A Spectral Clustering Approach to Optimally Combining Numerical Vectors with a Modular Network

Motoki Shiga, Ichigaku Takigawa, Hiroshi Mamitsuka

Bioinformatics Center, ICR, Kyoto University, Japan

KDD 2007, San Jose, California, USA, August 12-15 2007
# Table of Contents

1. Motivation
   Clustering for heterogeneous data
   (numerical + network)

2. Proposed method
   Spectral clustering (numerical vectors + a network)

3. Experiments
   Synthetic data and real data

4. Summary
Heterogeneous Data Clustering

Heterogeneous data: various information related to an interest

**Ex.** Gene analysis: gene expression, metabolic pathway, ..., etc.

Web page analysis: word frequency, hyperlink, ..., etc.

To improve clustering accuracy, combine numerical vectors + network

Related work: semi-supervised clustering

- Local property
  Neighborhood relation
  - must-link edge, cannot-link edge

- Hard constraint (K. Wagstaff and C. Cardie, 2000.)
- Soft constraint (S. Basu etc., 2004.)
  - Probabilistic model (Hidden Markov random field)

Proposed method

- Global property (network modularity)
- Soft constraint
  - Spectral clustering
Table of Contents

1. Motivation
   Clustering for heterogeneous data
   (numerical + network)

2. Proposed method
   Spectral clustering (numerical vectors + a network)

3. Experiments
   Synthetic data and real data

4. Summary
1. Compute affinity(dissimilarity) matrix $M$ from data

2. To optimize cost

$$J(Z) = \text{tr}\{Z^T M Z\} \text{ subject to } Z^T Z = I$$

*Trace optimization*

where $Z(i,k)$ is 1 when node $i$ belong to cluster $k$, otherwise 0, compute eigen-values and -vectors of matrix $M$ by relaxing $Z(i,k)$ to a real value

Each node is by one or more computed eigenvectors

3. Assign a cluster label to each node (by k-means)
Cost combining numerical vectors with a network

\[
J = \text{tr}\{Z^T M Z\} = (1 - \omega)J_{\text{num}}(Z) + \omega J_{\text{net}}(Z)
\]

Cost of numerical vector

\[
J_{\text{num}}(Z) = \frac{1}{2} - \text{tr}\left(\frac{Z^T (2N)^{-1} Y Z}{Z^T Z}\right)
\]

network

cosine dissimilarity

What cost?

N : #nodes,
Y : inner product of normalized numerical vectors

To define a cost of a network, use a property of complex networks
Complex Networks

Ex. Gene networks, WWW, Social networks, ..., etc.

Property
- Small world phenomena
- Power law
- Hierarchical structure
- Network modularity

Network Modularity

= density of intra-cluster edges

\[
Q(Z) = \sum_{k=1}^{K} \left\{ \frac{L(Z_k, Z_k)}{L} - \left( \frac{L(Z_k, Z)}{L} \right)^2 \right\}
\]

normalize by cluster size

# intra-edges

# total edges

Z : set of whole nodes

Z_k : set of nodes in cluster k

L(A,B) : # edges between A and B

Cost Combining Numerical Vectors with a Network

\[ J = \text{tr}\{Z^T M Z\} \]
\[ = (1 - \omega) J_{\text{num}}(Z) + \omega J_{\text{net}}(Z) \]

**Cost of numerical vector**

\[ J_{\text{num}}(Z) = \frac{1}{2} - \text{tr}\left( \frac{Z^T (2N)^{-1} Y Z}{Z^T Z} \right) \]

**Cosine dissimilarity**

**Network cost**

\[ J_{\text{net}}(Z) = -\text{tr}\left( \frac{Z^T N \left( \frac{1}{L^2} D - \frac{1}{L} W \right) Z}{Z^T Z} \right) \]

**Normalized modularity (Negative)**

\[ \tilde{Z} = \frac{Z}{\sqrt{Z^T Z}} \]

\[ = \text{tr}\{\tilde{Z}^T M \omega \tilde{Z}\} \]
Our Proposed Spectral Clustering

for $\omega = 0...1$

1. Compute matrix $M_\omega = \frac{\omega N}{L^2} D - \frac{\omega N}{L} W - \frac{1 - \omega}{2N} Y$

2. To optimize cost $J(Z) = \text{tr}\{Z^T M_\omega Z\}$ subj to $Z^T Z = I$, compute eigen-values and -vectors of matrix $M_\omega$ by relaxing elements of $Z$ to a real value.

Each node is represented by K-1 eigen-vectors.

