Supervised Clustering of Genes

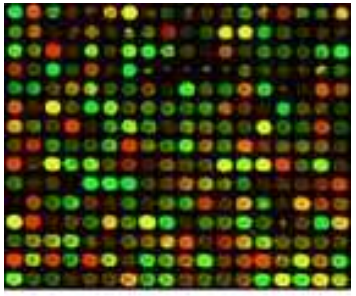
Marcel Dettling
in joint work with Peter Bühlmann
Seminar für Statistik
ETH Zürich, Switzerland

dettling@stat.math.ethz.ch
http://stat.ethz.ch/~dettling

July 13, 2003

Microarray Gene Expression Data

- Microarray technology & preprocessing steps $\rightarrow$ **gene expression matrix**



$$\rightarrow (x_{ij}) = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & \vdots & & \vdots \\ x_{p1} & x_{p2} & \dots & x_{pn} \end{pmatrix}$$

- n experimental tissues, showing expression levels or activities of p genes.
- Typically between 2'000-8'000 genes (variables), but only 20-80 experiments
- Tissue types as supervised categorical response $(y_j) = (y_1, y_2, \dots, y_n)$
- **Goal:** Detect gene clusters that are strongly associated with the response Y

Identification of Markers or Gene Clustering

Revealing gene cluster/markers is important for:

- gaining insight into biological processes & grasping how the genome works
- obtaining good predictors in medical diagnostics, allowing tailored treatment

Identification of Markers or Gene Clustering

Revealing gene cluster/markers is important for:

- gaining insight into biological processes & grasping how the genome works
- obtaining good predictors in medical diagnostics, allowing tailored treatment

Mathematics: A few marker components of genes determine a tissue's type.

$$P[Y = k|X] = f\left(X_{\mathcal{C}_1}, \ldots, X_{\mathcal{C}_q}\right), \qquad q \ll p$$

In words: "If in average, gene 534, gene 837 and gene 235 are overexpressed, and gene 24, gene 931 and gene 694 are underexpressed, this is typical for cancer subtype A"

Why is this difficult?

Identification of Markers or Gene Clustering

Revealing gene cluster/markers is important for:

- gaining insight into biological processes & grasping how the genome works
- obtaining good predictors in medical diagnostics, allowing tailored treatment

Mathematics: A few marker components of genes determine a tissue's type.

$$P[Y = k|X] = f\left(X_{\mathcal{C}_1}, \ldots, X_{\mathcal{C}_q}\right), \qquad q \ll p$$

In words: "If in average, gene 534, gene 837 and gene 235 are overexpressed, and gene 24, gene 931 and gene 694 are underexpressed, this is typical for cancer subtype A"

Why is this difficult?

- there are $2 \cdot 10^{30}$ possible clusters of 10 genes in a dataset of 5000 genes.
- we neither know cluster size, the number of clusters q, nor the function $f(\cdot)$
- how is the representative value $X_{\mathcal{C}_i}$ for cluster $\mathcal{C}_i$ defined?

$\longrightarrow$ exhaustive search by a penalized log-likelihood approach is impossible

Supervised vs. Unsupervised Clustering of Genes

Unsupervised Clustering: Hierarchical Clustering, k-Means, SOM, PCA, . . .

- genes clustered by similarity/correlation, or other criteria based on X-values
- no useful external information about the Y-variables, the response, is used

$\longrightarrow$ doesn't reveal groups of genes with special interest for tissue discrimination

Supervised vs. Unsupervised Clustering of Genes

Unsupervised Clustering: Hierarchical Clustering, k-Means, SOM, PCA, . . .

- genes clustered by similarity/correlation, or other criteria based on X-values
- no useful external information about the Y-variables, the response, is used

$\longrightarrow$ doesn't reveal groups of genes with special interest for tissue discrimination

Supervised Clustering: to be presented . . .

- grouping of variables (genes), controlled by information about the X and Y variables, thus including the supervised outcome of microarray experiments
- can be applied for variable (gene) clustering only, but not for tissue clustering

Supervised vs. Unsupervised Clustering of Genes

Unsupervised Clustering: Hierarchical Clustering, k-Means, SOM, PCA, . . .

- genes clustered by similarity/correlation, or other criteria based on X-values
- no useful external information about the Y-variables, the response, is used

$\longrightarrow$ doesn't reveal groups of genes with special interest for tissue discrimination

Supervised Clustering: to be presented . . .

- grouping of variables (genes), controlled by information about the X and Y variables, thus including the supervised outcome of microarray experiments
- can be applied for variable (gene) clustering only, but not for tissue clustering

$\longrightarrow$ supervised algorithms try to find gene clusters, whose average expression profile has great potential for explaining the response Y, i.e. for tissue discrimination

Supervised vs. Unsupervised Clustering of Genes

Unsupervised Clustering: Hierarchical Clustering, k-Means, SOM, PCA, . . .

