Kernel k-means, Spectral Clustering and Normalized Cuts

- KDD’04

Author: Inderjit S. Dhillon, Yuqiang Guan, Brian Kulis
Outline:

1. Background
   a. k-means
   b. Weighted Kernel k-means
   c. Spectral Clustering

2. Contribution
   a. Connection
   b. Implication
k-means

1. Centroid-based clustering

2. Objective function

\[
\text{minimize} \quad \sum_{k=1}^{K} \sum_{C(i) = k} \| x_i - m_k \|^2
\]
k-means

3. Algorithm
   a. Fix k centroids $u_i$, minimize objective function w.r.t $X$.

   $$C(i) = \arg\min_{1 \leq k \leq K} \| x_i - m_k \|^2, \quad i = 1, \ldots, N$$

   b. Fix $X$, minimize objective function w.r.t $u$.

   $$m_k = \frac{\sum_{i:C(i) = k} x_i}{N_k}, \quad k = 1, \ldots, K.$$
**k-means**

4. Drawback

It can’t separate clusters that are non-linearly separable in input space.
Weighted Kernel k-means

1. Objective function

\[
\text{minimize} \quad \sum_{j=1}^{k} \sum_{a \in \pi_j} w(a) \| \phi(a) - m_j \|^2
\]

\[
m_j = \frac{\sum_{b \in \pi_j} w(b) \phi(b)}{\sum_{b \in \pi_j} w(b)}
\]
Weighted Kernel k-means

2. Transformation

\[ \| \phi(a) - m_j \|^2 = \left| \frac{\sum_{b \in \pi_j} w(b) \phi(b)}{\sum_{b \in \pi_j} w(b)} \right|^2 = \phi(a) \cdot \phi(a) - \frac{2 \sum_{b \in \pi_j} w(b) \phi(a) \cdot \phi(b)}{\sum_{b \in \pi_j} w(b)} + \frac{\sum_{b, c \in \pi_j} w(b) w(c) \phi(b) \cdot \phi(c)}{\left( \sum_{b \in \pi_j} w(b) \right)^2}. \]

\[ K = \phi(a) \cdot \phi(b) \]
Weighted Kernel k-means

3. Algorithm

\textsc{Weighted-Kernel-kmeans}(K, k, w, C_1, \ldots, C_k)

\textbf{Input:} \( K \): kernel matrix, \( k \): number of clusters, \( w \): weights for each point

\textbf{Output:} \( C_1, \ldots, C_k \): partitioning of the points

1. Initialize the \( k \) clusters: \( C_1^{(0)}, \ldots, C_k^{(0)} \).
2. Set \( t = 0 \).
3. For each point \( \mathbf{a} \), find its new cluster index as

\[ j^*(\mathbf{a}) = \arg\min_j \|\phi(\mathbf{a}) - \mathbf{m}_j\|^2, \text{ using (2)}. \]

4. Compute the updated clusters as

\[ C_j^{t+1} = \{\mathbf{a} : j^*(\mathbf{a}) = j\}. \]

5. If not converged, set \( t = t + 1 \) and go to Step 3; Otherwise, stop.
Spectral Clustering, Normalized Cut

$W$: similarity matrix.

$W(i,j) = 0$ when $i, j$ are not connected.

$$\text{links}(A, B) = \sum_{i \in A, j \in B} W(i, j)$$

$$\text{degree}(A) = \text{links}(A, V)$$

$$\text{linkratio}(A, B) = \frac{\text{links}(A, B)}{\text{degree}(A)}$$
Spectral Clustering, Normalized Cut

Objective function:

Minimize \( \text{kncuts}(\Gamma_X^K) = \frac{1}{K} \sum_{l=1}^{K} \text{linkratio}(\mathcal{V}_l, \mathcal{V} \setminus \mathcal{V}_l) \)

Let \( X \) be the \( n \times k \) indicator matrix. \( D = \text{Diag}(W1_N) \)

Minimize \( \frac{1}{K} \sum_{l=1}^{K} \frac{X_l^T(D - W)X_l}{X_l^TDX_l} \)
Spectral Clustering, Normalized Cut

Objective function:

\[
\text{maximize} \quad \varepsilon(X) = \frac{1}{K} \sum_{l=1}^{K} \frac{X_l^T W X_l}{X_l^T D X_l}
\]

subject to \quad X \in \{0, 1\}^{N \times K}

\quad X 1_K = 1_N.
Spectral Clustering, Normalized Cut

Objective function:

\[ Z_l = D^{1/2} X_l^T \]

\[
\text{Maximize} \frac{1}{K} \sum_{l=1}^{K} Z_l^T D^{-1/2} W D^{-1/2} Z_l
\]

\[ Z^T Z = I \]
Outline:

