Predictive Mutual Clustering: Learning in Bioinformatics

K. Pelckmans, J.A.K. Suykens, B. De Moor
K.U.Leuven - ESAT - SCD/sista, Belgium
<kristiaan.pelckmans@esat.kuleuven.esat.be>

October, 2006

Overview

- Learning Task: Predictive Mutual Clustering
- Transductive Inference for Graphs
- Predictive Graph Cuts
- Toy Example
- Application in bioinformatics
- Discussion

Predictive Mutual Clustering

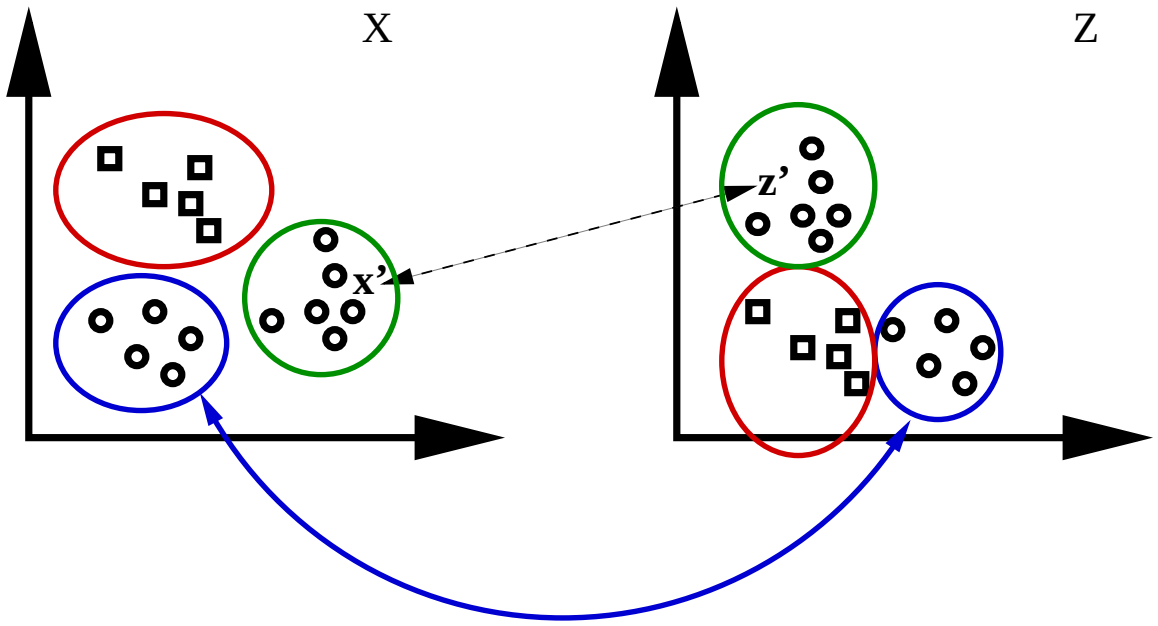
- Given iid sample $\{(X_i, Z_i)\}_{i=1}^{n} \sim F_{XZ}$, with $X_i \in \mathbf{X}$ and $Z_i \in \mathbf{Z}$.
- Task: search for K predictive mutual clusters (rules):

$$\left\{ \left(\mathcal{C}_k^X \subset \mathbf{X}, \mathcal{C}_k^Z \subset \mathbf{Z} \right) \right\}_{k=1}^{K}$$

such that for all $(X, Z) \sim F_{XZ}$:

$$I\left(X \in \mathcal{C}_k^X\right) \approx I\left(Z \in \mathcal{C}_k^Z\right), \quad \forall k = 1, \ldots, K.$$

- Example:



Predictive Mutual Clustering (Ct'd)

'Memberships of any couple $\{(\mathcal{C}_k^X, \mathcal{C}_k^Z)\}_k$ coincide':

- Actual Risk:

$$\mathcal{R}(\{\mathcal{C}_k^X, \mathcal{C}_k^Z\}_k) = \sup_k E\left[\ell\left(I\left(X \in \mathcal{C}_k^X\right), I\left(Z \in \mathcal{C}_k^Z\right)\right)\right],$$

  with loss $\ell : \{0,1\}^2 \to \{0,1\}$.