- genes clustered by similarity/correlation, or other criteria based on X-values
- no useful external information about the Y-variables, the response, is used

$\longrightarrow$ doesn't reveal groups of genes with special interest for tissue discrimination

Supervised Clustering: to be presented . . .

- grouping of variables (genes), controlled by information about the X and Y variables, thus including the supervised outcome of microarray experiments
- can be applied for variable (gene) clustering only, but not for tissue clustering

$\longrightarrow$ supervised algorithms try to find gene clusters, whose average expression profile has great potential for explaining the response Y, i.e. for tissue discrimination

Need to define a generic strategy and a criterion S for supervised clustering

A Generic Strategy for Supervised Clustering

The strategy is:

1) start from scratch with the best single gene

A Generic Strategy for Supervised Clustering

The strategy is:

1) start from scratch with the best single gene

2) grow the cluster incrementally by adding one gene after the other

A Generic Strategy for Supervised Clustering

The strategy is:

1) start from scratch with the best single gene

2) grow the cluster incrementally by adding one gene after the other

3) occasional stepwise pruning helps to remove spurious genes from the cluster

A Generic Strategy for Supervised Clustering

The strategy is:

1) start from scratch with the best single gene

2) grow the cluster incrementally by adding one gene after the other

3) occasional stepwise pruning helps to remove spurious genes from the cluster

4) if the current cluster cannot be improved by 2) & 3), start a new cluster

A Generic Strategy for Supervised Clustering

The strategy is:

1) start from scratch with the best single gene

2) grow the cluster incrementally by adding one gene after the other

3) occasional stepwise pruning helps to remove spurious genes from the cluster

4) if the current cluster cannot be improved by 2) & 3), start a new cluster

Example: a forward step

Assume that clusters $\mathcal{C}_1, \ldots, \mathcal{C}_p$ with predictor variables $x_1, \ldots, x_p$ are given, and repeat FOR all genes $j = 1, \ldots, p$:

a) construct the candidate cluster $\mathcal{C}_p^j$ and its predictor variable x_p^j

A Generic Strategy for Supervised Clustering

The strategy is:

1) start from scratch with the best single gene

2) grow the cluster incrementally by adding one gene after the other

3) occasional stepwise pruning helps to remove spurious genes from the cluster

4) if the current cluster cannot be improved by 2) & 3), start a new cluster

Example: a forward step

Assume that clusters $\mathcal{C}_1, \ldots, \mathcal{C}_p$ with predictor variables $x_1, \ldots, x_p$ are given, and repeat FOR all genes $j = 1, \ldots, p$:

a) construct the candidate cluster $\mathcal{C}_p^j$ and its predictor variable x_p^j

b) compute the clustering criterion S_j, a (penalized) goodness-of-fit measure

end FOR;

A Generic Strategy for Supervised Clustering

The strategy is:

1) start from scratch with the best single gene

2) grow the cluster incrementally by adding one gene after the other

3) occasional stepwise pruning helps to remove spurious genes from the cluster

4) if the current cluster cannot be improved by 2) & 3), start a new cluster

Example: a forward step

Assume that clusters $\mathcal{C}_1, \ldots, \mathcal{C}_p$ with predictor variables $x_1, \ldots, x_p$ are given, and repeat FOR all genes $j = 1, \ldots, p$:

a) construct the candidate cluster $\mathcal{C}_p^j$ and its predictor variable x_p^j

b) compute the clustering criterion S_j, a (penalized) goodness-of-fit measure

end FOR;

c) the gene $j^* = \arg\min_j S_j$ is the winner

A Generic Strategy for Supervised Clustering

The strategy is:

1) start from scratch with the best single gene

2) grow the cluster incrementally by adding one gene after the other

3) occasional stepwise pruning helps to remove spurious genes from the cluster

4) if the current cluster cannot be improved by 2) & 3), start a new cluster

Example: a forward step

Assume that clusters $\mathcal{C}_1, \ldots, \mathcal{C}_p$ with predictor variables $x_1, \ldots, x_p$ are given, and repeat FOR all genes $j = 1, \ldots, p$:

a) construct the candidate cluster $\mathcal{C}_p^j$ and its predictor variable x_p^j

b) compute the clustering criterion S_j, a (penalized) goodness-of-fit measure

end FOR;

c) the gene $j^* = \arg\min_j S_j$ is the winner

d) if $S_{j^*} < S_{old}$, gene j^* enters the cluster. $\mathcal{C}_p$ and x_p are updated

Implementations of Supervised Clustering

→ To implement the generic strategy, we need a (supervised) clustering criterion S

