Spectral clustering techniques for biological data

17 septembre 2014

Plan

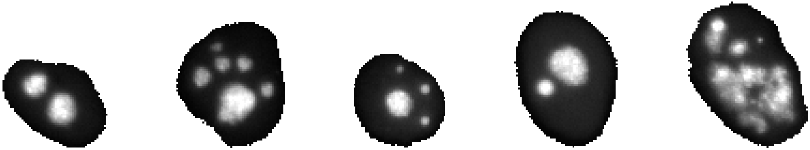
1. Project presentation
2. Spectral clustering
3. Results on synthetic data / biological data

Plan

1. **Project presentation**
2. Spectral clustering
3. Results on synthetic data / biological data

→ strong correlation between structure of the nucleolus of a cell and potential diseases of this cell

→ biologist have generated a database by annihilating some specific genes of the cells (silencers) and they have visually observed different conformations of the nucleolus



Well A12 Well A12 Well F12 Well F12 Well E02

1 well of cells = 1 silencer

Project Presentation

Objective :

- clustering the cells based on the conformation of their nucleolus
- maximize the number of cluster

Hypothesis :

- the cell of the same well should belong to the same cluster

Project Presentation

After an image analysis processing, each cell is represented by a 15-dimensional characteristics' vector $x_i \in \mathbb{R}^{15}$
Example : elliptic regularity, number of connected component, luminous intensity

Project Presentation

Presence of the noice :

- some cells of a well could not have reacted to the silencer
- 2D representation of a 3D cell

Plan

Graph clustering

data points : $x_1, ..., x_n \in \mathbb{R}^p$
similarity matrix : $W = (w_{ij})_{i,j=1..n} = w(x_i, x_j)$
similarity graph $G = (V, E)$
V : vertices (data points)   E : edges with weight w_{ij}

Problem of clustering $\leftrightarrow$ Partition the graph so that edges within a group have large weights and edges across groups have small weights.

Construction of the connectivity matrix

For each vertices, selection of the m-nearest neighbors $\rightarrow$ $C(i,j) = 1$ if j is one of the m-nearest neighbors of i and 0 otherwise.

C is not symmetric :

- $C_{norm} = max(C, C') \rightarrow C(i,j) = 1$ if i is one of the m-nearest neighbors of j OR if j is one of the m-nearest neighbors of i : each vertice has at least m neighbors (normal graph)
- $C_{mut} = min(C, C') \rightarrow C(i,j) = 1$ if i is one of the m-nearest neighbors of j AND if j is one of the m-nearest neighbors of i : each point has at most m neighbors (mutual graph)

$\rightarrow$ Connectivity matrix C_{norm} or C_{mut} : sparse matrix

Construction of the similarity matrix

If i and j are connected $w_{ij} = e^{-\frac{||x_i - x_j||^2}{2\sigma^2}} \rightarrow \sigma$ controls the size of the neighborhood

How to choose σ :

- human-specified parameter
- local scaling ([1] Zelnik-Manor, 2005) : one value of σ for each point. Ex : $\sigma_i = \max_j(||x_i - x_j||)$ for j in the neighborhood of i

Definitions

Degree of a vertice i : $d_i = \sum_{j=1}^{n} w_{ij}$
Degree diagonal-matrix with coefficients d_i : D
Laplacian matrix :

$$L = D - W$$

Normalized Laplacian matrix :

- $L_{rw} = D^{-1}L$
- $L_{sym} = D^{-1/2}LD^{-1/2}$

Numbers of connected components and spectrum of L_{rw}

The multiplicity k of the eigenvalue 0 of L_{rw} equals the number of connected components $A_1, ..., A_k$ in the graph. The eigenspace of the eigenvalue 0 is spanned by the indicator vector $\mathbb{1}_{A_1}, ..., \mathbb{1}_{A_k}$

