m is the number of sampled columns for Nystrom approximation

2 Proposed Method

The proposed approach builds up on previous works on matrix approximation using alternating least squares (ALS) [2, 7]. For a general matrix M, ALS

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8. Available from http://www.i6doc.com/en/.

gives approximation of the form $M \approx UV^T$ by solving the following non-convex optimization problem:

$$\min_{U,V} \sum_{i,j} (M_{ij} - e_i^T U V^T e_j)^2 \tag{1}$$

where e_i represents the canonical basis vector $[0, \ldots, 0, 1, 0, \ldots 0]^T$. The problem is solved by iterating over steps of fixing U and solving for V and vice-versa. Each alternating step is convex, but overall the objective function is non-convex. Jain et al. [2] show that if 1 is solved over enough number of entries of M, convergence to the optimal solution is guaranteed.

Algorithm 1 Kernel approximation via ALS

Require: Matrix $K_{n\times n}$, target rank r
Ensure: $\hat{U}_{n\times r}$
1: $\Omega \leftarrow$ sample entries of K.
2: $\hat{U}^{(0)} \leftarrow Nystrom(K, r)$
3: **for** $t = 0, \ldots, T-1$ **do**
4: $\quad \hat{V}^{(t+1)} \leftarrow \text{argmin}_V \sum_{(i,j)\in\Omega} (K_{ij} - e_i^T \hat{U}^{(t)} V^T e_j)^2$
5: $\quad \hat{U}^{(t+1)} \leftarrow \frac{\hat{U}^t + \hat{V}^{t+1}}{2}$

The proposed method, kernel approximation via ALS (KALS), is shown in algorithm 1. For kernel matrix approximation we are interested in solving $\min_U \sum_{i,j} (K_{ij} - e_i^T U U^T e_j)^2$. Algorithm 1 solves the problem stated in 1 but approximates U as $\hat{U}^{(t+1)} \leftarrow \frac{\hat{U}^t + \hat{V}^{t+1}}{2}$ at the end of $t + 1^{th}$ iteration. Also, for kernel matrices, the initialization step can be carried out using Nystrom method which gives a good initialization point for the alternating steps.

Initialization requires running Nystrom which takes $O(nr^2)$ time [3]. Step 4 and 5 of algorithm 1 require solving n least square problems each of which take $O(|\Omega_j|r^2 + r^3)$ time, where $|\Omega_j|$ is the number of entries sampled from the j^{th} column of K. Thus, overall time complexity of running 1 iteration of KALS is $O(|\Omega|r^2 + nr^3)$. . In our experiments we found $|\Omega| = nr \log n$ to give good results. Also,in practice we found running as low as 4 iterations of ALS is enough.

2.1 Theoretical Analysis

In this section we give a summary of geometric convergence of KALS to optimal SVD solution. For the proofs, we consider uniform sampling in step 1 of algorithm 1. The analysis is an extension of analysis in [2]. For rank r approximation at the end of the t^{th} iteration, let $\hat{K} = \hat{U^t}\hat{U^t}^T$ be the approximation. U^t denotes an orthonormal basis spanning the column space of $\hat{U^t}$. Let the optimal rank r approximation of K be $K_r = U^* \Sigma_K^{(r)} U^{*T}$, where $U^* = U_K^{(r)}$ are the top-r singular vectors of K and $\Sigma_K^{(r)}$ is a diagonal matrix of top-r diagonal entries of Σ_K in non-increasing order. $dist(\hat{U}, \hat{W}) = ||U_\perp^T W|| = ||W_\perp^T U||$ denotes the principal angle based distance between the subspaces spanned by the columns of $\hat{U}$

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8. Available from http://www.i6doc.com/en/.

and $\hat{W}$, where $U_\perp$ and $W_\perp$ denote the orthonormal basis spanning the subspace perpendicular to $\hat{U}$ and $\hat{W}$ respectively and $||.||$ denotes the L_2 norm [1].

