Block-Quantized Kernel Matrix for Fast Spectral Embedding

Kai Zhang     James T. Kwok

Department of Computer Science and Engineering
The Hong Kong University of Science and Technology
Hong Kong
Outline

1 Introduction
   - Eigendecomposition of Kernel Matrix
   - Scale-Up Methods

2 The Proposed Method
   - Gram Matrix of Special Forms
   - Basic Idea
   - Matrix Approximation
   - Matrix Quantization
   - Density Weighted Nyström Extension

3 Experiments
   - Kernel Principal Component Analysis
   - Image Segmentation

4 Conclusion
Outline

1 Introduction
   - Eigendecomposition of Kernel Matrix
   - Scale-Up Methods

2 The Proposed Method
   - Gram Matrix of Special Forms
   - Basic Idea
   - Matrix Approximation
   - Matrix Quantization
   - Density Weighted Nyström Extension

3 Experiments
   - Kernel Principal Component Analysis
   - Image Segmentation

4 Conclusion
Eigen-decomposition of Kernel Matrix

When do we need to eigen-decompose the kernel matrix?

**Kernel Principle Component Analysis**
A powerful tool to extract nonlinear structures in the high dimensional feature space (Schölkopf 1998).

**Spectral Clustering**
A global, pairwise clustering method based on graph partitioning theories (Shi & Malik, 2000).

**Manifold Learning and Dimensionality Reduction**
Laplacian Eigenmap, ISOMAP, Locally linear Embedding...
Scale-Up Methods

Low-rank approximation of the form $L = GG'$, where $L \in \mathbb{R}^{N \times N}$, $G \in \mathbb{R}^{N \times m}$ and $m \ll N$ is the rank

- Incomplete Cholesky decomposition (Bach & Jordan, 2002; Fine & Scheinberg, 2001)
- Sparse greedy kernel methods (Smola & Bartlett, 2000)

Sampling-based methods

- Nyström: randomly selects columns of the kernel matrix (Williams & Segger, 2001; Lawrence & Herbrich, 2005)
- Drineas & Mahoney (2005): chooses the columns based on a data-dependent probability
- Ouimet and Bengio (2005): uses a greedy sampling scheme based on the feature space geometry
Gram Matrix of Special Forms

Block Quantized Matrices

\[
W = \begin{cases} 
    a & a & b & b & b \\
    a & a & b & b & b \\
    c & c & d & d & d \\
    c & c & d & d & d \\
    c & c & d & d & d 
\end{cases}
\]

**Definition**

1. The block-quantized matrix \( W \) contains \( m^2 \) constant blocks.
2. The block at the \( i \)th row and \( j \)th column, \( C_{ij} \), has dimension \( n_i \times n_j \), with entry value \( \beta_{ij} \).
3. E.g., \( n_1 = 2, n_2 = 3, \beta_{11} = a, \beta_{12} = b, \beta_{21} = c, \beta_{22} = d \).

**Note**

Block quantization can be performed by:

1. partition the data set into \( m \) clusters;
2. set \( \beta_{ij} = K(t_i, t_j) \) \((i, j = 1, 2, \ldots, m)\), where \( t_i \) is the representative of the \( i \)th cluster.
Gram Matrix of Special Forms

Properties of Block Quantized Matrices

**Eigensystem of** $\mathbf{W}$, $\mathbf{W}\phi = \lambda\phi$

$$
\begin{bmatrix}
a & a & b & b & b \\
a & a & b & b & b \\
c & c & d & d & d \\
c & c & d & d & d \\
c & c & d & d & d \\
\end{bmatrix}
\begin{bmatrix}
\phi_1 \\
\phi_2 \\
\phi_3 \\
\phi_4 \\
\phi_5 \\
\end{bmatrix}
= \lambda
\begin{bmatrix}
\tilde{\phi}_1 \\
\tilde{\phi}_2 \\
\tilde{\phi}_3 \\
\tilde{\phi}_4 \\
\tilde{\phi}_5 \\
\end{bmatrix}
$$

The first $n_1$ equations are the same, so are the next $n_2$ equations,..., and so on. It is equal to the $m \times m$ system

$$
\mathbf{\tilde{W}}\tilde{\phi}_j = \lambda\tilde{\phi}_i, \text{ where } \mathbf{\tilde{W}}_{ij} = \beta_{ij}n_j.
$$

**How to recover the eigensystem of** $\mathbf{W}$ **from that of** $\mathbf{\tilde{W}}$?

- **Eigenvalues**: $\mathbf{W}$ and $\mathbf{\tilde{W}}$ have the same eigenvalues.
- **Eigenvectors**: repeat the $k$th entry of $\tilde{\phi}$ $n_k$ times, then we get $\phi$ (i.e., $\phi$ is piecewise constant).
Basic Idea

Utilize the blockwise structure of the kernel matrix $W$ to compute the eigen-decomposition more efficiently.

Procedure

1. Find a blockwise-constant matrix $\tilde{W}$ to approximate $W$.
   - Use the Frobenius norm $\|W - \tilde{W}\|_F$ as the approximation criteria.

