On Speeding Up Computation In Information Theoretic Learning

Sohan Seth and José C. Príncipe
Computational NeuroEngineering Lab, University of Florida, Gainesville

Introduction

With the recent progress in kernel based learning methods, computation with Gram matrices has gained considerable attention. Given n samples $\{x_i\}_{i=1}^n$ and a positive definite function $\kappa(x, y)$, Gram matrix $\mathbf{K}_{XX}$ is defined as,

$$\mathbf{K}_{XX} = \begin{bmatrix} \kappa(x_1, x_1) & \cdots & \kappa(x_1, x_n) \\ \vdots & \ddots & \vdots \\ \kappa(x_n, x_1) & \cdots & \kappa(x_n, x_n) \end{bmatrix}.$$

However, the complexity of computing the entire Gram matrix is quadratic in n. Therefore, a considerable amount of work has been focused on extracting relevant information from the Gram matrix without accessing all the elements [1, 2].
Although information theoretic learning (ITL) is conceptually different from kernel based learning, several ITL estimators can be written in terms of Gram matrices [4]. For example, the estimator of Rényi's quadratic entropy is given by

$$\hat{H}_2(X) = \frac{1}{n^2}\mathbf{1}^\top \mathbf{K}_{XX}\mathbf{1}.$$

However, the difference between ITL and kernel based methods is that ITL estimators might involve a different type of matrix which is neither positive definite nor symmetric. Given samples $\{x_i\}_{i=1}^n$ and $\{y_i\}_{i=1}^n$ and a positive definite function κ, this matrix, $\mathbf{K}_{XY}$ is defined as

$$\mathbf{K}_{XY} = \begin{bmatrix} \kappa(x_1, y_1) & \cdots & \kappa(x_1, y_n) \\ \vdots & \ddots & \vdots \\ \kappa(x_n, y_1) & \cdots & \kappa(x_n, y_n) \end{bmatrix}.$$

This typical matrix appear in several ITL estimators such as in the estimator of cross-information potential which is defined as

$$\hat{\mathcal{CIP}}(X, Y) = \frac{1}{n^2}\mathbf{1}^\top \mathbf{K}_{XY}\mathbf{1}.$$

Incomplete Cholesky decomposition

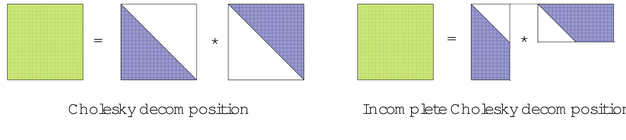
Any $n \times n$ symmetric positive definite matrix $\mathbf{K}$ can be expressed as

$$\mathbf{K} = \mathbf{G}\mathbf{G}^\top.$$

where $\mathbf{G}$ is a $n \times n$ lower triangular matrix with positive diagonal entries. This decomposition is known as the *Cholesky decomposition*. However, if the eigenvalues of $\mathbf{K}$ drops rapidly then the matrix can be approximated by a $n \times d$ $(d \leq n)$ lower triangular matrix $\mathbf{G}$ with arbitrary accuracy i.e.

$$\left\|\mathbf{K} - \mathbf{G}\mathbf{G}^\top\right\| < \epsilon$$

where ϵ is a small positive number of choice and $||\cdot||$ is a suitable matrix norm. This decomposition is called the *incomplete Cholesky decomposition* (ICD). The complexity of computing $\mathbf{G}$ is $\mathcal{O}(nd^2)$ [2].



Evaluation

Using $\mathbf{G}$, $\hat{H}_2(X)$ can be written as

$$\hat{H}_2(X) \approx \frac{1}{n^2}\mathbf{1}^\top \mathbf{G}_{XX}\mathbf{G}_{XX}^\top \mathbf{1} = \frac{1}{n^2}||\mathbf{1}^\top \mathbf{G}_{XX}||_2^2.$$

Thus, complexity of computing $\hat{H}_2(X)$ reduces from $\mathcal{O}(n^2)$ to $\mathcal{O}(nd^2 + nd + d) = \mathcal{O}(nd^2)$. However, similar trick can not be applied to $\hat{\mathcal{CIP}}$.

However, consider the matrix

$$\mathbf{K}_{ZZ} = \begin{bmatrix} \mathbf{K}_{XX} & \mathbf{K}_{XY} \\ \mathbf{K}_{YX} & \mathbf{K}_{YY} \end{bmatrix}$$

where $\mathbf{K}_{YX} = \mathbf{K}_{XY}^\top$. This $2n \times 2n$ matrix can also be generated by the samples

$$\{z_1, \ldots, z_n, z_{n+1}, \ldots, z_{2n}\} = \{x_1, \ldots, x_n, y_1, \ldots, y_n\}$$

such that

$$\mathbf{K}_{ZZ} = \begin{bmatrix} \kappa(z_1, z_1) & \cdots & \kappa(z_1, z_n) \\ \vdots & \ddots & \vdots \\ \kappa(z_n, z_1) & \cdots & \kappa(z_n, z_n) \end{bmatrix}$$

Therefore, this matrix is again symmetric positive definite and we can perform ICD.

