Isoperimetric Cut on a Directed Graph

Mo Chen, IE Dept. CUHK
Joint work with
Ming Liu, Jianzhuang Liu and Xiaoou Tang
Outline

- Brief review of previous graph based clustering algorithms (spectral clustering)
- Probabilistic interpretation of spectral clustering
- New method for constructing graphs
- New algorithm for partitioning directed graphs
- Summary
Glossary

- Undirected Graph \( G=(V,E) \)
- Affinity matrix \( W \), \( w_{ij} \) is the weight of edge \( e_{ij} \)
- Diagonal degree matrix \( D \), \( d_{ii} = \sum_j w_{ij} \)
- Transition matrix \( P=D^{-1}W \)
- Laplacian of undirected graph \( L=D-W \)
- \( \text{Vol}(\partial S) = \sum_{i,j} w_{ij}, i \in S, j \in \overline{S} \)
General Procedure of Spectral Clustering

1. Construct an undirected graph from vectors. The edge weights represent the similarities (usually Gaussian kernel is used)

\[ w_{ij} = \exp[-\frac{||x_i - x_j||^2}{(2\sigma^2)}] \]

2. Solve an eigen-decomposition problem of a certain matrix \( (L, W, P...) \)

3. Discretize the eigenvectors \( (k\text{-means}...) \)
Normalized Cut

Criterion: Normalized Cut (Shi & Malik, ’97)
- Normalize the association between groups by volume

\[
NCut(G) = \min_S \frac{\text{Vol}(\partial S)}{\text{Vol}(S)} + \frac{\text{Vol}(\partial \bar{S})}{\text{Vol}(\bar{S})}
\]

\(\text{Vol}(S)\): The total weight of edges originating from group \(S\).
Normalized Cut

Criterion: Normalized Cut (Shi & Malik, ’97)

- Normalize the association between groups by volume

\[ NCut(G) = \min_S \frac{Vol(\partial S)}{Vol(S)} + \frac{Vol(\partial S)}{Vol(S)} \]

- Vol(S): The total weight of edges originating from group S.

Why use this criterion?
- Works well in practice
- Has good theoretical property: consistency
Solution of NCut

- Solve a relaxed convex optimization problem

\[ \arg\min_y y^T L y \]

\[ s.t. \quad y^T D y = 1 \]

- Apply eigen-decomposition

\[ L y = \lambda D y \]
Graph Isoperimetric Ratio

- The eigen-decomposition solution is also an upper bound of the graph Cheeger isoperimetric ratio

\[ h_G = \min_S \frac{\text{Vol}(\partial S)}{\text{Vol}(S)} \]

- Therefore, spectral clustering is a way to optimize isoperimetric ratio
Problems of Spectral Clustering

► Three steps of the algorithms are uncorrelated
► How graph construction step affects clustering results is not clear
► Bad parameter setting leads to bad results
► Not work well on multiscale data sets even parameter is carefully tuned
► There is no principal way for the parameter selection
A Failed Case
An Alternative Probabilistic View
Kernel Density Estimation

The kernel density estimator is given by

$$f_h(x) = \frac{1}{nh} \sum_{j=1}^{n} K\left(\frac{x - x_j}{h}\right)$$

$h$ is the bandwidth (smoothing parameter) depending only on the sample size.

Gaussian Kernel is to use

$$K\left(\frac{x - y}{h}\right) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\|x - y\|^2}{2h^2}\right)$$
Bayes Error

Bayes error rate for two-class classification problem is

\[ P(error) = \int_{T_1} p(x|c_2)p(c_2)dx + \int_{T_2} p(x|c_1)p(c_1)dx \]

Optimal classification boundary if the conditional distribution \( p(x|c_i) \) for each class \( c_i \) is known
Nonparametric density modeling

► Model conditional density by KDE

\[ p(x|c_i) = \frac{1}{|S_i|h\sqrt{2\pi}} \sum_{j, x_j \in S_i} \exp\left( -\frac{||x - x_j||^2}{2h^2} \right) \]