Partitioning a graph

For two subsets A, B of V : $W(A,B) = \sum_{i \in A, j \in B} w_{ij}$
Two ways for mesuring the "size" of a subset A :

- $|A|$: number of vertices in A
- $vol(A) = \sum_{i \in A} d_i$

Two criteria to partitioning a graph :

$$cut(A_1, ..., A_k) = \frac{1}{2} \sum_{i=1}^{k} W(A_i, \overline{A_i})$$

$$Ncut(A_1, ..., A_k) = \frac{1}{2} \sum_{i=1}^{k} \frac{W(A_i, \overline{A_i})}{vol(A_i)} = \sum_{i=1}^{k} \frac{cut(A_i, \overline{A_i})}{vol(A_i)}$$

Partitioning a graph

- Minimize cut leads to solution which separate one individual vertex from the rest of the graph.
- By dividing the cut by $vol(A_i)$, we explicity request that the sets $A_1, ...A_k$ are reasonably large.

Problem : minimizing Ncut is NP-Hard $\rightarrow$ Spectral clustering is a way to solve relaxed version of this problem.

Spectral clustering

Algorithm Normalized spectral clustering (Shi and Malik 2000)

L=D-W

Compute the k first eigenvector $u_1, ..., u_k$ of $L_{rw} = D^{-1}L$ by solving $Lu = \lambda Du$

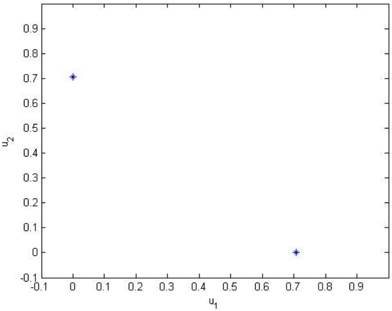
$$U = \begin{pmatrix} u_1(1) & \dots & u_k(1) \\ \vdots & \dots & \vdots \\ u_1(i) & \dots & u_k(i) \\ \vdots & \dots & \vdots \\ u_1(n) & \dots & u_k(n) \end{pmatrix} = \begin{pmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{pmatrix} y_i \in \mathbb{R}^k$$

$[C_1, ..., C_k] \leftarrow kmeans(\{y_i\}_{i=1..n}, k)$

Output : Clusters $A_1, ...A_k$ with $A_i = \{x_j | y_j \in C_i\}$

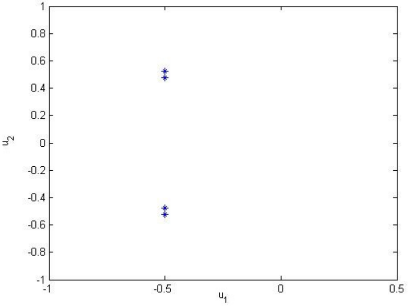
Why does it work ?

$$W = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix} \quad U = \begin{pmatrix} 0.7 & 0 \\ 0.7 & 0 \\ 0 & 0.7 \\ 0 & 0.7 \end{pmatrix}$$



Why does it work ?

$$W = \begin{pmatrix} 1 & 1 & 0.2 & 0 \\ 1 & 1 & 0 & 0.1 \\ 0.2 & 0 & 1 & 1 \\ 0 & 0.1 & 1 & 1 \end{pmatrix} \quad U = \begin{pmatrix} -0.5 & -0.4745 \\ -0.5 & -0.5243 \\ -0.5 & 0.4745 \\ -0.5 & 0.5243 \end{pmatrix}$$



Plan

Information theoritic measures for clustering

Pb : How to evaluate the quality of a clustering ? ([3] Vinh 2010)

Basing the following array, we can compare two clusterings $K = (K_1, ..., K_p)$ et $C = (C_1, ..., C_r)$