- Empirical Risk:

$$\mathcal{R}_n(\{\mathcal{C}_k^X, \mathcal{C}_k^Z\}_k) = \sup_k \frac{1}{n}\sum_{i=1}^{n} \ell\left(I\left(X_i \in \mathcal{C}_k^X\right), I\left(Z_i \in \mathcal{C}_k^Z\right)\right).$$

- Towards prediction if all elements in cluster $\mathcal{C}_k^X$ and $\mathcal{C}_k^Z$ for all k cannot deviate too much (concentrated):

$$\forall x, x' \in \mathbf{X} : \mathcal{C}_k^X(x), \mathcal{C}_k^X(x') \Rightarrow \|x - x'\| \leq \rho.$$

- Trade-off large $\mathcal{C}_k^X$ (*membership!*) vs. small $\mathcal{C}_k^X$ (*prediction given membership!*).

Predictive Mutual Clustering: Bioinformatics

Motivations:

- Uncertainty in data $\leftrightarrow$ set membership
- Filling in missing information $(x, ?)$ and $(?, z)$
- Precise goal clustering (verifiable/falsifiable)
- E.g. genes represented in *graph* from Microarray experiments, and same genes represented in *graph* based on text corpus
- Looking for overrepresented (*invariant*) clusters in various representations (cfr. data fusion)

Transductive Inference for Graphs

Setting:

- Fixed amount of $n \in \mathbb{N}$ nodes (objects) $V = \{v_1, \ldots, v_n\}$
- Organized in *deterministic* graph $\mathcal{G}_n = (V, E)$ with edges $E = \{x_{ij} \geq 0\}_{i \neq j}$ (symmetrical $x_{ij} = x_{ji}$, no loops $x_{ii} = 0$)
- Fixed label $y_i \in \{-1, 1\}$ for any node $i = 1, \ldots, n$, but only partly observed: $\mathcal{S} \subset \{1, \ldots, n\}$

$$y_{\mathcal{S}} = \{y_i \in \{-1, 1\}\}_{i \in \mathcal{S}}\,.$$

- Predict the remaining labels

$$y_{\neg\mathcal{S}} = \{y_i \in \{-1, 1\}\}_{i \notin \mathcal{S}}\,.$$

Transductive Inference (Ct'd)

- Hypothesis set:

$$\mathcal{H} = \{q_i \in \{-1, 1\}\}$$

  with $|\mathcal{H}| = 2^n$

- Given a restricted hypothesis set $\mathcal{H}' \subset \mathcal{H}$ with $|\mathcal{H}'| \ll |\mathcal{H}|$, and a few observations $y_{\mathcal{S}}$ where $\mathcal{S}$ is iid without replacement, bound?

- Actual risk

$$\mathcal{R}(q) = E[I(y_* q_* < 0)],$$

  with E_* over iid choice of choice $y_* \in \{y_i\}_i$.

- Empirical risk

$$\mathcal{R}_{\mathcal{S}}(q) = \frac{1}{n} \sum_{i \in \mathcal{S}} I(y_i q_i < 0).$$

Transductive Inference (Ct'd)

- **Generalization Bound:**

  Let $\mathcal{S} \subset \{1, \ldots, n\}$ be iid sampled without replacement. Consider a set of hypothetical labelings $\mathcal{H}' \subset \mathcal{H}$ having a cardinality of $|\mathcal{H}'| \in \mathbb{N}$. Then the following inequality holds with probability higher than $(1 - \delta) < 1$.

$$\sup_{q \in \mathcal{H}'} \mathcal{R}(q) - \mathcal{R}_{\mathcal{S}}(q) \leq \sqrt{\frac{2(n - n_s + 1)}{n_s n} \log(|\mathcal{H}'|) - \log(\delta)}, \tag{1}$$

  where n_s equals the number of observed samples $|\mathcal{S}|$.