2. The eigen-system of the $N \times N$ matrix $\tilde{W}$ can be fully recovered from that of the $m \times m$ matrix $\tilde{W}$.
   - Use this as an approximate solution to the eigen-decomposition of $W$. 
Matrix perturbation theory [Bhatia, 1992]

- Difference between two matrices can bound the difference between their singular value spectra.
- If $A, E \in \mathbb{R}^{m \times n}$, and $\sigma_k(A)$ is the $k$th singular value of $A$, then

\[
\max_{1 \leq t \leq n} |\sigma_t(A + E) - \sigma_t(A)| \leq \|E\|_2,
\]

\[
\sum_{k=1}^{n} (\sigma_k(A + E) - \sigma_k(A))^2 \leq \|E\|_F^2.
\]
Approximation of Eigenvectors

Our Analysis

- In some cases the eigenvectors are of greater importance, such as in manifold embedding, spectral clustering, etc.
- Let $W$ and $\overline{W}$ be the original and block-quantized matrices, with eigen-value/vector pair $(\alpha, \mu)$ and $(\beta, \nu)$, respectively. Then we have

\[
\|\mu - \nu\| \leq \begin{cases} 
\left(\frac{1}{\alpha} + \frac{1}{\beta}\right) \|W\|_2 + \frac{1}{\beta} \|E\|_2, & \alpha \leq \beta, \\
\left(\frac{3}{\beta} - \frac{1}{\alpha}\right) \|W\|_2 + \frac{1}{\beta} \|E\|_2, & \alpha > \beta.
\end{cases}
\]

- Since $\|E\|_2 \leq \|E\|_F$, therefore by minimizing $\|E\|_F$, we can also bound the approximation error of the eigenvectors.
Minimization of the Matrix Approximation Error

- The objective $E = \| W - \overline{W} \|_F$ can be written as
  
  $$E = \sum_{i,j=1}^{N} \left( W_{ij} - \overline{W}_{ij} \right)^2 = \sum_{i,j=1}^{m} \sum_{x_p \in S_i, x_q \in S_j} (W_{pq} - \beta_{ij})^2 .$$

- Can be minimized by setting $\frac{\partial E}{\partial \beta_{ij}} = 0$ to obtain
  
  $$\beta_{ij} = \frac{1}{n_in_j} \sum_{x_p \in S_i, x_q \in S_j} K(x_p, x_q).$$

- Takes $O(N^2)$ time to compute the $\beta_{ij}$'s.
Matrix Quantization

Data Partitioning

---

**Assumption**

Suppose the data set is partitioned into clusters in the input space:

- Local cluster $S_i$ has a minimum enclosing ball (MEB) with radius $r_i$.
- The cluster representative $t_i$ should fall into this MEB.

---

**Question**

How does the partitioning influence the matrix approximation quality?
Approximation error vs. Data Partitioning

Upper Bound

The approximation error $E$ is bounded by

$$ E \leq 64N^2 \xi^2 R^2 \frac{1}{\sigma^4} \left( \overline{D^2} + 4R^2 + 4\overline{DR} \right), $$

- $\sigma$, width of the (stationary) kernel $K(x, y) = k \left( \| \frac{x-y}{\sigma} \| \right)$.
- $\xi = \max |k'(x)|$.
- $R = \max_{i=1,2,...,m} r_i$, maximum MEB radius.
- $\overline{D} = \frac{1}{N^2} \sum_{ij} n_i n_j D_{ij}$, average pairwise distance.
- $\overline{D^2} = \frac{1}{N^2} \sum_{ij} n_i n_j D_{ij}^2$, average pairwise squared distance.
Sequential Sampling

**Objective**

Partition the data set into compact local clusters, such that every point is close to its cluster center.

**Procedure**

1. Randomly select a sample to initialize the cluster center set $C = \{t_1\}$. For $i = 1, 2, \ldots, N$, do the following.
2. Compute $l_{ij} = \|x_i - t_j\|$, $t_j \in C$. Once $l_{ij} \leq r$, assign $x_i$ to $S_j$, let $i = i + 1$, and go to the next step.
3. If $\|x_i - t_j\| > r$, $\forall t_j \in C$, add $x_i$ to $C$ as a new center. Let $i = i + 1$ and go to the next step.
4. On termination, count the number of samples, $n_j$, in $S_j$, and update each $t_j \in C$ as $t_j = \frac{1}{n_j} \sum_{x_i \in S_j} x_i$. 
Example: Sequential Sampling

Data (left); Small threshold $r$ (middle); Large threshold (right)

Property

- The local clusters are bounded by the hypercube of side length $2r$, where $r$ is the partitioning parameter.
- The complexity is $O(N \log m)$ by using a hierarchical implementation.
The approximation error $E = \| W - \overline{W} \|_F$ can be written as a function of the cluster representatives $t_i$’s,

$$E = \sum_{i,j=1}^{m} \sum_{p \in S_i, q \in S_j} (K(x_p, x_q) - K(t_i, t_j))^2$$

which can be optimized using gradient descent

$$t_k = \frac{\sum_{j \neq k} t_j \left( B_{kj} K \left( \| t_k - t_j \|^2 / \sigma^2 \right) - A_{kj} K^2 \left( \| t_k - t_j \|^2 / \sigma^2 \right) \right)}{\sum_{j \neq k} B_{kj} K \left( \| t_k - t_j \|^2 / \sigma^2 \right) - A_{kj} K^2 \left( \| t_k - t_j \|^2 / \sigma^2 \right)}.$$

