Nyström Based Kernel Classification for Big Data

September 18, 2015
1. Introduction
   - Background
   - Challenge for Big Data

2. Approximation Kernel Using Nyström Method
   - Nyström Method
   - Nyström based Kernel Classification

3. Experimental Results

4. Conclusion
Background

Classical Setting in Machine Learning

- $n$ training examples: $\{(x_i, y_i)\}_{i=1}^n$ where $x_i \in \mathcal{X} \subseteq \mathbb{R}^d, y_i \in \mathcal{Y}$. 

Background

Classical Setting in Machine Learning

- $n$ training examples: $\{(x_i, y_i)\}_{i=1}^n$ where $x_i \in \mathcal{X} \subseteq \mathbb{R}^d, y_i \in \mathcal{Y}$. 

Nyström Based Kernel Classification for Big Data
Background

Classical Setting in Machine Learning

- $n$ training examples: $\{(x_i, y_i)\}_{i=1}^n$ where $x_i \in \mathcal{X} \subseteq \mathbb{R}^d, y_i \in \mathcal{Y}$. 

feature representation
Background

Classical Setting in Machine Learning

- $n$ training examples: $\{(x_i, y_i)\}_{i=1}^n$ where $x_i \in \mathcal{X} \subseteq \mathbb{R}^d, y_i \in \mathcal{Y}$. 

target variable
Background

Classical Setting in Machine Learning

- $n$ training examples: $\{(x_i, y_i)\}_{i=1}^{n}$ where $x_i \in \mathcal{X} \subseteq \mathbb{R}^d, y_i \in \mathcal{Y}$.
- The goal of machine learning is to learn a predictive function $f(x) : \mathcal{X} \rightarrow \mathcal{Y}$

- Classification: $\mathcal{Y} = \{-1, +1\}$
- Regression: $\mathcal{Y} \subseteq \mathbb{R}$
Background

Linear Model

- Predictive function $f(x) = w^T x$
- Optimization problem

$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \ell(w^T x_i, y_i) + \lambda R(w)$$

- Example: SVM, Logistic Regression
- Pros: efficient to solve
- Cons: suffer lower performance when data points are not linearly separable
Background

Non-Linear Model
Background

Non-Linear Model

- Predictive function $f(x) = \mathbf{w}^T \phi(x)$, where $\phi(\cdot)$ is mapping function
- Optimization problem (Primal form):

$$
\min_{f \in \mathcal{H}_K} \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i) + \frac{\lambda}{2} \|f\|^2_{\mathcal{H}_K} \tag{1}
$$

- Optimization problem (Dual form):

$$
\max_{\alpha \in \mathbb{R}^n} - \frac{1}{n} \sum_{i=1}^{n} \ell^*(\alpha_i) - \frac{1}{2\lambda n^2} \alpha^T K \alpha \tag{2}
$$

- Where $\ell^*$ is the conjugate function of loss function $\ell$ and $K \in \mathbb{R}^{n \times n}$ is kernel matrix
Background

Non-Linear Model

- Kernel trick $\kappa(\mathbf{x}, \mathbf{y}) = \phi(\mathbf{x})^T \phi(\mathbf{y})$

$$K = \begin{bmatrix}
\phi(\mathbf{x}_1)^T \phi(\mathbf{x}_1) & \cdots & \phi(\mathbf{x}_1)^T \phi(\mathbf{x}_n) \\
\phi(\mathbf{x}_2)^T \phi(\mathbf{x}_1) & \cdots & \phi(\mathbf{x}_2)^T \phi(\mathbf{x}_n) \\
\vdots & \vdots & \vdots \\
\phi(\mathbf{x}_n)^T \phi(\mathbf{x}_1) & \cdots & \phi(\mathbf{x}_n)^T \phi(\mathbf{x}_n)
\end{bmatrix} = \begin{bmatrix}
\kappa(\mathbf{x}_1, \mathbf{x}_1) & \cdots & \kappa(\mathbf{x}_1, \mathbf{x}_n) \\
\kappa(\mathbf{x}_2, \mathbf{x}_1) & \cdots & \kappa(\mathbf{x}_2, \mathbf{x}_n) \\
\vdots & \vdots & \vdots \\
\kappa(\mathbf{x}_n, \mathbf{x}_1) & \cdots & \kappa(\mathbf{x}_n, \mathbf{x}_n)
\end{bmatrix}$$

