Mutual Spectral Clustering: Microarray Experiments versus Text Corpus

K. Pelckmans, S. Van Vooren, B. Coessens,
J.A.K. Suykens, B. De Moor
<kristiaan.pelckmans@esat.kuleuven.esat.be>

june, 2006

Overview

- Mutual Clustering
- Mutual Spectral Clustering
- Preliminary experiments
- Extensions
- Discussion

Mutual Clustering

Learning problem:

Given iid sample $\{(X_i, Z_i)\}_{i=1}^n \sim F_{XZ}$,

search for mutual rules $\{(\mathcal{C}_k^X, \mathcal{C}_k^Z)\}_{k=1}^K$ such that

$$\mathcal{C}_k^X(X) \Leftrightarrow \mathcal{C}_k^Z(Z), \quad (X, Z) \sim F_{XZ}$$

why:

- *Transducing* information over different representations
- e.g. genes represented in graph from Microarray experiments, and same genes represented in graph based on text corpus
- Looking for overrepresented clusters in various representations
- If given $(X_i, ?)$ and $\mathcal{C}_k^X(X_i)$, then with high probability $\mathcal{C}_k^Z(?)$, *and vice versa!*.

Mutual Spectral Clustering

2 Undirected graphs G_1 and G_2 having the same nodes and symmetrical weights $\{w_{ij}^1\}_{i\neq j}$ and $\{w_{ij}^2\}_{i\neq j}$.

Minimal cut in two representations:

$$\min_{q\in\{-1,1\}^n} \frac{\pi_1}{4}\sum_{i\neq j} w_{ij}^{(1)}(q_i-q_j)^2+\frac{\pi_2}{4}\sum_{i\neq j} w_{ij}^{(2)}(q_i-q_j)^2$$

Spectral relaxation $q\in\{-1,1\}^n \rightarrow \|q\|=1$:

$$L_{\pi_1,\pi_2}q=\lambda q$$

with extended Laplacian

$$L_{\pi_1,\pi_2}=\left(\pi_1 D^{(1)}+\pi_2 D^{(2)}\right)-\left(\pi_1 W^{(1)}+\pi_2 W^{(2)}\right)$$

Fiedler vector $\hat{q}=q_{(2)}$ associated with second eigenvalue of L_{π_1,π_2}.

Microarray Experiments and Text Corpus

Given set of genes $g = \{g_1, \ldots, g_n\}$.

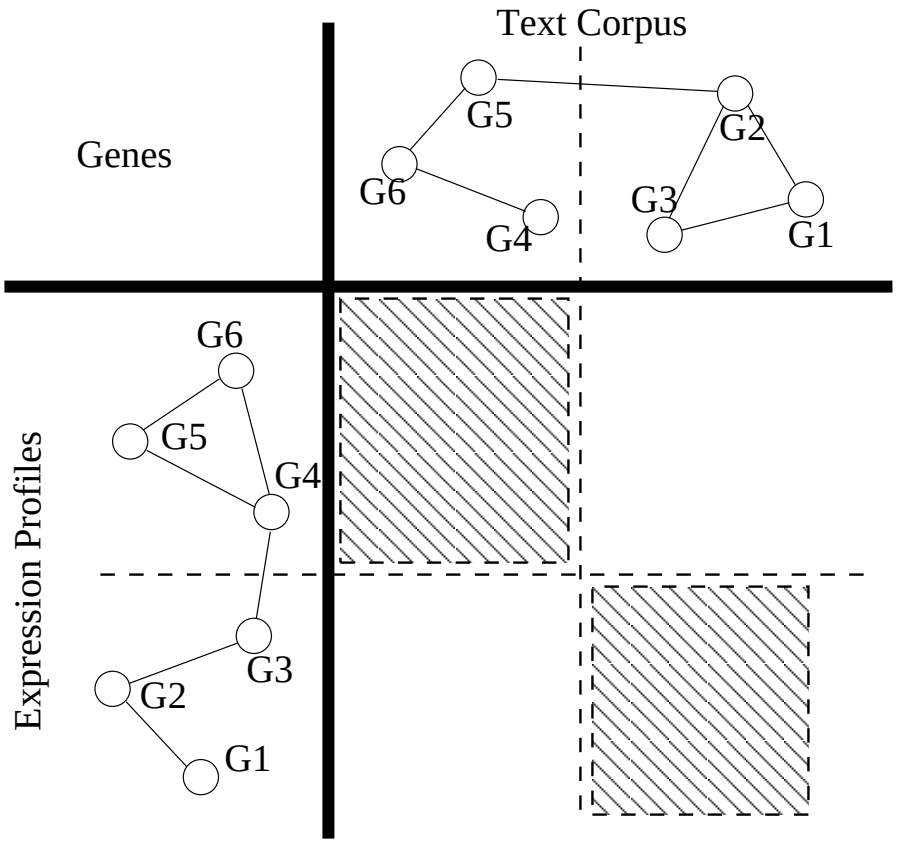
G_1: graph of g based on similarity between citing abstracts in PubMed: w_{ij}^1 by cosine rule on vector term representation of abstracts citing g_i and g_j, respectively.