Here $A_{ij} = n_i n_j$, $B_{ij} = \sum_{p \in S_i, q \in S_j} K(x_p, x_q)$. The iteration can fine tune the cluster representatives especially when $m$ is small.
We refine $\phi$ through the Nyström extension. But we can incorporate the “cluster” information by

$$\phi_k(x) = \frac{1}{N\lambda_k} \sum_{i=1}^{m} n_i \phi_k(x_i) W(x, t_i),$$

- It is believed to be difficult to directly use the density information in high-dimensional problems.
- However, $n_i$’s can be deemed as coefficients of a multidimensional histogram. It greatly improves the convergence behavior of the Nyström extension.
Outline

1. **Introduction**
   - Eigendecomposition of Kernel Matrix
   - Scale-Up Methods

2. **The Proposed Method**
   - Gram Matrix of Special Forms
   - Basic Idea
   - Matrix Approximation
   - Matrix Quantization
   - Density Weighted Nyström Extension

3. **Experiments**
   - Kernel Principal Component Analysis
   - Image Segmentation

4. **Conclusion**
Kernel Principal Component Analysis

MNIST digit image data set (digit 0 and 1)
- Image size: $28 \times 28$ (dimension 784)
- 2000 training samples and 2000 testing samples
- Gaussian kernel with bandwidth $\sigma = 30$

Our algorithm is compared with the Nyström method using different sampling methods:

1. random subset;
2. sequential sampling;
3. vector quantization.

Embedding results are aligned with the standard KPCA and the mean squared error computed.
Embedding on the 3 leading eigen-directions of KPCA (left); Gradient method using 3 representatives (middle) and sequential sampling using 10 representatives (right).

Our embedding results are faithful although the number of representatives are quite few.
In-sample (left) and out-of-sample (right) embedding errors on the 3 leading principal directions of different methods, with reference to the standard KPCA embedding. In both cases, our method is superior to the other algorithms.
Approximation errors of the top (left), second (middle) and third (right) principal eigenvectors of different methods.

1. The larger the eigenvalue, the easier the approximation;
2. Our method is superior to the other algorithms.
Kernel Principal Component Analysis

**Total time (in secs) and #Representatives (m)**

Machine: 2.26GHz Pentium-3 PC.

Time consumption (secs) and #Representatives under different partitioning thresholds ($r$)

<table>
<thead>
<tr>
<th>$r^2$</th>
<th>140</th>
<th>110</th>
<th>90</th>
<th>70</th>
<th>60</th>
<th>52</th>
<th>44</th>
<th>36</th>
</tr>
</thead>
<tbody>
<tr>
<td>$m$</td>
<td>3</td>
<td>10</td>
<td>19</td>
<td>47</td>
<td>87</td>
<td>150</td>
<td>226</td>
<td>382</td>
</tr>
<tr>
<td>time</td>
<td>0.04</td>
<td>0.09</td>
<td>0.17</td>
<td>0.48</td>
<td>0.95</td>
<td>1.82</td>
<td>3.57</td>
<td>8.55</td>
</tr>
</tbody>
</table>

- Standard KPCA takes about 87 secs.
- For our method, the approximation quality is satisfactory when $r^2 < 90$. The corresponding time consumption is one order of magnitude smaller than KPCA.
Embedding error versus the number of representatives. Our method has the lowest errors.
Experimental Setting

Berkeley image segmentation benchmark dataset

- Image size $481 \times 321$

Normalized cut

- Similarity measures: pixel color (RGB) and position (XY) (both normalized to the domain $[0,255]$)

Gaussian kernel

- Bandwidth $\sigma \in [20, 40]$
Image Segmentation

Segmentation Results (1)

\[
m = 114, \ 0.45s
\]

\[
m = 162, \ 0.91s
\]
Image Segmentation

Segmentation Results (2)

$m = 175, 0.66s$

$m = 89, 0.31s$
Comparison of Segmentation Results

(1) original image; (2a) our segmentation; (2b) boundary; (3a) Nyström (random sampling); (3b) boundary; (4a) Nyström (sequential sampling); (4b) boundary; (5a) Nyström (VQ); (5b) boundary.
Outline

1 Introduction
   - Eigendecomposition of Kernel Matrix
   - Scale-Up Methods

2 The Proposed Method
   - Gram Matrix of Special Forms
   - Basic Idea
   - Matrix Approximation
   - Matrix Quantization
   - Density Weighted Nyström Extension

3 Experiments
   - Kernel Principal Component Analysis
   - Image Segmentation

4 Conclusion
Conclusions

Summary

- Proposed an efficient approach for eigen-decomposition of kernel matrices.
- The complexity $O(mN)$ is lower than most existing approaches.
- By incorporating density information, our method greatly reduces the data representatives needed, and improves the convergence behavior of the Nyström algorithm.