Sparse Kernel Models for Spectral Clustering Using the Incomplete Cholesky Decomposition

Carlos Alzate
Johan Suykens

June 5, 2008

ESAT-SCD-SISTA
K. U. Leuven
Leuven, Belgium
Outline

1. Spectral Clustering and Kernel PCA
2. Sparse Kernels Models for Spectral Clustering
3. Experimental Results
4. Conclusions
Motivation

Spectral Clustering Facts

- Can detect complex nonlinear structures.
- Weighted version of kernel PCA.
- Out-of-sample extensions via projections.
- The projections are dense and every training point contributes.
- The affinity matrix is huge for large-scale problems.
Spectral Clustering

Definition

- Graph nodes → data points, edges → affinities.
- The objective is to partition the graph into $k$ disjoint sets by minimizing some cut criterion (cutting edges).
- Graph partitioning is NP-hard (combinatorial problem).
- Relaxing the problem leads to eigenvalue problems where the eigenvectors contain clustering information.
- Converting the eigenvectors back to clusters is not straightforward.
Typical Spectral Clustering Algorithm
Multiway LS-SVM formulation related to kernel PCA [Alzate and Suykens, 2007]

**Primal form**

\[
\min_{w^{(l)}, e^{(l)}, b_l} \frac{1}{2N} \sum_{l=1}^{k-1} \gamma_l e^{(l)T} D^{-1} e^{(l)} - \frac{1}{2} \sum_{l=1}^{k-1} w^{(l)T} w^{(l)}
\]

such that

\[
\begin{align*}
    e^{(1)} &= \Phi w^{(1)} + b_1 1_N \\
    &\vdots \\
    e^{(k-1)} &= \Phi w^{(k-1)} + b_{k-1} 1_N
\end{align*}
\]

**Advantages**

- Out-of-sample extensions via projections.
- Bias terms lead to a different centering.
- Additional constraints can be added.
- Model selection can be performed using the projections.

**Dual form**

\[
M \Omega \alpha = \lambda \alpha,
\]

where \( M = D^{-1} - \frac{D^{-1} 1_N 1_N^T D^{-1}}{1_N^T D^{-1} 1_N} \).
Spectral Clustering and Kernel PCA

Multiway LS-SVM formulation related to kernel PCA [Alzate and Suykens, 2007]

Computational Issues solving $M\Omega\alpha = \lambda\alpha$ for large $N$

- $M$ and $\Omega$ are $\mathbb{R}^{N \times N}$ matrices where $N$ is the number of data points.
- For a $200 \times 200$ pixels image $\rightarrow N = 40,000$.
- The projections of an out-of-sample data point $x$

$$z(x) = w^T\varphi(x) + b = \sum_{i=1}^{N} \alpha_i K(x_i, x) + b,$$

are dense kernel expansions.
The Incomplete Cholesky Factorization

Definition

- Given $A \in \mathbb{R}^{N \times N}, A = A^T, A > 0$, the Cholesky factorization is $A = LL^T$ where $L \in \mathbb{R}^{N \times N}$.
- If $A \geq 0$, an incomplete Cholesky factorization is still possible:

$$A \approx GG^T, G \in \mathbb{R}^{N \times R},$$

where the quality of the approximation is given by:

$$||A - GG^T||_2^2 \leq \eta, R \leq N$$

and $\eta$ is a user-defined threshold.
- If the eigenspectrum of $A$ decays fast $\rightarrow R \ll N$

---

[Fine and Scheinberg, 2001] (interior point methods for SVM) and [Alzate and Suykens, 2008] (contrast function for ICA).
How to solve $M\Omega\alpha = \lambda\alpha$ when $N$ is large?

**Ideas**
- The eigenspectrum of RBF kernel matrices is known to decay fast.
- The kernel matrix can be approximated by $\Omega \approx GG^T$.
- $G$ can be computed iteratively (i.e. no need to store $\Omega$).
- The rows of $M$ can be computed on demand so no need to store it.