- *Proof* $\rightarrow$ Serfling's inequality and union bound over finite hypothesis set.

- Uniformly sampled without replacement $\rightarrow$ knowledge of $\frac{1}{n} \sum_{i=1}^{n} I(y_i q_i < 0)$.

Transductive Inference: Restrictions

- *Graph MINCUT:*

$$\mathcal{H}'_{\rho} = \left\{ q \in \{-1,1\}^n \quad \text{s.t.} \quad \frac{1}{n} \sum_{q_i \neq q_j} x_{ij} \leq \rho \right\}$$

- *Consistent Predictor Rule:*

$$q_i f(v_i) = q_i \sum_{ij} x_{ij} q_i = q_i \left(x_i^T q \right) > 0, \ \forall i = 1, \ldots, n$$

- *Balancing constraints:*

$$\sum_{i=1}^{n} (1 + q_i) \leq 2B$$

- *Fixed labels*
- *Others:* MAXCUT, coloring, matching,...

$\longrightarrow$ Imposing sufficient regularization/prior problem knowledge tightens statistical guarantee.

Transductive Inference: Restrictions (Ct'd)

Restricted set $\mathcal{H}'$ with CUT$< \rho$:

$$\begin{aligned}\mathcal{H}'_\rho &= \left\{ q \in \{-1,1\}^n \quad \text{s.t.} \quad \frac{1}{n} \sum_{q_i \neq q_j} x_{ij} \leq \rho \right\} \\ &= \left\{ q \in \{-1,1\}^n \quad \text{s.t.} \quad \frac{q^T(D-X)q}{q^T q} \leq 2\rho \right\},\end{aligned}$$

where $X_{ij} = x_{ij}$ and $D = \mathrm{diag}(X 1_N)$. Can be *relaxed* as

$$\mathcal{H}''_\rho = \left\{ q = \mathrm{sign}(w) \quad \text{s.t.} \quad \frac{w^T(D-X)w}{w^T w} \leq 2\rho \right\},$$

thus w in eigenspace U'' corresponding to lowest eigenvalues $\Sigma = \{\sigma_i \leq 2\rho\}$ of Laplacian $(D-X)$. Thus

$$|\mathcal{H}'_\rho| \leq |\{q = \mathrm{sign}(U''w)\}| \leq \left(\frac{ne}{|\Sigma|}\right)^{|\Sigma|}.$$

Transductive Inference vs. Clustering

- Isomorphism cluster $\leftrightarrow$ hypothesis q, label '$+1$' and '-1' indicates whether node v_i in cluster.
- Disjunct clusters $\leftrightarrow$ every label $+1$ in exactly one hypothesis q.
- Not optimal hypothesis $\hat{q}$, but examining hypothesis space $\mathcal{H}'$ explicitly.
- Domain knowledge shapes $\mathcal{H}'$
- Clusters concentrated (MINCUT, consistent rule,...)
- Given a node v_i labeled '$+1$', pick the corresponding hypothesis q. Then generalization bound $|\mathcal{R}(q) - \mathcal{R}_i(q)|$ will give bound on the remaining cluster members *(transductive setting)*.

Predictive Graph Cuts

Setting:

- Nodes (objects, genes,...) $V = \{v_1, \ldots, v_n\}$ organized in graphs $G_x = (V, E_x)$ and $G_z = (V, E_z)$, and edges $E_x = \{x_{ij}\}_{i \neq j}$ and $E_z = \{z_{ij}\}_{i \neq j}$.
- How to incorporate this structure in an appropriate restricted $\mathcal{H}'$? Define rules

$$\begin{cases} f(v_i) = \operatorname{sign}\left(\sum_{j=1}^{n} x_{ij} q_j\right) = \operatorname{sign}(x_i^T q), & \forall i \\ g(v_i) = \operatorname{sign}\left(\sum_{j=1}^{n} z_{ij} q_j\right) = \operatorname{sign}(z_i^T q), & \forall i. \end{cases}$$