G_2: graph of g based on correlations in microarray experiments: w_{ij}^2 by RBF-distance on expression level for different conditions for g_i and g_j, respectively.

$\rightarrow$ Preliminary experiment:
51 genes of motor activity and visual perception.

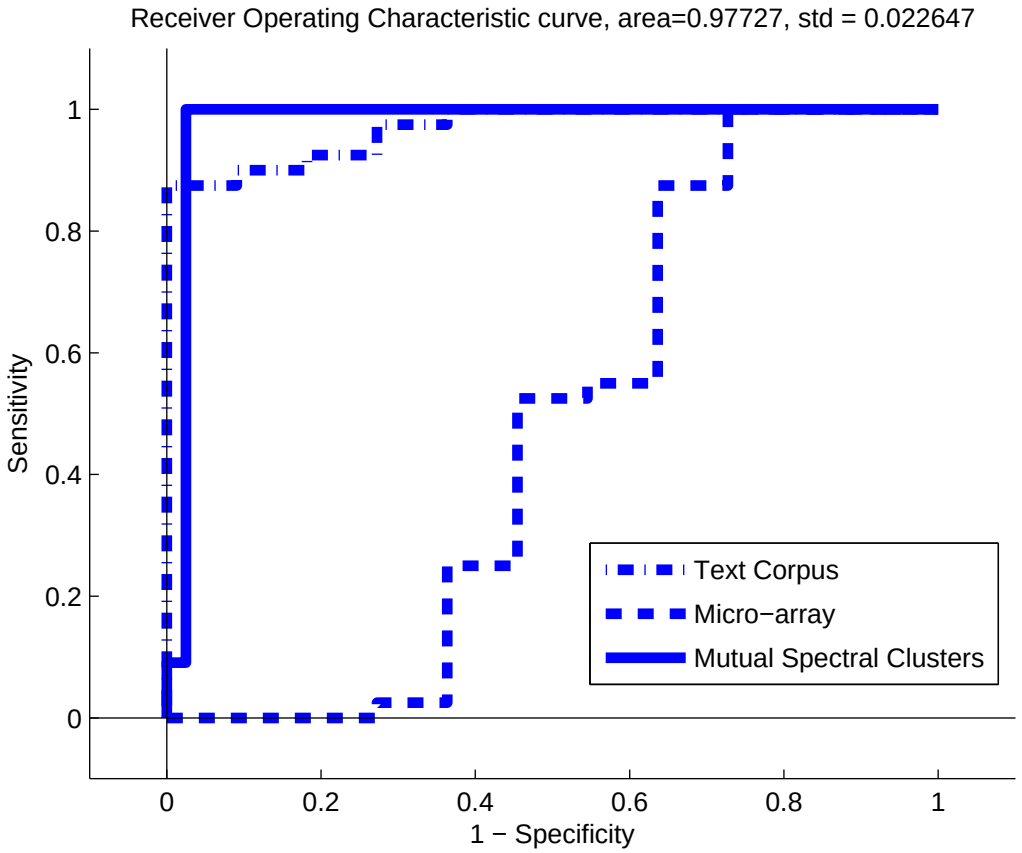
Microarray Experiments and Text Corpus (Ct'd)

Schematical example:



Microarray Experiments and Text Corpus (Ct'd)

ROC curve relating $\hat{q}$ with known label *motor* or *visual* (no need for thresholding)



Extensions

Extension operator:

$$\begin{cases} R^1(n_*;q) = \operatorname{sign}\left(\sum_{j=1}^n q_j w_{*j}^{(1)}\right) \\ R^2(n_*;q) = \operatorname{sign}\left(\sum_{j=1}^n q_j w_{*j}^{(2)}\right). \end{cases}$$

Label q_i consistent with $\operatorname{sign}(W_i q)$ if $q_i(W_i q) > \rho$:

$$H_\rho = \left\{ q \in \{-1,1\}^n \;\middle|\; \forall i : q_i W_i^1 q \geq \rho,\ q_i W_i^2 q \geq \rho \right\}$$

- Stochastic model: $(W^1, W^2) \sim F_{W^1 W^2}$
- q act as 'parameters' with hypothesis space H_ρ
- Future use of rule $\operatorname{sign}(W_*^1 q)$ likely to correspond with unknown label q_*^1 without need for reclustering
- Probably $\operatorname{sign}(W_*^1 q) = \operatorname{sign}(W_*^2 q)$ for all $*$
- If $|H_\rho| < 2^n$:$\Rightarrow$, efficient learning possible.

Discussion

- Mutual Clustering as learning paradigm
- Appriate for graph mining: between unsupervised and supervised learning
- Missing values - prediction
- Gene prioritization and fusing data sources (Endeavour)
- Zoom on small but coherent groups of relevant cluster(s)
- Weakly connected nodes
- Quantifying probabilistic confidence