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

Abstract

Clustering, a major research subject in data mining, has been successfully applied to a wide variety of areas in the real world. In this paper, we address the issue of clustering numerical vectors with a network. This is a general setting which can be found in a lot of applications and basically equivalent to constrained clustering by Wagstaff and Cardie [Wagstaff2000] and semi-supervised clustering by Basu et al. [Basu2004], but our focus is more on the optimal combination of two heterogeneous data sources, numerical vectors and a network.

An application of this setting is web pages which can be numerically vectorized by their contents, e.g. term frequencies, and which are hyperlinked to each other, showing a network. Another typical application is genes whose behavior can be numerically measured and a gene network can be given from another data source.

We first define a new graph clustering measure which we call normalized network modularity, by balancing the cluster size of the original modularity. We then propose a new clustering method which integrates the cost of clustering numerical vectors with the cost of maximizing the normalized network modularity into a spectral relaxation problem. Our learning algorithm is based on spectral clustering which makes our issue an eigenvalue problem and uses k-means for final cluster assignments. A significant advantage of our method is that we can optimize the weight parameter for balancing the two costs from the given data by choosing the minimum total cost.

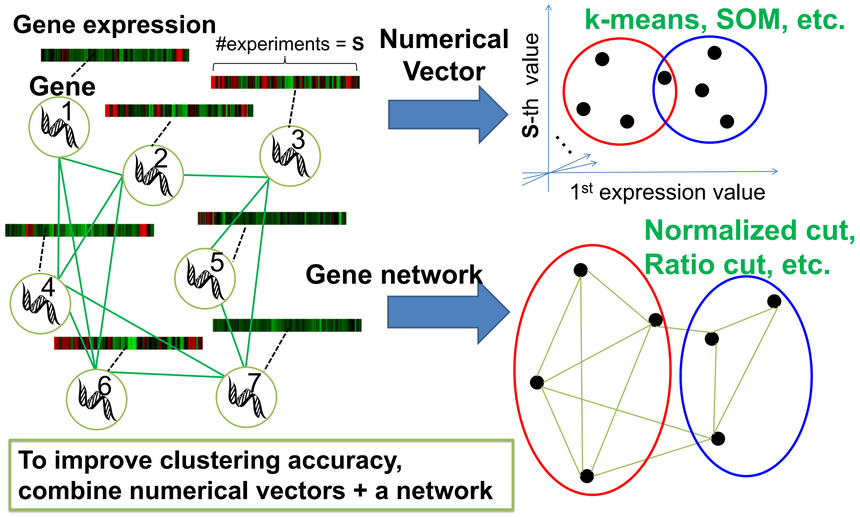
We evaluated the performance of our proposed method using a variety of datasets including synthetic data as well as real-world data from molecular biology. Experimental results showed that our method is effective enough to have good results for clustering by numerical vectors and a network.

1. 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 + a network

2. Cost Combining Numerical Vectors with a Network

$$\mathrm{J}(\mathbf{Z}) = (1-\omega)\,\mathrm{J}_{\mathrm{num}}(\mathbf{Z}) + \omega\,\mathrm{J}_{\mathrm{net}}(\mathbf{Z})$$

cosine dissimilarity

$$\mathrm{J}_{\mathrm{num}}(\mathbf{Z}) = \frac{1}{2} - \mathrm{tr}\left(\frac{\mathbf{Z}^{\mathrm{T}}(2N)^{-1}\mathbf{Y}\mathbf{Z}}{\mathbf{Z}^{\mathrm{T}}\mathbf{Z}}\right)$$

(Negative) Normalized modularity

$$\mathrm{J}_{\mathrm{net}}(\mathbf{Z}) = -\mathrm{tr}\left(\frac{\mathbf{Z}^{\mathrm{T}}N\left(\frac{1}{L^2}-\mathbf{D}\frac{1}{L}\mathbf{W}\right)\mathbf{Z}}{\mathbf{Z}^{\mathrm{T}}\mathbf{Z}}\right)$$

