An Out-of-Sample Extension for Spectral Clustering based on Weighted Kernel PCA

Carlos Alzate Johan A. K. Suykens

ESAT-SCD-SISTA
Katholieke Universiteit Leuven
Leuven, Belgium

September 25, 2006

First Research Contact Day of the Computational Intelligence and Learning (CIL) doctoral school

Outline

Motivation

- It is not clear what we are optimizing when doing spectral clustering.
- Due to the lack of a clear optimization problem, the parameters selection is not straightforward.
- Clustering of new points should rely on approximation techniques.

Spectral Clustering

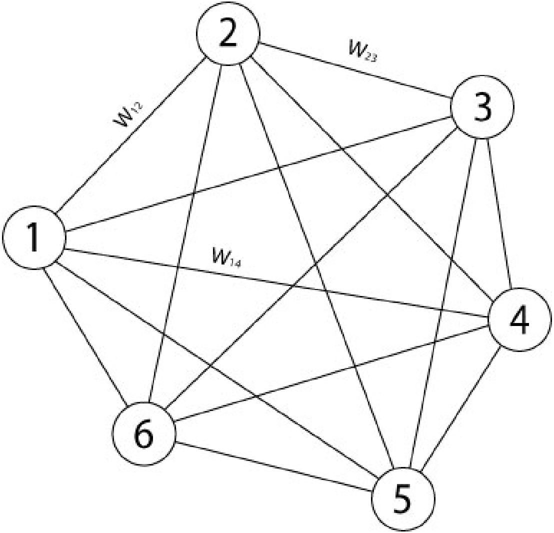
- Class of clustering algorithms that use the eigenvectors of an affinity matrix derived from the data.
- The data are represented as an undirected graph.
- The objective is to minimize the cost of cutting the graph into two disjoint sets $\mathcal{A}, \mathcal{B}$.

The Cut

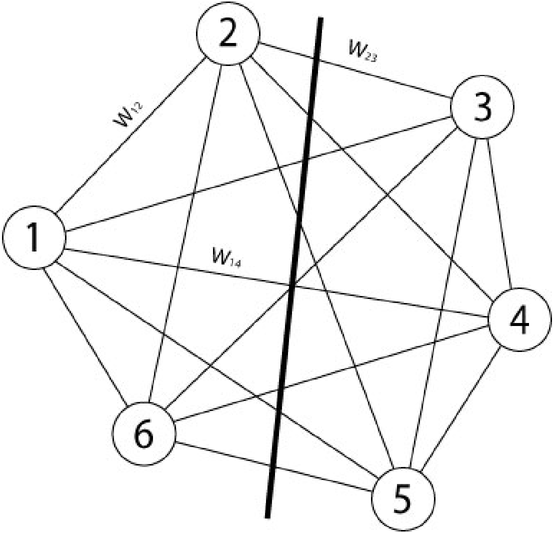
$$\mathrm{cut}(\mathcal{A}, \mathcal{B}) = \sum_{a \in \mathcal{A}, b \in \mathcal{B}} w(a, b)$$

where $w(a, b)$ is the weight between node a and b.

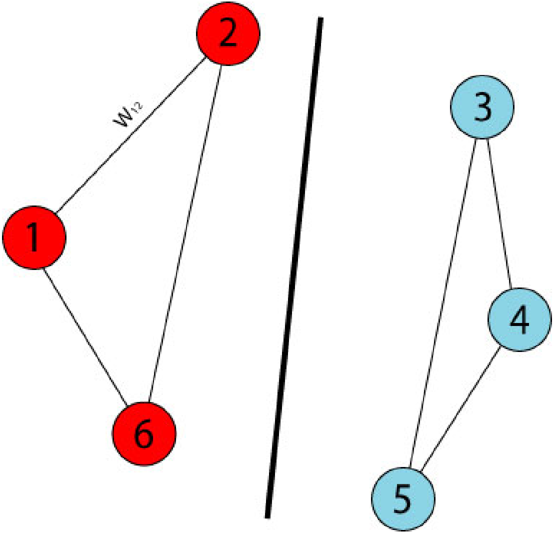
Spectral Clustering



Spectral Clustering



Spectral Clustering



The Mincut

$$\begin{aligned} \min_q J_{mincut} &= q^T(D-W)q \\ \text{such that} \quad & q \in \{-1,1\}^N \end{aligned}$$

D: degree matrix, W: affinity matrix, q: cluster membership indicator.