- Kernel function: $\kappa(\mathbf{x}, \mathbf{y}) = \exp(-\gamma \|\mathbf{x} - \mathbf{y}\|^2)$, $\kappa(\mathbf{x}, \mathbf{y}) = (a\mathbf{x}^T \mathbf{y} + c)^d$

- Pros: enjoys high performance

- Cons: hard to solve
Challenge in Big Data

- $X$ is $n \times d$ matrix, $n \rightarrow$ millions, billions, · · ·
- Need to compute kernel matrix $K$

$$K = \begin{bmatrix}
\kappa(x_1, x_1) & \cdots & \kappa(x_1, x_n) \\
\kappa(x_2, x_1) & \cdots & \kappa(x_2, x_n) \\
\vdots & \vdots & \vdots \\
\kappa(x_n, x_1) & \cdots & \kappa(x_n, x_n)
\end{bmatrix}$$

- Very very large kernel matrix $K$
- Computation and memory cost
Nyström Method

Nyström based Kernel Classification for Big Data

\[ K \approx K_bK_m^\dagger K_b^T \]
Theorems about Nyström Method

Theorem [Drineas and Mahoney, 2005]

For any $m$ uniformly sampled columns, with a high probability,

$$
\| K - K_b K_m^\dagger K_b^T \|_2 = O\left( \frac{n}{\sqrt{m}} \right)
$$

Theorem [Jin et al., 2011]

For any $m$ uniformly sampled columns, assume there exists $\rho \in (0, 1/2)$ such that $\lambda_m = \Omega(n/m^\rho)$ and $\lambda_{m+1} = O(n/m^{1-\rho})$, with a high probability,

$$
\| K - K_b K_m^\dagger K_b^T \|_2 = O\left( \frac{n}{m^{1-\rho}} \right)
$$
Nyström based Kernel Classification

- Nyström approximation dual optimization

\[
\max_{\alpha \in \mathbb{R}^n} - \frac{1}{n} \sum_{i=1}^{n} \ell_i^*(\alpha_i) - \frac{1}{2\lambda n^2} \alpha^T (K_b K_m^T K_b)\alpha
\]

- Short feature representation

\[
\hat{K} = K_b K_m^T K_b^T
\]
\[
= K_b V D^{-1} V^T K_b^T
\]
\[
= (D^{-1/2} V^T K_b^T)^T (D^{-1/2} V^T K_b^T)
\]
\[
= \hat{X}^T \hat{X}
\]

- Recall \( K = \Phi(X)^T \Phi(X) \) while \( \hat{K} = \hat{X}^T \hat{X} \)
Nyström based Kernel Classification

- Nyström approximation dual optimization with short feature representation

\[
\max_{\alpha \in \mathbb{R}^n} - \frac{1}{n} \sum_{i=1}^{n} \ell_i^*(\alpha_i) - \frac{1}{2\lambda n^2} \alpha^T \hat{X}^T \hat{X} \alpha
\]  

(4)

\[
\hat{X} = D^{-1/2} V^T K_b^T
\]

- \(X \in \mathbb{R}^{n \times d}, \Phi(X) \in \mathbb{R}^{n \times hd}, \hat{X} \in \mathbb{R}^{n \times m}\)
Nyström based Kernel Classification

- Draw $m$ samples $\{x_{i_1}, \cdots, x_{i_m}\}$ from $n$ examples
- Compute sub-kernel matrix $K_m \in \mathbb{R}^{m \times m}$ among $m$ samples and $K_b \in \mathbb{R}^{n \times m}$ between all examples and $m$ samples
- Singular Value Decomposition on $K_m = VDV^T$
- $\hat{X} = D^{-1/2}V^TK_b^T$
- Reduced the kernel problem to linear model
### Statistic of Datasets