Let

$$\mathbf{I} = \begin{bmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{bmatrix} \text{ and } \mathbf{0} = \begin{bmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{bmatrix}.$$

denote the identity and zero matrix respectively. Then

$$\mathbf{K}_{XY} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{K}_{XX} & \mathbf{K}_{XY} \\ \mathbf{K}_{YX} & \mathbf{K}_{YY} \end{bmatrix} \begin{bmatrix} \mathbf{0} \\ \mathbf{I} \end{bmatrix}$$

Define

$$\mathbf{I}_1 = \begin{bmatrix} \mathbf{I} \\ \mathbf{0} \end{bmatrix} \text{ and } \mathbf{I}_2 = \begin{bmatrix} \mathbf{0} \\ \mathbf{I} \end{bmatrix}$$

Then

$$\hat{\mathcal{CIP}} = \frac{1}{n^2}\mathbf{1}^\top \mathbf{I}_1 \mathbf{G}_{ZZ}\mathbf{G}_{ZZ}^\top \mathbf{I}_2 \mathbf{1} = \frac{1}{n^2}(\mathbf{e}_1^\top \mathbf{G}_{ZZ})(\mathbf{G}_{ZZ}^\top \mathbf{e}_2)$$

where

$$\mathbf{e}_1 = \{\underbrace{1, \ldots, 1}_{n}, \underbrace{0, \ldots, 0}_{n}\}^\top \text{ and } \mathbf{e}_2 = \{\underbrace{0, \ldots, 0}_{n}, \underbrace{1, \ldots, 1}_{n}\}^\top.$$

Therefore in the same way, the complexity of computing CIP reduces to $\mathcal{O}(2nd_z^2 + 2nd_z + d_z) \approx O(2nd_z^2)$ from $O(n^2)$.

The similar approach can be extended to other estimators such as estimators of divergence, mutual information and centered correntropy [4]. This approach is particularly useful when we have an estimator that requires $\mathbf{K}_{XX}$,$\mathbf{K}_{YY}$ and $\mathbf{K}_{XY}$ at the same time such as the estimator of correntropy coefficient [3]. In such cases we use

$$\mathbf{K}_{XX} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{K}_{XX} & \mathbf{K}_{XY} \\ \mathbf{K}_{YX} & \mathbf{K}_{YY} \end{bmatrix} \begin{bmatrix} \mathbf{I} \\ \mathbf{0} \end{bmatrix} \text{ and } \mathbf{K}_{YY} = \begin{bmatrix} \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{K}_{XX} & \mathbf{K}_{XY} \\ \mathbf{K}_{YX} & \mathbf{K}_{YY} \end{bmatrix} \begin{bmatrix} \mathbf{0} \\ \mathbf{I} \end{bmatrix}$$

and apply similar approach.

Simulation

Parameters: $\kappa(x, y) = \frac{1}{\sqrt{\pi}} \exp\left(-(x - y)^2\right)$, $\epsilon = 10^{-6}$.

Table 1: Description of the datasets

	IRIS	WINE	CANCER	YEAST	ABALONE
Features	4	13	32	8	8
Samples	150	178	198	1484	4177

Table 2: Total time of computing correntropy coefficient between all possible pairs of variables

	Direct Method		Optimized method	
Dataset	Value	Time (s)	Value	Time (s)
IRIS	1.5685	0.67	1.5685	0.04
CANCER	3.8530	12.15	3.8530	0.6
WINE	76.2108	95.0	76.2108	4.4
YEAST	0.3031	301.9	0.3031	3.19
ABALONE	19.0452	2447.2	19.0452	12.7

Table 3: Total time of computing Cauchy-Schwartz quadratic mutual information between all possible pairs of variables

	Direct Method		Optimized method	
Dataset	Value	Time (s)	Value	Time (s)
IRIS	1.5109	0.36	1.5109	0.04
CANCER	5.5022	6.76	5.5022	0.7
WINE	20.0637	53.7	20.0637	4.8
YEAST	0.1142	201.2	0.1142	1.65
ABALONE	6.711	2162.4	6.711	8.5

Summary

We suggest the use of incomplete Cholesky decomposition to reduce the computational cost of ITL estimators. We experimentally verify that the proposed approach reduces the computation cost drastically. However, it should be noted that we assume the existence of a low rank approximation of the Gram matrix which might not be always available in practice. Finally, a bound on the absolute difference between the actual and estimated statistic in terms of the precision parameter ϵ would be interesting to see.

References

[1] F. R. Bach and M. I. Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, 3:1–48, 2002.

[2] Shai Fine, Katya Scheinberg, Nello Cristianini, John Shawe-taylor, and Bob Williamson. Efficient svm training using low-rank kernel representations. *Journal of Machine Learning Research*, 2:243–264, 2001.

[3] J.-W. Xu, H. Bakardjian, A. Cichocki, and J. C. Principe. A new nonlinear similarity measure for multichannel signals. *Neural Netowrks*, 21:222–231, 2008.

[4] J. W. Xu, A. R. C. Paiva, I. Park, and J. C. Principe. A reproducing kernel hilbert space framework for information-theoretic learning. *Signal Processing, IEEE Transactions on*, 56(12):5891–5902, 2008.