  restrict together space $\mathcal{H}$ as follows

$$\mathcal{H}_\rho^2 = \left\{ q \in \{-1, 1\}^n \quad \text{s.t.} \quad \frac{q_i(x_i^T q)}{n} \geq \rho, \ \frac{q_i(z_i^T q)}{n} \geq \rho, \ \forall i \right\}$$

  where $\frac{q_i(x_i^T q)}{\|q\|_2}$ and $\frac{q_i(x_i^T q)}{\|q\|_2}$ are the margin of f and g
- But difficult to work with...

Predictive Graph Cuts (Average CUT)

Sum CUT$< c$

$$\mathcal{H}_c = \left\{ q \in \{-1,1\}^n \quad \text{s.t.} \quad \frac{1}{n} \sum_{q_i \neq q_j} x_{ij} + \frac{1}{n} \sum_{q_i \neq q_j} z_{ij} \leq c \right\}$$

Relaxation

$$\mathcal{H}'_c = \left\{ q = \text{sign}(w) \quad \text{s.t.} \quad \frac{w^T (L_x + L_z) w}{w^T w} \leq c \right\},$$

with $L_x = (\text{diag}(X 1_N) - X)$ and $L_z = (\text{diag}(Z 1_N) - Z)$. Bound on cardinality

$$|\mathcal{H}_c| \leq |\{\text{sign}(U_{\vartheta(c)}) w\}| \leq \left(\frac{en}{\vartheta(c)} \right)^{\vartheta(c)}$$

Predictive Graph Cuts (Product Margin)

Product of margin $\frac{q_i f(v_i)}{\|q\|_2}$ and $\frac{q_i g(v_i)}{\|q\|_2}$ becomes

$$m_i(q) = \frac{1}{\|q\|_2^2} q^T (x_i z_i^T) q, \quad \forall i$$

Average product margin

$$\overline{m}(q) = \frac{1}{n\|q\|_2^2} \sum_{i=1}^{n} q^T (x_i z_i^T) q = \frac{1}{n} \frac{q^T (XZ) q}{q^T q}$$

Hypothesis set

$$\mathcal{H}_\rho^p = \left\{ q \in \{-1, 1\}^n \quad \text{s.t.} \quad \frac{1}{n} \frac{q^T (XZ) q}{q^T q} \geq \rho \right\}$$

or

$$\mathcal{H}_\rho^p = \left\{ q = \operatorname{sign}(U_{\vartheta(\rho)} w) \right\}$$

Predictive Graph Cuts (Algorithm)

Pick hypothesis $\hat{q}$ corresponding optimally with observations $y_{\mathcal{S}}$

$$\hat{w} = \arg\max_{w \in \mathcal{H}_\rho} \frac{1}{n^2} \sum_{i \in \mathcal{S}} y_i (x_i^T w)$$

or

$$\hat{w}' = \arg\max_{w'} \sum_{i \in \mathcal{S}} y_i \left(x_i^T U_\vartheta w' \right) \quad \text{s.t.} \quad \|U_\vartheta w'\|_2 = 1$$

Solved easily as

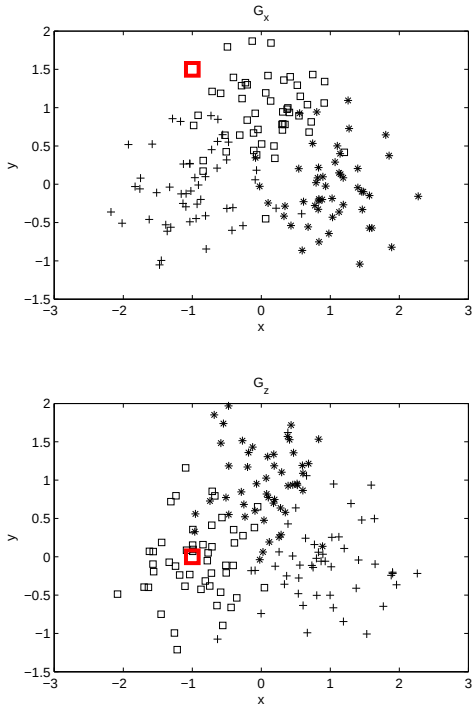
$$w' = c Y^T X U_\vartheta$$

with $Y \in \{-1, 0, 1\}^n$ where $Y_i = y_i$ for $i \in \mathcal{S}$, and zero elsewhere. Predict labels of remaining nodes as