<table>
<thead>
<tr>
<th>Name</th>
<th>usps</th>
<th>letter</th>
<th>ijcnn1</th>
<th>webspm</th>
<th>cod-rna</th>
<th>covtype</th>
</tr>
</thead>
<tbody>
<tr>
<td>#Training</td>
<td>7,291</td>
<td>12,000</td>
<td>91,701</td>
<td>280,000</td>
<td>271,617</td>
<td>464,810</td>
</tr>
<tr>
<td>#Testing</td>
<td>2,007</td>
<td>6000</td>
<td>49,990</td>
<td>70,000</td>
<td>59,535</td>
<td>116,202</td>
</tr>
<tr>
<td>#Features</td>
<td>256</td>
<td>16</td>
<td>22</td>
<td>254</td>
<td>8</td>
<td>54</td>
</tr>
</tbody>
</table>
Compare linear-SVM, Kernel SVM, and Nyström Method using $m = 512, 1024, 2048$ and $4096$
Adding $\ell_1$ regularization

- Approximation error

$$\max_{\alpha \in \mathbb{R}^n} - \frac{1}{n} \sum_{i=1}^{n} \ell_i^*(\alpha_i) - \frac{1}{2\lambda n^2} \alpha^T \hat{X}^T \hat{X} \alpha - \frac{\tau}{n} \|\alpha\|_1$$ (5)
Analysis

- Approximation error

\[
\max_{\alpha \in \mathbb{R}^n} - \frac{1}{n} \sum_{i=1}^{n} \ell_i^*(\alpha_i) - \frac{1}{2\lambda n^2} \alpha^T \hat{X}^T \hat{X} \alpha - \frac{\tau}{n} \| \alpha \|_1
\]  

(6)

Lemma

Let \( S \) be the support set of \( \alpha_* \) and \( S^c \) denote its complement. By setting \( \tau \geq \frac{2}{\lambda n} \sum_{i=1}^{n} |[\alpha_*]_i| \| \hat{K}_{*i} - K_{*i} \|_\infty \), we have

\[
\| [\tilde{\alpha}_* - \alpha_*]_{S^c} \|_1 \leq 3 \| [\tilde{\alpha}_* - \alpha_*]_{S} \|_1.
\]
Analysis

- Proof:
Analysis

Lemma

Let \( q = \frac{1}{n}X^T(A^TA - I)e \). With a probability at least \( 1 - \delta \), we have

\[
\|q\|_{\infty} \leq \frac{c\eta R}{n} \sqrt{\frac{\log(d/\delta)}{m}},
\]

where \( c \) is the universal constant in the JL lemma, \( \|e\|_2 \leq \eta \) and \( \max_{1 < j < d} \|x_j\|_2 \leq R \)

- In our case, we need to similar \( \Delta = \frac{1}{\lambda n}(\hat{X}^T\hat{X} - X^TX)\alpha^* \), Can we use the same strategy?
Analysis
Adding $\ell_1$ regularization
Two Strategies to Refine Nystrom

- Probability sampling data points

\[ Pr(X_i \text{ is selected}) = \frac{|\tilde{\alpha}_i|}{\sum_{i=1}^{n} |\tilde{\alpha}_i|} \]  \hfill (8)

- Weighted kmean to constructed data points

\[ \min \sum_{i=1}^{n} [\alpha_i]^2 ||x_i - c_{\pi_i}||^2 \]  \hfill (9)
adding $\ell_1$ Nyström, $\ell_1$-pro-Nyström and weight-kmean-Nyström
Time complexity

- Linear SVM, Standard-Nyström, Refined-Nyström and Kernel SVM

![Graph showing time complexity for different methods on two datasets: web spam and cod-rna.](image-url)
Conclusion and Future Work

- Nyström method is a powerful method for matrix approximation.
- Nyström based kernel classification can achieve high performance with less computation and memory.
- Adding $\ell_1$ norm on Nyström based on kernel classification can improve test accuracy.
- Theoretical analysis for Adding $\ell_1$ norm on Nyström based on kernel classification is the future work.