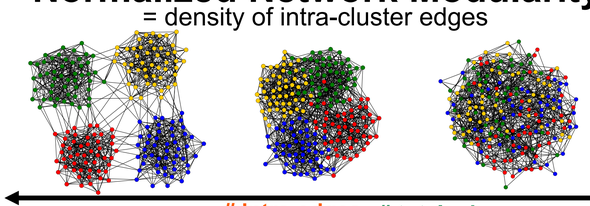
$$\widetilde{\mathbf{Z}} = \mathbf{Z}\Big/\sqrt{\mathbf{Z}^{\mathrm{T}}\mathbf{Z}}$$

$$= \mathrm{tr}\left\{\widetilde{\mathbf{Z}}^{\mathrm{T}}\left(\frac{\omega N}{L^2}\mathbf{D} - \frac{\omega N}{L}\mathbf{W} - \frac{1-\omega}{2N}\mathbf{Y}\right)\mathbf{Z}\right\}$$

$\mathbf{M}_\omega$

Normalized Network Modularity

= density of intra-cluster edges



High ← → Low

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

(# intra-edges; # total edges)

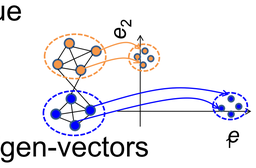
normalize by cluster size

Z : set of whole nodes, Z_k : set of nodes in cluster k, L(A,B) : #edges between A and B

3.Our Proposed Spectral Clustering

for ω = 0…1

1. Compute matrix $\mathbf{M}_\omega = \frac{\omega N}{L^2}\mathbf{D} - \frac{\omega N}{L}\mathbf{W} - \frac{1-\omega}{2N}\mathbf{Y}$
2. To optimize cost $J(\mathbf{Z}) = \mathrm{tr}\{\mathbf{Z}^{\mathrm{T}}\mathbf{M}_\omega\mathbf{Z}\}$ subjet to $\mathbf{Z}^{\mathrm{T}}\mathbf{Z}=\mathbf{I}$, compute eigen-values and -vectors of matrix $\mathbf{M}_\omega$ relaxing elements of $\mathbf{Z}$ to a real value



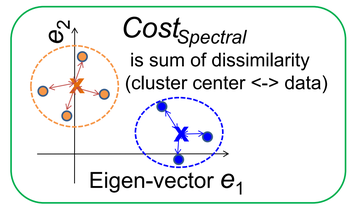
Each node is represented by K-1 eigen-vectors

3. Assign a cluster label to each node by k-means. (k-means outputs $Cost_{Spectral}$ in spectral space.)

end

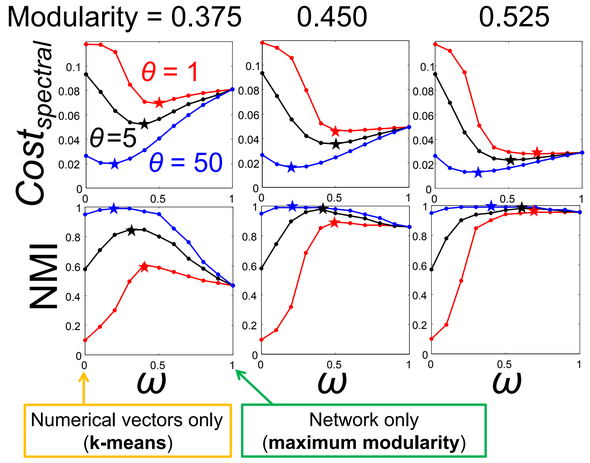
· Optimize weight ω

$$\omega^* \leftarrow \arg\min_{0\le\omega\le 1} Cost_{\mathrm{Spectral}}$$

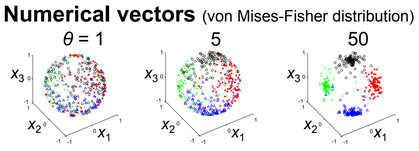


4. Experiments

4-1. Synthetic Data

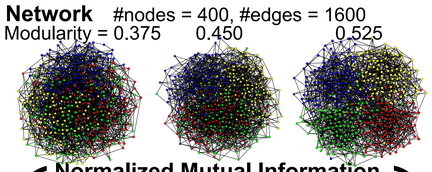


Numerical vectors (von Mises-Fisher distribution)