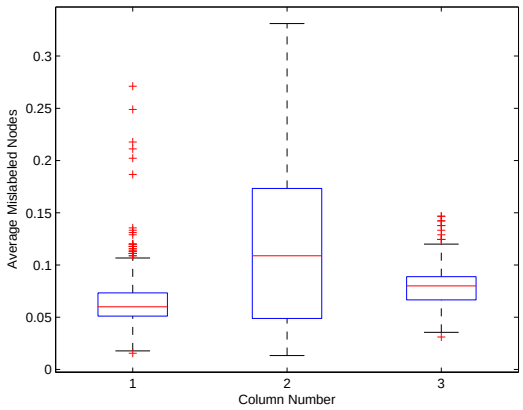
$$\hat{y}_i = \operatorname{sign}\left(U_\vartheta \hat{w}' \right), \quad i \notin \mathcal{S}$$

Toy example

2D display of random graph with 3 classes ('+','□','*'), $n = 150$:



(a)



(b)

Reconstructing 3 hypothesis (clusters) using 3 eigenvectors:

- Average Product Margin
- Average CUT$< c$
- (Sum of) Distance from true cluster center

MC iteration

Microarray Experiments and Text Corpus

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

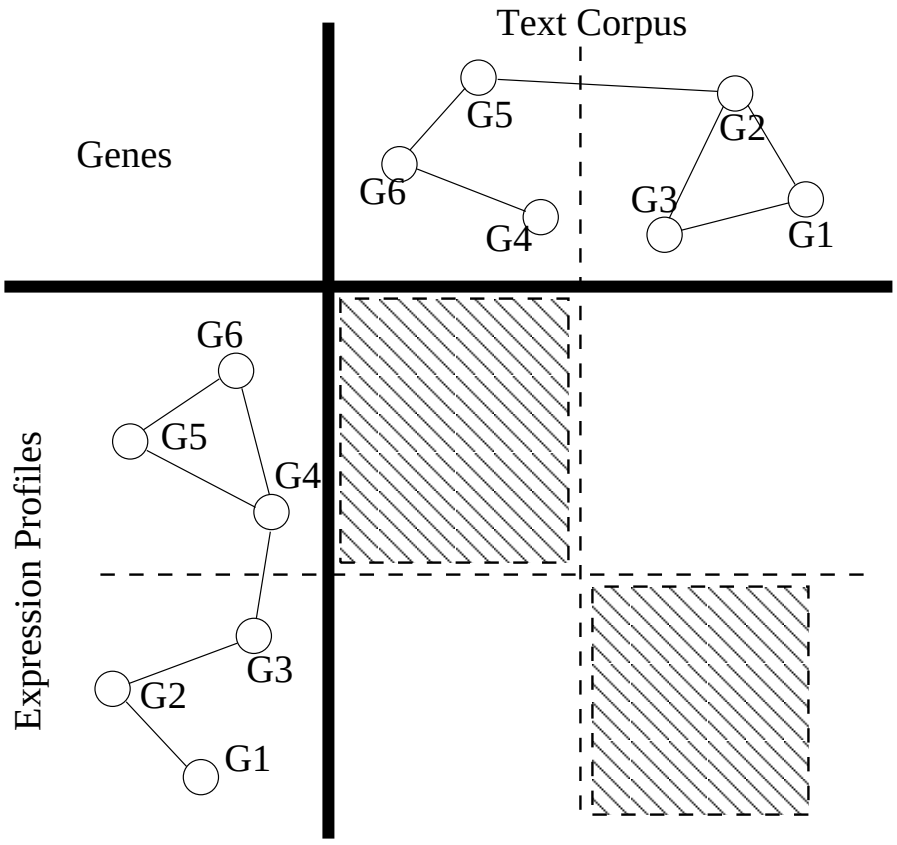
G_1: graph of g based on similarity between citing abstracts in PubMed: w_{ij}^1 by cosine rule on vector term representation of abstracts citing g_i and g_j, respectively.