**Approximated Eigenvectors**
- The SVD of $G \rightarrow G = U\Lambda V^T$ therefore $\Omega \approx U\Lambda^2 U^T$
- $M\Omega\alpha = \lambda\alpha$ can be written as
  \[ U^T MU \Lambda^2 \rho = \lambda \rho \text{ with } \rho = U^T \alpha \]

which is a $R \times R$ eigenvalue problem and $R \ll N$. 
How to sparsify the projections?

**Ideas**

- **Approximate** $w = \sum_{i=1}^{N} \alpha_i \varphi(x_i) = \Phi^T \alpha$ with a reduced set method:

  $$\tilde{w} = \sum_{m=1}^{N_s} \beta_m \varphi(\tilde{x}_m) = \Psi^T \beta$$

  where $\beta = [\beta_1, \ldots, \beta_{N_s}]$ is the vector of reduced set coefficients, $\mathcal{R} = \{\tilde{x}_m\}_{m=1}^{N_s}$ is the reduced set and $N_s < N$.

- **Two subproblems**: How to find the reduced set coefficients $\beta$ and how to obtain the reduced set $\mathcal{R}$. 
How to sparsify the projections?

Subproblem 1: Finding the reduced set coefficients $\beta$

Minimize the squared two-norm of the residuals with respect to $\beta$:

$$\min_{\beta} ||w - \tilde{w}||_2^2.$$ 

Using $\frac{\partial ||w - \tilde{w}||_2^2}{\partial \beta} = 0$ leads to

$$\Omega_{\Psi\Psi} \beta^{(l)} = \Omega_{\Psi\Phi} \alpha^{(l)}$$

which is a linear system $^a$ to be solved in $\beta$.

---

$^a$A different approach for sparse expansions was also mentioned in [Schölkopf et al., 1999]. This approach uses $L_1$ shrinkage penalizers and the reduced set coefficients can be obtained by solving a QP problem.
How to sparsify the projections?

Subproblem 2: Finding the reduced set $\mathcal{R}$
- The incomplete Cholesky factorization of $\Omega$ already provides a reduced set (the pivots).
- The degree of sparseness is controlled via the Cholesky error threshold $\eta$.

Sparse projections

The sparse projections for out-of-sample data become:

$$z_t^{(l)} \approx \sum_{m=1}^{R} \beta_m^{(l)} K(x_t^{\text{test}}, \tilde{x}_m)$$

where the reduced set points $\tilde{x}_m$ are given by the pivots of the incomplete Cholesky factorization.
Experimental Results

Toy problem 1: Three gaussian clouds in 3-D

$N = 6,000, \sigma^2 = 0.5, \eta = 10^{-3}, R_s = 78.$
Experimental Results

Toy problem 2: Large-scaled intertwined spirals

\[ N = 40,000, \sigma^2 = 0.5, \eta = 0.5, R_s = 237. \]
Experimental Results

Toy problem 2: Large-scaled intertwined spirals

<table>
<thead>
<tr>
<th>Training set size $N$</th>
<th>Reduced set size $R$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,000</td>
<td>168 (83.2%)</td>
</tr>
<tr>
<td>5,000</td>
<td>210 (95.8%)</td>
</tr>
<tr>
<td>10,000</td>
<td>223 (97.8%)</td>
</tr>
<tr>
<td>50,000</td>
<td>242 (99.5%)</td>
</tr>
<tr>
<td>100,000</td>
<td>261 (99.7%)</td>
</tr>
</tbody>
</table>
Experimental Results

Image segmentation

\[ N = 28,000, \ (200 \times 140), \ \eta = 0.01, \ R_S = 124, \ \text{comp. time 403 seconds.} \]
Experimental Results

Image segmentation

\[ N = 112,000, \ (400 \times 280), \ \eta = 0.01, \ Rs = 500, \ \text{comp. time 13 seconds}. \]
Conclusions

A sparse kernel model for spectral clustering is proposed.
The incomplete Cholesky factorization is used to build a low-rank approximation of the kernel matrix.
The out-of-sample extensions can be made sparse via a reduced set method.
The proposed method makes spectral clustering applicable to large-scale datasets.
References