1. Background
   a. k-means
   b. Weighted Kernel k-means
   c. Spectral Clustering

2. Contribution
   a. Connection
   b. Implication
Connection

1. By choosing the weights in particular way, the weighted kernel k-means objective function is identical to the spectral clustering normalized cut.

\[ m_j = \frac{\sum_{b \in \pi_j} w(b)\phi(b)}{\sum_{b \in \pi_j} w(b)} \quad \rightarrow \quad m_j = \Phi_j \frac{W_j e}{s_j} \]

\[ \Phi = [\phi(a_1, \phi(a_2), \ldots, \phi(a_n)] \quad s_j = \sum_{a \in \pi_i} w(a) \]

W is diagonal matrix of all weights.
Connection

\[ D(\{\pi_j\}_{j=1}^k) = \sum_{j=1}^k d(\pi_j) \]

\[ d(\pi_j) = \sum_{\text{a} \in \pi_j} w(\text{a}) \| \phi(\text{a}) - m_j \|^2 \]

\[ = \sum_{\text{a} \in \pi_j} w(\text{a}) \| \phi(\text{a}) - \Phi_j \frac{W_j e e^T}{s_j} \|^2 \]

\[ = \| (\Phi_j - \Phi_j \frac{W_j e e^T}{s_j}) W_j^{1/2} \|_F \]

\[ = \| (\Phi_j W_j^{1/2} (I - \frac{W_j^{1/2} e e^T W_j^{1/2}}{s_j}) ) \|_F \]
Connection

1. \( \text{trace}(AA^T) = \text{trace}(A^TA) = \|A\|_F^2 \)

2. \( I - \frac{W_j^{1/2}ee^T W_j^{1/2}}{s_j} = P \) since \( s_j = e^T W_j e \) we know \( P^2 = P \)

\[
d(\pi_j) = \text{trace} \left( \Phi_j W_j^{1/2} \left| I - \frac{W_j^{1/2}ee^T W_j^{1/2}}{s_j} \right|^2 W_j^{1/2} \Phi_j^T \right)
\]

\[
= \text{trace} \left( \Phi_j W_j^{1/2} I - \frac{W_j^{1/2}ee^T W_j^{1/2}}{s_j} W_j^{1/2} \Phi_j^T \right)
\]

\[
= \text{trace}(W_j^{1/2} \Phi_j^T \Phi_j W_j^{1/2}) - \frac{e^T W_j \Phi_j^T \Phi_j W_j e}{\sqrt{s_j}} \frac{W_j e}{\sqrt{s_j}}.
\]
Connection

\[ D(\{\pi_j\}_{j=1}^k) = \text{trace}(W^{1/2} \Phi^T \Phi W^{1/2}) - \text{trace}(Y^T W^{1/2} \Phi^T \Phi W^{1/2} Y) \]

\[ Y = \begin{bmatrix} \frac{W_1^{1/2} e}{\sqrt{s_1}} \\ \frac{W_2^{1/2} e}{\sqrt{s_2}} \\ \ldots \\ \frac{W_k^{1/2} e}{\sqrt{s_k}} \end{bmatrix} . \]

First term is constant, \( K = \Phi^T \Phi \)
Connection

So, the problem:

Minimize $\mathcal{D}(\{\pi_j\}_{j=1}^k)$

becomes:

Maximize $\text{trace}(Y^T W^{1/2} K W^{1/2} Y)$

s.t. $Y^T Y = I$

$Y$ is an $n*k$ orthonormal matrix, so the optimal $Y$ is the top $k$ eigenvectors of $W^{1/2} K W^{1/2}$
Implication

1. Compute normalized cuts using weighted kernel k-means.
Implication

2. Techniques in kernel k-means could be used to compute normalized cuts.
   
a. Local search.
   b. Pruning procedure.

\[ \|x - m^n_j\| \geq \|x - m^0_j\| - \|m^n_j - m^0_j\| \]

Use right side to compute the lower bound as estimation. If estimation is smaller than the distance from \(x\) to its cluster center, compute the distance from \(x\) to \(m_j\).
Implication

Pruning procedure.
Implication

3. Compute kernel k-means using eigenvectors.
   a. run spectral clustering to get an initial partition.
   b. run kernel k-means on this partition.

<table>
<thead>
<tr>
<th></th>
<th>initial</th>
<th>final</th>
<th>NMI</th>
</tr>
</thead>
<tbody>
<tr>
<td>random</td>
<td>.0213</td>
<td>.0062</td>
<td>.666</td>
</tr>
<tr>
<td>spectral</td>
<td>.0081</td>
<td>.0059</td>
<td>.698</td>
</tr>
</tbody>
</table>
Summary

1. Connection between weighted kernel k-means and spectral clustering.

2. Implications of this connection.
Thanks