	K_1	$\dots$	K_i	$\dots$	K_p	Sum
C_1	$\lvert C_1 \cap K_1 \rvert$	$\dots$	$\lvert C_1 \cap K_i \rvert$	$\dots$	$\lvert C_1 \cap K_p \rvert$	a_1
$\vdots$		$\dots$		$\dots$		
$C_{i'}$	$\lvert C_{i'} \cap K_1 \rvert$	$\dots$	$\lvert C_{i'} \cap K_i \rvert$	$\dots$	$\lvert C_{i'} \cap K_p \rvert$	$a_{i'}$
$\vdots$		$\dots$		$\dots$		
C_r	$\lvert C_r \cap K_1 \rvert$	$\dots$	$\lvert C_r \cap K_i \rvert$	$\dots$	$\lvert C_r \cap K_p \rvert$	a_r
Sum	b_1	$\dots$	b_i	$\dots$	b_p	$\sum_{ij} n_{ij} = n$

with $n_{ij} = |C_i \cap K_j|$

Information theoritic measures for clustering

$$H(C) = -\sum_{i=1}^{r} \frac{a_i}{n} \log\left(\frac{a_i}{n}\right) \qquad \text{Entropy}$$

$$H(C|K) = -\sum_{i=1}^{r}\sum_{j=1}^{p} \frac{n_{ij}}{n} \log\left(\frac{\frac{n_{ij}}{n}}{\frac{b_j}{n}}\right) \qquad \text{Conditional entropy}$$

$$I(C,K) = \sum_{i=1}^{r}\sum_{j=1}^{p} \frac{n_{ij}}{n} \log\left(\frac{\frac{n_{ij}}{N}}{\frac{a_i b_j}{N^2}}\right) \qquad \text{Mutual Information}$$

$$I(C,K) = H(C) - H(C|K) = H(K) - H(K|C)$$

Information theoritic measures for clustering

$$NMI(K, C) = \frac{I(K, C)}{\sqrt{H(C)H(K)}}$$

Normalized Mutual Information

$0 \leq NMI(K, C) \leq 1$
if $K = C$ then $NMI(K, C) = 1$

Spectral Clustering vs Kmeans

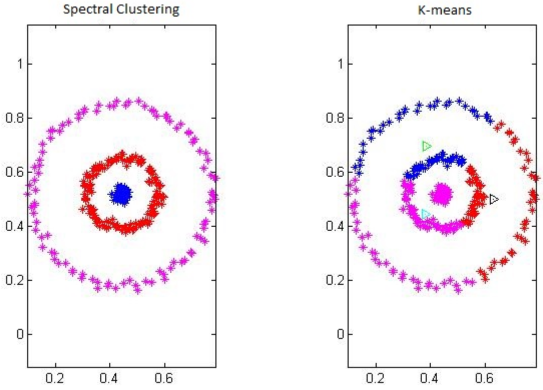
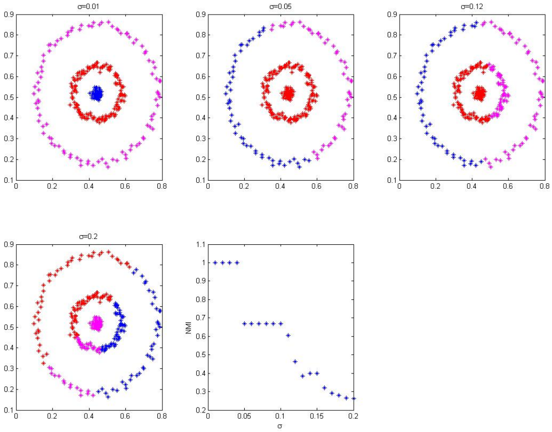
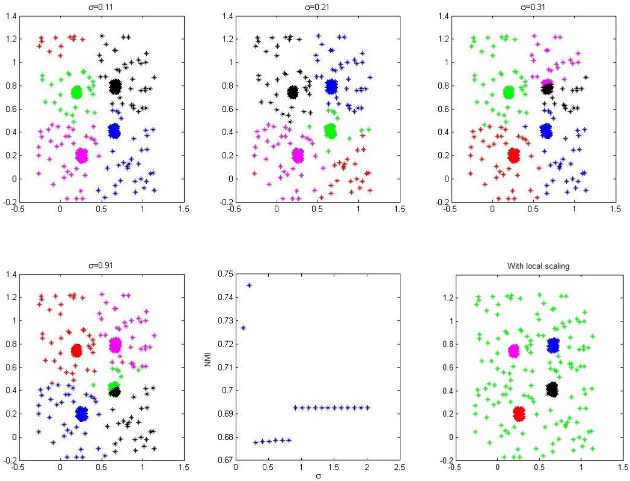