► Approximate integral by summation

\[ \int_{T_i} p(x|c_i) dx \approx \frac{\sum_{j, x_j \in S_i} p(x_j|c_i)}{\sum_{j=1}^{n} p(x_j|c_i)} \]
Assuming equal priors \( p(c_1) = p(c_2) \) we have

\[
P(error) = \int_{T_1} p(x|c_2)p(c_2)dx + \int_{T_2} p(x|c_1)p(c_1)dx
\]

\[
\approx \frac{\sum_{j,x_j \in S_1} p(x_j|c_1)}{\sum_{j=1}^{n} p(x_j|c_1)} + \frac{\sum_{j,x_j \in S_2} p(x_j|c_2)}{\sum_{j=1}^{n} p(x_j|c_2)} = NCut(G)
\]
Local Density!
Directed Graph Construction

- Use variable bandwidth kernel

\[ f_b(x) = \sum_{j=1}^{n} \frac{1}{nh_j} K \left( \frac{x - x_j}{h_j} \right) \]

- Equivalent to construct a directed graph (using Gaussian kernel) with edge weights

\[ w_{ij} = \frac{1}{h_i} \exp \left( -\frac{\| x_i - x_j \|^2}{2h_i^2} \right) \]

- Model selection: Leave-one-out cross validation likelihood
Random Walks

Transition probability: \( p_{ij} = \frac{w_{ij}}{d_i} \)

Define volumes of a directed graph by random walks

\[
\text{Vol}(\partial S) = \sum_{i \in S, j \in \overline{S}} \pi_i p_{ij}
\]

\[
\text{Vol}(S) = \sum_{i \in S} \pi_i
\]
Random Walk Isoperimetric Ratio

- Isoperimetric ratio of a directed graph

\[ h_R = \inf_{S} \frac{\text{Vol}(\partial S)}{\text{Vol}(S)} = \min_{x} \frac{x^T \Pi (I - P)x}{x^T \Pi 1} \]

- Undirected graph is a special case of this
- Relaxed problem solved by linear system

\[ 2(I - P)x = 1 \]

- Singular :(
Isocut

- Designate one vertex to be included in $S$, then solve

\[ L_0 x_0 = 1 \]

- We are done for bi-partitioning problem now

- Multi-class clustering result: recursively cut the subgraphs with the smallest isoperimetric constant
Random walk expected first hitting time

\[ h(i|i) = 0, \]
\[ h(j|i) = 1 + \sum_{k=1}^{n} p_{ik} h(j|k), \quad i \neq j. \]

Equivalent to

\[ h_0 = 1 + P_0 h_0 \]
\[ L_0 x_0 = 1 \]
Useful Facts

Fact 1: if the ground vertex is selected such that for any other vertex there exists a path from any other vertex to the ground then the linear system is well posed.

Fact 2: the subgraph containing the ground vertex is connected, regardless of how a threshold (i.e., cut) is chosen.
Evaluation

► Image data sets
  ▪ handwritten digits
  ▪ human faces
  ▪ natural scenes
  ▪ satellite images

► Other Benchmark data sets
  ▪ UCI
  ▪ ...
# Experimental Results

Table 1: NMI comparison results.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Kmeans</th>
<th>NCut</th>
<th>NJW</th>
<th>StNCut</th>
<th>StNJW</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iris</td>
<td>0.7582</td>
<td>0.7571</td>
<td>0.7661</td>
<td>0.6524</td>
<td>0.7857</td>
<td>0.8449</td>
</tr>
<tr>
<td>Wine</td>
<td>0.4288</td>
<td>0.4624</td>
<td>0.4351</td>
<td>0.3665</td>
<td>0.4199</td>
<td>0.4496</td>
</tr>
<tr>
<td>WDBC</td>
<td>0.4672</td>
<td>0.5754</td>
<td>0.5358</td>
<td>0.4679</td>
<td>0.4845</td>
<td>0.5868</td>
</tr>
<tr>
<td>Satimage</td>
<td>0.6138</td>
<td>0.6749</td>
<td>0.6373</td>
<td>0.6336</td>
<td>0.6307</td>
<td>0.6932</td>
</tr>
<tr>
<td>Segment</td>
<td>0.6124</td>
<td>0.6465</td>
<td>0.6629</td>
<td>0.5852</td>
<td>0.6801</td>
<td>0.7440</td>
</tr>
<tr>
<td>UMist-all</td>
<td>0.6726</td>
<td>0.6157</td>
<td>0.8009</td>
<td>0.5364</td>
<td>0.6512</td>
<td>0.8785</td>
</tr>
<tr>
<td>UMist-10</td>
<td>0.6161</td>
<td>0.5769</td>
<td>0.8214</td>
<td>0.4918</td>
<td>0.5850</td>
<td>0.8634</td>
</tr>
<tr>
<td>UMist-5</td>
<td>0.7065</td>
<td>0.8903</td>
<td>0.8655</td>
<td>0.6384</td>
<td>0.6371</td>
<td>1</td>
</tr>
<tr>
<td>USPS-all</td>
<td>0.4038</td>
<td>0.4517</td>
<td>0.5180</td>
<td>0.1894</td>
<td>0.3606</td>
<td>0.6880</td>
</tr>
<tr>
<td>USPS-5</td>
<td>0.4469</td>
<td>0.5789</td>
<td>0.4247</td>
<td>0.2536</td>
<td>0.3197</td>
<td>0.6910</td>
</tr>
<tr>
<td>Scene</td>
<td>0.3951</td>
<td>0.4100</td>
<td>0.4471</td>
<td>0.3605</td>
<td>0.4204</td>
<td>0.4695</td>
</tr>
</tbody>
</table>
## Experimental Results