We make the standard assumption regarding the coherence of the kernel matrix i.e. $||u^{t(i)}|| \leq \frac{\mu_r\sqrt{r}}{\sqrt{n}}, \forall i \in [n]$, where $u^{t(i)}$ denotes i^{th} row of U^t. The proof proceeds by showing that the subspaces spanned by the solution of ALS gets iteratively closer to the optimal subspace spanned by the top-r singular vectors of the original kernel matrix. We use the following inductive structure for the proof:

1. Base case: U^0 is close to U^* and U^0 is incoherent
2. Inductive hypothesis: U^t is close to U^* and U^t is incoherent
3. Inductive step: U^{t+1} is closer to U^* and U^{t+1} is incoherent

Lemma 1. *(Initialization) For rank r Nystrom approximation given by $\hat{K} = KS(S^TKS)_r^\dagger S^TK$, the distance between optimal U^* and the subspace spanned by the columns of $\hat{K}$, $U_{\hat{K}}$ is bounded by:*

$$dist(U^*, U_{\hat{K}}) \leq \sqrt{\frac{\sigma_{r+1}(K)}{\sigma_r(\hat{K})}} \tag{2}$$

where $\sigma_{r+1}(K)$ denotes $r+1^{th}$ largest singular value of K.

Further for large gap in r^{th} eigenvalue of K, using theorem 6 from [8], we get following high probability convergence bound which shows the effectiveness of initialization via Nystrom:

Corollary 1. *(Initialization) Let $\Delta = \frac{\sigma_r(K)}{\sigma_{r+1}(K)}$ be the ratio of consecutive eigenvalues of K. If $\sigma_r(K) - \sigma_{r+1}(K) \geq \frac{12n\ln(2/\delta)}{\sqrt{r}}$ then with probability $1-\delta$*

$$dist(U^*, U_{\hat{K}}) \leq \sqrt{\frac{3}{1+2\Delta}} \tag{3}$$

For the inductive hypothesis in step 2 of the proof, we leverage the results for LRMC from theorem 5.1 of [2].

Corollary 2 (Extending Theorem 2.5 and 5.1 from [2])**.** *In the sampling step, let every entry of K be sampled uniformly and independently with probability,*

$$p \geq C\frac{(\frac{\sigma_1(K)}{\sigma_r(K)})^2\mu^2 r^{2.5}\log n \log\frac{r||K||_F}{\epsilon}}{n\delta_{2r}^2} \tag{4}$$

where $\delta_{2r} \leq \frac{\sigma_r(K)}{12r\sigma_1(K)}$ and $C \geq 0$ is a global constant. The $(t+1)^{th}$ iterate, $\hat{V}^{t+1}$, satisfies the following with probability atleast $1-\frac{1}{n^3}$:

$$dist(\hat{U}^{(t+1)}, U^*) \leq \frac{5}{8}dist(\hat{U}^{(t)}, U^*) \tag{5}$$

Similarly, the coherence of iterates can also be shown.

Corollary 3 (Coherence)**.** *Let $\hat{U}^{(t)}$ be μ_1 incoherent. Then with probability at least $1-\frac{1}{n^3}$, iterate $\hat{U}^{(t+1)}$ is also μ_1 incoherent.*

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8. Available from http://www.i6doc.com/en/.

2.2 Empirical Results

For empirical evaluation we consider the popular RBF kernel. We evaluate our method on several standard datasets taken from LIBSVM website. For empirical comparisons we consider sampling based on mixture of uniform sampling, column selection based on clustering and deterministic selection of diagonal entries. For initialization step, we perform k-means Nystrom. Table 2 shows the mean and standard deviation of spectral error ($||K - \hat{K}||$) for 1000 randomly sampled data points for different kernel approximation methods given the number of parameters (nr) required to store the approximation. Under the space constraints of storing the approximation, KALS shows both lower errors as well as stabler results as indicated by lower standard deviation.

Figure 1, shows the spectral error as the number of parameters are varied by changing the rank r of the approximation. As the number of parameters is increased (by raising rank r of approximation), the spectral error goes down for all the methods. KALS is worse only to SVD which is provably optimal.

Table 2: Comparison of different algorithms given number of parameters for storing approximation.

Dataset	$\#param$	stdNys	kNys	MEKA	KALS	SVD
germanSc	50000	3.36 ± 0.29	2.92 ± 0.11	7.00 ± 2.64	$\mathbf{2.09 \pm 0.07}$	1.67
satimageSc	443500	0.72 ± 0.28	0.30 ± 0.04	1.73 ± 0.24	$\mathbf{0.20 \pm 0.01}$	0.12
wine	649700	1.12 ± 0.23	0.26 ± 0.04	1.16 ± 0.79	$\mathbf{0.17 \pm 0.08}$	0.004
cpusmall ($\times 10^{-6}$)	819200	191.6 ± 211.1	4.64 ± 2.05	875.0 ± 735.5	$\mathbf{1.28 \pm 0.56}$	0.004
cadataSc	4128000	0.67 ± 0.30	0.60 ± 0.15	0.072 ± 0.02	$\mathbf{0.067 \pm 0.04}$	0.0003
ijcnnSc	9998000	0.79 ± 0.07	0.57 ± 0.06	1.03 ± 0.45	$\mathbf{0.38 \pm 0.02}$	0.1307