- NP-hard!
- Efficient solution by relaxing $q \rightarrow \tilde{q}^T\tilde{q} = 1$
- Bias for small sets.

The Mincut Relaxation

$$L\tilde{q} = \lambda\tilde{q}$$

L: graph *Laplacian*.
Solution: Fiedler vector.

Normalized Cut

$$\min_q J_{ncut} = \frac{q^T L q}{q^T D q}$$

$$\text{such that} \quad \begin{cases} q \in \{-b, 1\}^N \\ q^T D 1_N = 0 \end{cases}$$

- NP-complete!
- Efficient solution by relaxing $q \rightarrow \tilde{q}^T \tilde{q} = 1$
- Size of the clusters is taken into account.

Normalized Cut Relaxation

- Generalized eigenvalue problem:

$$L\tilde{q} = \lambda D\tilde{q}$$

Markov Random Walks

- Probabilistic interpretation.
- $P = D^{-1}W$.
- ij-th entry of $P \rightarrow$ probability of moving from node i to node j.

Solution

$$Pr = \xi r.$$

Solution is the eigenvector corresponding to the second largest eigenvalue.

- Equivalent to the normalized cut:

$$r = \tilde{q}, \lambda = 1 - \xi$$

Kernel Alignment

- Measure of similarity between a kernel and an objective function:

$$\begin{aligned} \max_q A(K, q) &= q^T \Omega q \\ \text{such that} \quad & q \in \{-1, 1\}^N \end{aligned}$$

where Ω is the kernel matrix.

Kernel Alignment Relaxation

After relaxing q the dual solution is an eigenvalue problem:

$$\Omega \tilde{q} = \lambda \tilde{q}$$

which corresponds to kernel PCA!

LS-SVM Approach to Kernel PCA

Clear primal optimization problem to which kernel PCA is the dual.
Underlying loss function is explicit $\rightarrow L_2$.

$$\begin{aligned} \max_{w,e} J_p(w,e) &= \gamma \frac{1}{2} e^T e - \frac{1}{2} w^T w \\ \text{such that} \quad & e = \Phi_c w \end{aligned}$$

Φ_c is the $N \times n_h$ feature matrix:

$$\Phi_c = \begin{bmatrix} \varphi(x_1)^T - \hat{\mu}_\varphi^T \\ \varphi(x_2)^T - \hat{\mu}_\varphi^T \\ \vdots \\ \varphi(x_N)^T - \hat{\mu}_\varphi^T \end{bmatrix}$$

Dual

Eigendecomposition of the centered kernel matrix Ω_c:

$$\Omega_c \alpha = \lambda \alpha$$

Weighted Kernel PCA

Introducing a weighting matrix V into the formulation:

$$\max_{w,e} J_p(w,e) = \gamma \frac{1}{2} e^T V e - \frac{1}{2} w^T w$$
$$\text{such that} \quad e = \Phi w$$

$\Phi = [\varphi(x_1)^T; \varphi(x_2)^T; \ldots; \varphi(x_N)^T]$, $V = V^T > 0$.

Dual

Non-symmetric eigenvalue problem:

$$V \Omega \alpha = \lambda \alpha$$

Equivalence

If $V = D^{-1}$ then weighted kernel PCA is equivalent to the random walks algorithm.

Relation with Spectral Clustering

Changing the weighting matrix V leads to different spectral clustering algorithms:

Method	Original Problem	V	Relaxed Solution
Alignment	$\Omega q = \lambda q$	I_N	$\alpha^{(1)}$
NCut	$Lq = \lambda D q$	D^{-1}	$\alpha^{(2)}$
Random walks	$D^{-1} W q = \lambda q$	D^{-1}	$\alpha^{(2)}$
NJW	$D^{-\frac{1}{2}} W D^{-\frac{1}{2}} q = \lambda q$	D^{-1}	$D^{\frac{1}{2}} \alpha^{(2)}$