Figure: Spectral Clustering vs Kmeans

Influence of σ (normal graph - 20 neighbors)

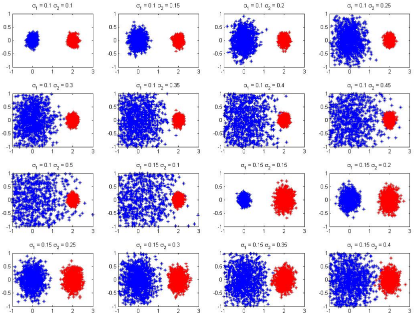


Local scaling (normal graph - 20 neighbors)

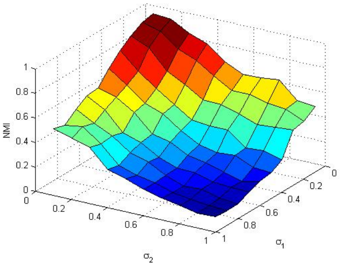


Influence of the noise

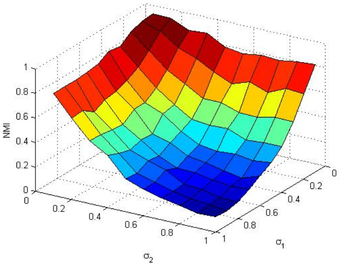
Experimental protocol :
- given two gaussian distributions (1000 points in each) (μ_1, σ_1) and (μ_2, σ_2) where μ_1 and μ_2 are fixed so that $||\mu_1 - \mu_2|| = 1$. We test our algorithm by varying σ_1 and σ_2 from 0.1 to 1