Network #nodes = 400, #edges = 1600

Modularity = 0.375 0.450 0.525



Normalized Mutual Information (NMI)

$$\mathrm{NMI} \leftarrow \frac{\mathrm{H}(C)+\mathrm{H}(G)-\mathrm{H}(C,G)}{\sqrt{\mathrm{H}(C)}\sqrt{\mathrm{H}(G)}}$$

$H(\mathbf{C})$: Entropy of probability variable **C**,
C : Resultant cluster, **G** : KEGG metabolic function

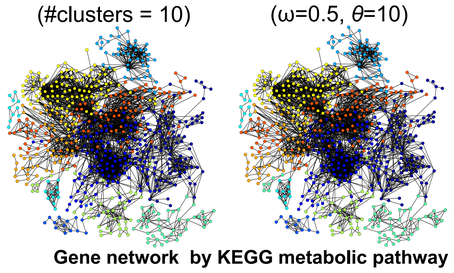
The more **similar** clusters **C** and **G** are, **the larger the NMI**.

- **Best NMI is in 0 < ω < 1**
- **Can optimize weight ω using $Cost_{\mathrm{spectral}}$**

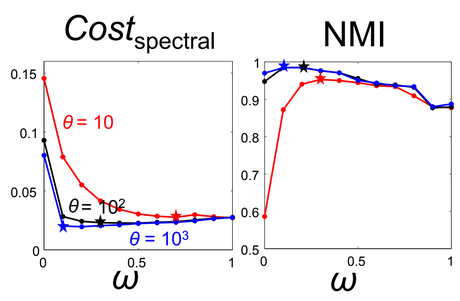
4-2. Synthetic Data (numerical vector) + Real Data (gene network)

True cluster (#clusters = 10)

Resultant cluster (ω=0.5, θ=10)

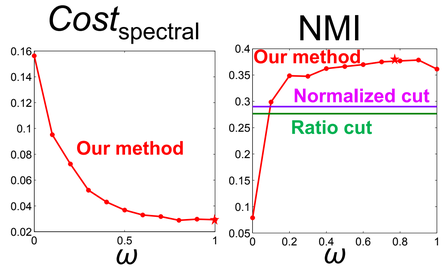


Gene network by KEGG metabolic pathway



4-3. Real Genomic Data

- Numerical vectors : Hughes' expression data *(Hughes, et al., cell, 2000)*
- Gene network : KEGG metabolic pathway *(M. Kanehisa, etc. NAR, 2006)*



5. Reference

[1] S. Basu, M. Bilenko, and R. J. Mooney. Aprobabilistic framework for semi-supervised clustering. In *KDD*, p. 59–68, August 2004.

[2] T. R. Hughes et al. Functional discovery via a compendium of expression profiles *Cell*, 102(1):109–126, 2000.

[3] M. Kanehisa et al. From genomics to chemical genomics: new developments in KEGG. *NAR*, 34:D354–357, 2006.

[4] E. Ravasz et al. Hierarchical organization of modularity in metabolic networks. *Science*, 297(5589):1551–1555, 2002.

[5] M. Shiga, I. Takigawa and H. Mamitsuka. Annotating gene function by combining expression data with a modular gene network. In *ISMB/ECCB 2007*.

[6] K. Wagstaff and C. Cardie. Clustering with instance-level constraints. In *ICML, p. 1103–1110,* 2000.

KDD 2007, San Jose, California, USA, August 12-15 2007