3. Assign a cluster label to each node by k-means. (k-means outputs $\text{Cost}_{\text{spectral}}$ in spectral space.)

end

- Optimize weight $\omega$

$\omega^* \leftarrow \text{argmin}_{0 \leq \omega \leq 1} \text{Cost}_{\text{spectral}}$

$\text{Cost}_{\text{spectral}}$ is sum of dissimilarity (cluster center <-> data)

Eigen-vector $e_1$
Table of Contents

1. Motivation
   Clustering for heterogeneous data
   (numerical + network)

2. Proposed method
   Spectral clustering (numerical vectors + a network)

3. Experiments
   Synthetic data and real data

4. Summary
Synthetic Data

Numerical vectors (von Mises-Fisher distribution)

\[ \theta = 1 \]

Network (Random graph) #nodes = 400, #edges = 1600

Modularity = 0.375 0.450 0.525
Results for Synthetic Data

Modularity $= 0.375$

**Numerical vectors**

Best NMI (Normalized Mutual Information) is in $0 < \omega < 1$.

Can be optimized using Cost$_{spectral}$.
Results for Synthetic Data

Modularity = 0.375

- Best NMI (Normalized Mutual Information) is in $0 < \omega < 1$
- Can be optimized using $\text{Cost}_{\text{spectral}}$

Cost$_{\text{spectral}}$

NMI

$\theta = 1$
$\theta = 5$
$\theta = 50$

Numerical vectors only
(k-means)

Network only
(maximum modularity)
**Synthetic Data (Numerical Vector) + Real Data (Gene Network)**

**True cluster**  
(#clusters = 10)

**Resultant cluster**  
($\omega=0.5, \theta=10$)

Gene network by KEGG metabolic pathway

- Best NMI is in $0 < \omega < 1$
- Can be optimized using

- Best NMI is in $0 < \omega < 1$
- Can be optimized using
Summary

• **New spectral clustering method proposed**
  combining numerical vectors with a network
  • **Global network property** (normalized network modularity)
  • Clustering can be optimized by the weight

• **Performance confirmed experimentally**
  • Better than numerical vectors only and a network only
  • **Optimizing the weight** with synthetic dataset and semi-real dataset
Thank you for your attention!
Spectral Representation of $M_\omega$

(concentration $\theta = 5$, Modularity = 0.375)

$\omega = 0$  \hspace{1cm} $\omega = 0.3$  \hspace{1cm} $\omega = 1$

Cost $J_\omega = 0.0932$  \hspace{1cm} 0.0538  \hspace{1cm} 0.0809

Select $\omega$ by minimizing Cost$_{\text{spectral}}$

(clusters are divided most separately)
Result for Real Genomic Data

- Numerical vectors: Hughes’ expression data (Hughes, et al., cell, 2000)
- Gene network: Constructed using KEGG metabolic pathways (M. Kanehisa, etc. NAR, 2006)
Evaluation Measure

**Normalized Mutual Information (NMI)**

between estimated cluster and the standard cluster

\[
NMI = \frac{H(C) + H(G) - H(C, G)}{\sqrt{H(C)} \sqrt{H(G)}}
\]

\(H(C)\) : Entropy of probability variable \(C\),
\(C\) : Estimated clusters,
\(G\) : Standard clusters

The more similar clusters \(C\) and \(G\) are, the larger the NMI.
## Web Page Clustering

<table>
<thead>
<tr>
<th>Word</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$</td>
<td>$n_{(A,1)}$</td>
</tr>
<tr>
<td>$B$</td>
<td>$n_{(B,1)}$</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

To improve accuracy, combine heterogenous data

### Numerical Vector

- Frequency of word $A$
- Frequency of word $B$

### Network

1. 2. 3. 4. 5. 6. 7.
Spectral Clustering for Graph Partitioning

- **Ratio cut**
  \[
  \sum_k \frac{L(Z_k, Z \setminus Z_k)}{|Z_k|}
  \]
  \[\text{Subject to } \tilde{Z}^T \tilde{Z} = I_K\]
  \[
  \text{tr}(\tilde{Z}^T (D_d - W) \tilde{Z})
  \]

- **Normalized cut**
  \[
  \sum_k \frac{L(Z_k, Z \setminus Z_k)}{L(Z_k, Z)}
  \]
  \[\text{Subject to } \tilde{Z}^T D_d \tilde{Z} = I_K\]
  \[
  \text{tr}(\tilde{Z}^T (D_d - W) \tilde{Z})
  \]