Clustering New Points

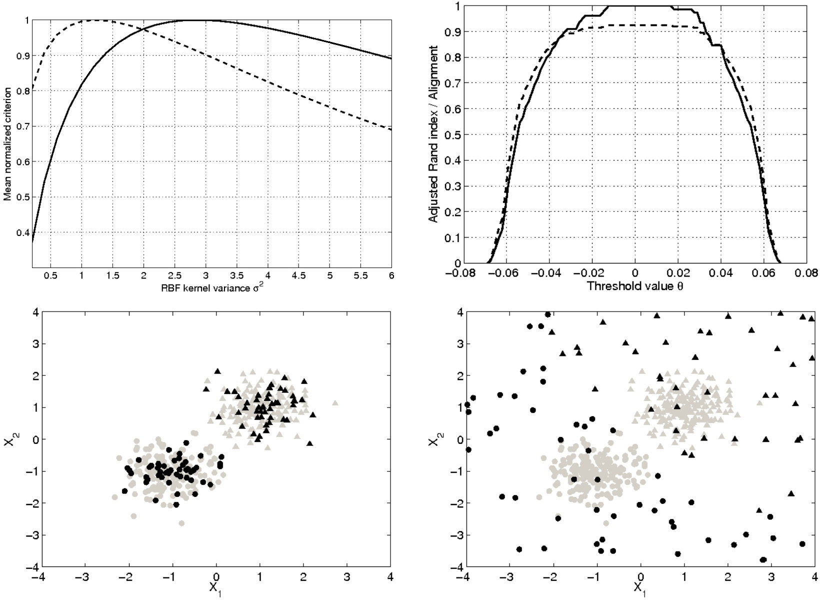
- No straightforward extensions for out-of-sample data points in the spectral clustering framework.
- Extensions can be done via approximation techniques such as Nyström *[Bengio et al., 2003]*.

Score Variables

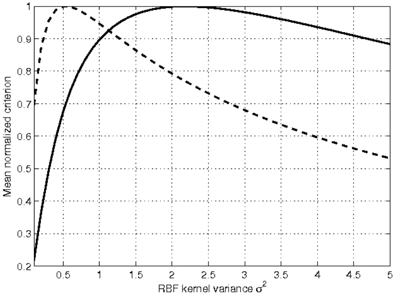
- No approximation needed! New points can be clustered using the projection onto the eigenvector solution:

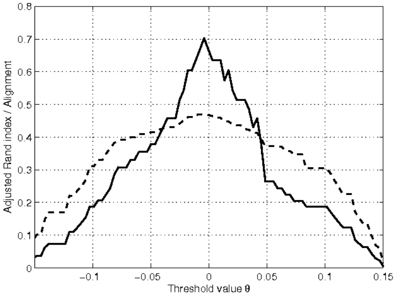
$$\begin{aligned} z(x_{new}) &= w^T \varphi(x_{new}) = \sum_{l=1}^{N} \alpha_l K(x_l, x_{new}). \\ q_{x_{new}} &= \operatorname{sign}(z(x_{new}) - \theta) \end{aligned}$$

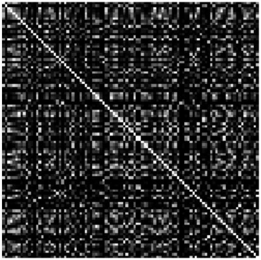
Toy Example

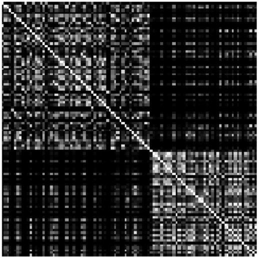

Mean normalized criterion
RBF kernel variance σ^2
Adjusted Rand index / Alignment
Threshold value θ
X_2
X_1
X_2
X_1

Iris Dataset

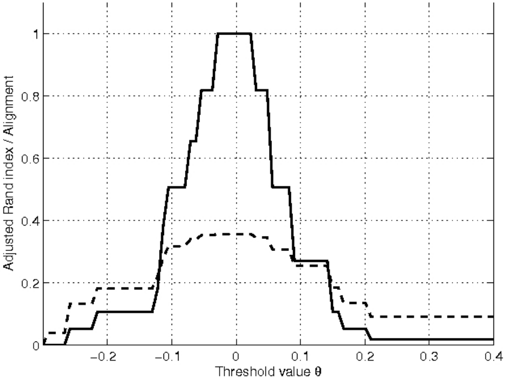








Golub Microarray



Method	Nyström RBF k.	Weighted KPCA RBF k.
Alignment	0.6	0.6
NCut	0.63	0.72
NJW	0.72	0.86

Conclusions

- Unifying view of spectral clustering based on the weighted kernel PCA formulation.
- Out-of-sample extension based on the primal-dual formulation insights.
- Model selection criterion using the variance of the projections on a validation set.

Future Work

- Extensions to K-way clustering (more than two clusters).
- Semi-supervised clustering (when some training points have labels).