Table 1: Error comparison results.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Kmeans</th>
<th>NCut</th>
<th>NJW</th>
<th>StNCut</th>
<th>StNJW</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iris</td>
<td>0.1067</td>
<td>0.0933</td>
<td>0.1000</td>
<td>0.4867</td>
<td>0.0933</td>
<td>0.0533</td>
</tr>
<tr>
<td>Wine</td>
<td>0.2978</td>
<td>0.2697</td>
<td>0.2921</td>
<td>0.2921</td>
<td>0.2865</td>
<td>0.2472</td>
</tr>
<tr>
<td>WDBC</td>
<td>0.1459</td>
<td>0.0879</td>
<td>0.109</td>
<td>0.1388</td>
<td>0.1248</td>
<td>0.0796</td>
</tr>
<tr>
<td>Satimage</td>
<td>0.3310</td>
<td>0.2544</td>
<td>0.2457</td>
<td>0.2810</td>
<td>0.2737</td>
<td>0.2197</td>
</tr>
<tr>
<td>Segment</td>
<td>0.3342</td>
<td>0.4004</td>
<td><strong>0.2740</strong></td>
<td>0.5165</td>
<td>0.3407</td>
<td>0.2922</td>
</tr>
<tr>
<td>UMist-all</td>
<td>0.5339</td>
<td>0.5791</td>
<td>0.3948</td>
<td>0.6348</td>
<td>0.5739</td>
<td><strong>0.2661</strong></td>
</tr>
<tr>
<td>UMist-10</td>
<td>0.5509</td>
<td>0.5208</td>
<td>0.3057</td>
<td>0.5849</td>
<td>0.5547</td>
<td><strong>0.2604</strong></td>
</tr>
<tr>
<td>UMist-5</td>
<td>0.2214</td>
<td>0.0857</td>
<td>0.1214</td>
<td>0.3786</td>
<td>0.3071</td>
<td>0</td>
</tr>
<tr>
<td>USPS-all</td>
<td>0.6008</td>
<td>0.6404</td>
<td>0.4882</td>
<td>0.8396</td>
<td>0.6388</td>
<td><strong>0.3398</strong></td>
</tr>
<tr>
<td>USPS-5</td>
<td>0.3468</td>
<td>0.4140</td>
<td>0.4256</td>
<td>0.6224</td>
<td>0.4572</td>
<td><strong>0.2232</strong></td>
</tr>
<tr>
<td>Scene</td>
<td>0.5056</td>
<td>0.4835</td>
<td>0.4014</td>
<td>0.5443</td>
<td>0.4725</td>
<td><strong>0.3857</strong></td>
</tr>
</tbody>
</table>
Experimental Results

The number of samples vs. Running times (second). The graph shows the running times for NCut and IsoCut algorithms as a function of the number of samples. The running times increase with the number of samples for both algorithms, but NCut shows a steeper increase compared to IsoCut.
Future Work

► Multiclass isocut formulation

► Other density models (mixture model, etc)

► Large scale clustering (sampling + local density estimation)

► Theoretical result: consistency
Thanks.