3 Conclusion

In this paper we proposed kernel approximation via alternating least squares (KALS). The algorithm shows better performance than several other kernel approximation schemes. In future we want to evaluate more sophisticated but efficient sampling schemes which could overcome the incoherency assumption in our analysis. [1]

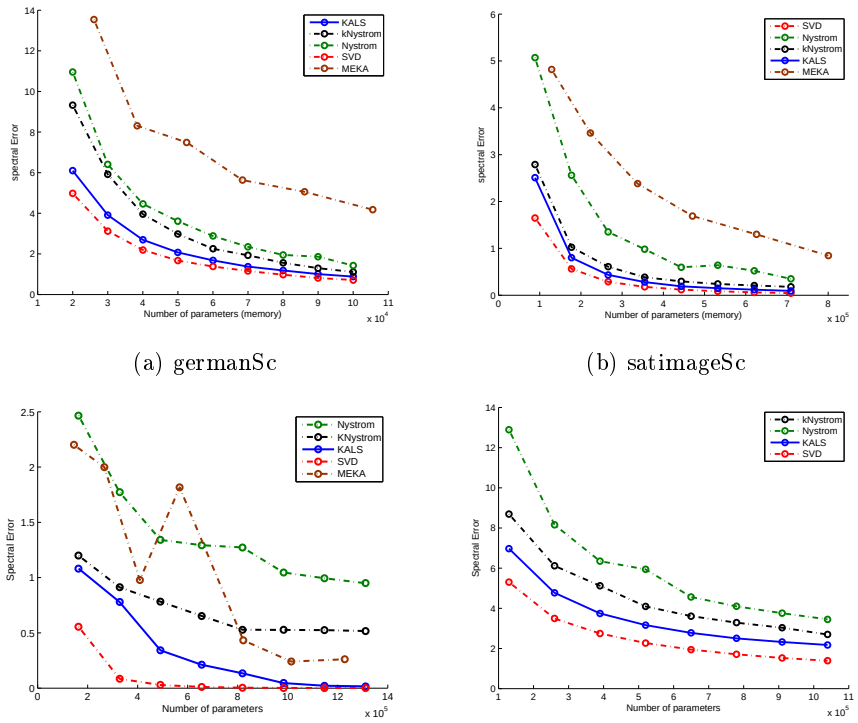
References

[1] Gene H. Golub and Charles F. Van Loan. *Matrix Computations (3rd Ed.)*. Johns Hopkins University Press, Baltimore, MD, USA, 1996.

[2] Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 665–674. ACM, 2013.

[1] Acknowledgement: At the time of work PB was a masters student at IIT Kanpur. We would like to thank Prateek Jain from Microsoft Research India for several fruitful discussions.

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8.
Available from http://www.i6doc.com/en/.



(a) germanSc (b) satimageSc

(c) cpusmallSc (d) wineSc

Fig. 1: Number of parameters (Memory) vs Error for different datasets.

[3] Christopher Williams and Matthias Seeger. Using the nyström method to speed up kernel machines. In *Proceedings of the 14th Annual Conference on Neural Information Processing Systems*, number EPFL-CONF-161322, pages 682–688, 2001.

[4] Petros Drineas and Michael W Mahoney. On the nyström method for approximating a gram matrix for improved kernel-based learning. *The Journal of Machine Learning Research*, 6:2153–2175, 2005.

[5] Kai Zhang, Ivor W Tsang, and James T Kwok. Improved nyström low-rank approximation and error analysis. In *Proceedings of the 25th international conference on Machine learning*, pages 1232–1239. ACM, 2008.

[6] Si Si, Cho-Jui Hsieh, and Inderjit Dhillon. Memory efficient kernel approximation. In *Proceedings of The 31st International Conference on Machine Learning*, pages 701–709, 2014.

[7] Srinadh Bhojanapalli, Prateek Jain, and Sujay Sanghavi. Tighter low-rank approximation via sampling the leveraged element. In *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 902–920. SIAM, 2015.

[8] Mehrdad Mahdavi, Tianbao Yang, and Rong Jin. An improved bound for the nystrom method for large eigengap. *arXiv preprint arXiv:1209.0001*, 2012.