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.

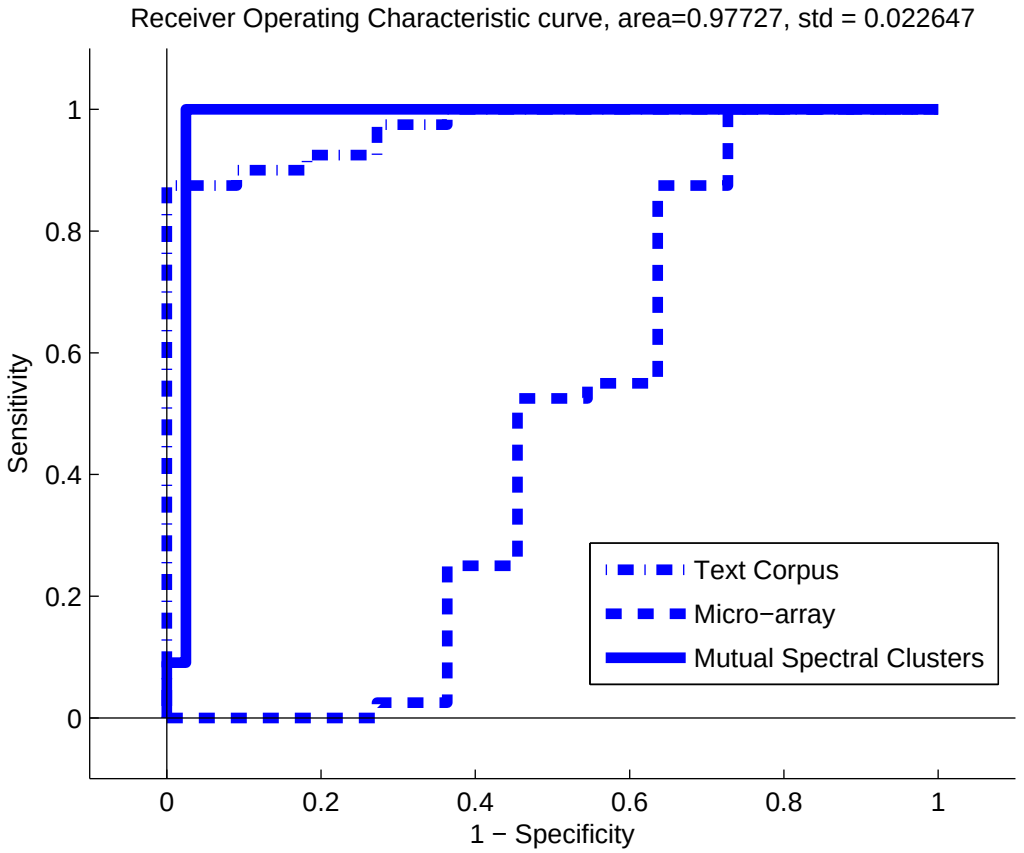
Microarray Experiments and Text Corpus (Ct'd)

Schematical example:



Microarray Experiments and Text Corpus (Ct'd)

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



Discussion

- Predictive Clustering as learning paradigm
- Missing values - prediction
- Appropriate for graph mining: between transductive inf. and clustering
- Quantifying probabilistic confidence
- Gene prioritization and fusing data sources (Endeavour)
- Zoom on small but coherent groups of relevant cluster(s)
- Weakly connected nodes
- Application?
- Metric space $\mathbf{X}$ and $\mathbf{Z}$?