Comparaison Normal Graph - Mutual Graph 2D

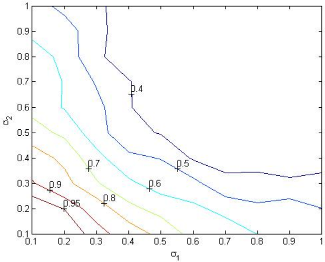


Normal graph

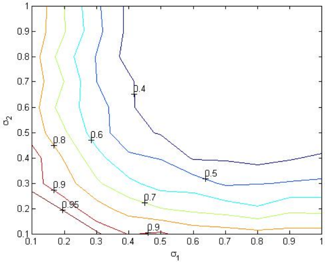


Mutual graph

Contour line Normal Graph - Mutual Graph 2D

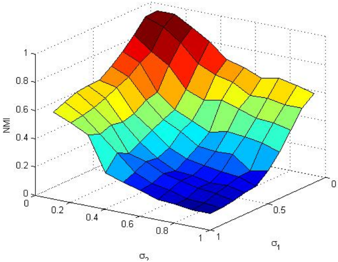


Normal graph

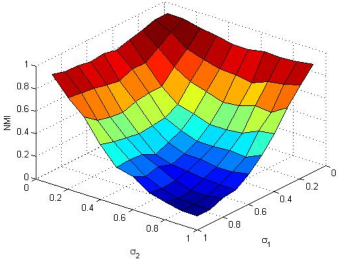


Mutual graph

Comparaison Normal Graph - Mutual Graph 3D

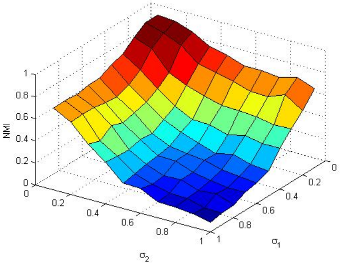


Normal graph

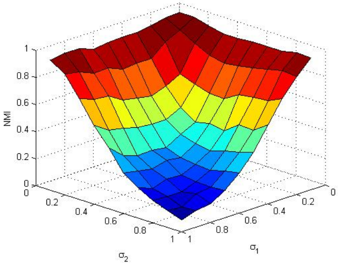


Mutual graph

Comparaison Normal Graph - Mutual Graph 4D



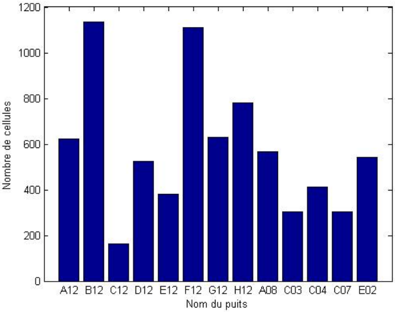
Normal graph



Mutual graph

Results on biological data

Pb : the number of cluster k is unknown.
$\rightarrow$ We test our algorithm for different values of k and we keep which has the largest value of NMI
Database :



Results on biological data

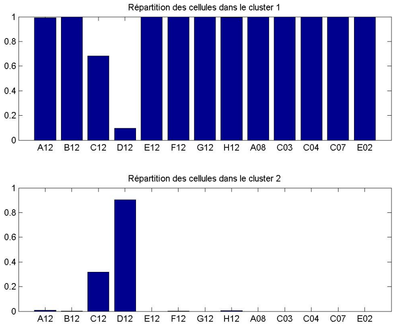
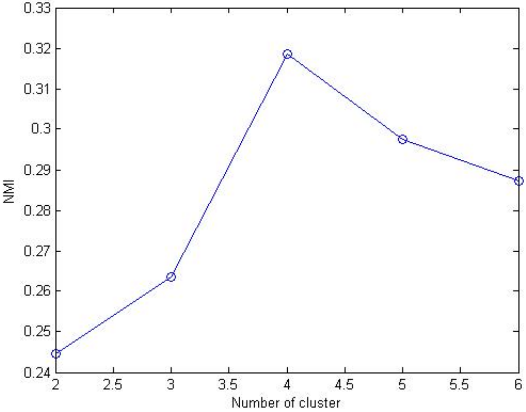


Figure: 2 Clusters (normal graph - 100 neighbors)

Results on biological data



Results on biological data

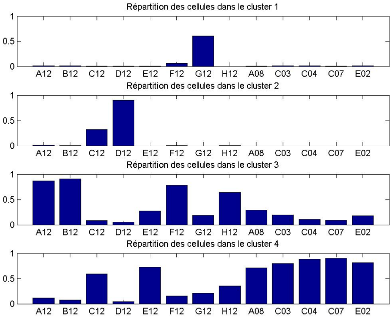


Figure: 4 Clusters

Conclusion

Advantages of spectral clustering :

- quite simple to implement
- good results on our dataset

Future work :

- use other algorithm than kmeans to separate eigenvector
- clustering on one well of cells to identify the noise

Thanks for your attention
Any questions ?

Bibliography

[1] Zelnik-Manor L. Perona P. 'Self-tuning on Spectral Clustering' (2005)
[2] Von Luxburg U. 'A Tutorial on Spectral Clustering' *Statistics and Computing, 17 (4)* (2007)
[3] Vinh N. Epps J. 'Information Theoretic Measures for Clusterings Comparaison : Variants, Properties, Normalization and Correction for Chance' *Journal of Machine Learning Research 11 2837-2854*(2010)
[4] Shi J. Jambo J. 'Normalized Cut and Image Segmentation' (2000)