Implementations of Supervised Clustering

$\rightarrow$ To implement the generic strategy, we need a (supervised) clustering criterion S

Implementation 1: **WILMA** (MD & P. Bühlmann, 2002)
S is the value of the Wilcoxon test statistic, refined by the margin function

Implementations of Supervised Clustering

→ To implement the generic strategy, we need a (supervised) clustering criterion S

Implementation 1: **WILMA** (MD & P. Bühlmann, 2002)
S is the value of the Wilcoxon test statistic, refined by the margin function

Implementation 2: **PELORA** (MD & P. Bühlmann, in preparation)
Based on probabilities $p_\theta(x_i)$ from Penalized logistic regression. S is a penalized goodness-of-fit measure, the negative log-likelihood plus the ℓ_2-Penalty

$$S = -\sum_{i=1}^{n} (y_i \cdot \log p_\theta(x_i) + (1 - y_i) \cdot \log(1 - p_\theta(x_i))) + \lambda \theta^T P \theta$$

Implementations of Supervised Clustering

$\rightarrow$ To implement the generic strategy, we need a (supervised) clustering criterion S

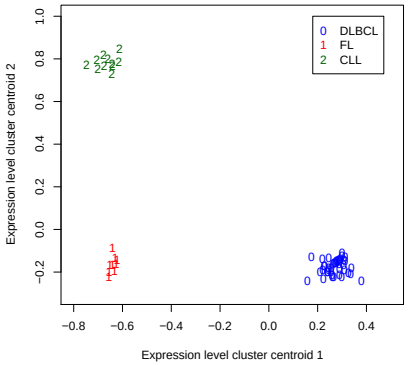
Implementation 1: **WILMA** (MD & P. Bühlmann, 2002)
S is the value of the Wilcoxon test statistic, refined by the margin function

Implementation 2: **PELORA** (MD & P. Bühlmann, in preparation)
Based on probabilities $p_\theta(x_i)$ from Penalized logistic regression. S is a penalized goodness-of-fit measure, the negative log-likelihood plus the ℓ_2-Penalty

$$S = -\sum_{i=1}^{n} (y_i \cdot \log p_\theta(x_i) + (1 - y_i) \cdot \log(1 - p_\theta(x_i))) + \lambda \theta^T P \theta$$

Advantages and improvements of Pelora vs. Wilma:
- clusters can be non-disjoint, this allows to capture multiple pathways
- better interaction between the clusters, they are not "independent"
- milder form of supervision (more robustness) in inhomogeneous problems
- additional clinical variables can be incorporated into the clustering process
- can be adapted to continuous response by using the ℓ_2-loss/ridge regression
- comprises a built-in classifier that yields probabilities for sample prediction

Typical Output and Evaluation of the Algorithm

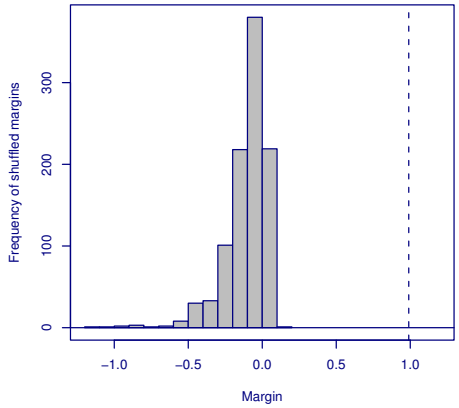


Typical output on more than 10 datasets

- Cluster size: Wilma 3-9 genes, Pelora 15-20 genes
- Very clear separation of the tissue types
- Error-free classification of training data

Permutation test on microarray datasets

- No clearly separating clusters on noise data
- p-value of zero, clusters thus no noise artifact
- Bootstrapping: Clusters are reasonably stable



Predictive Potential for Test Data

10-fold cv	Leukemia	Estro	Nodal	Colon	Prostate	Lymph
Pelora	6.92%	7.25%	18.13%	14.10%	6.53%	0.57%
Wilma	2.62%	8.25%	34.88%	15.05%	7.41%	0.57%
1-NN	2.46%	15.38%	43.25%	15.90%	12.82%	0.67%
SVM	0.92%	11.12%	36.88%	17.62%	8.35%	0.48%

Technical details:

- Pelora: $q = 10$ clusters for the built-in penalized logistic regression classifier
- Wilma: $q = 10$ clusters as predictors in a 1-NN classifier
- 1-NN and SVM with rbf kernel: 200 "best" single genes as predictors

Supervised clustering vs. single gene classifiers

- very competitive, often better than sophisticated state-of-the-art classifiers
- works well with simple classifiers, due to low dimensionality

Pelora vs. Wilma

- the new implementation Pelora has an edge over Wilma
- difference is biggest on difficult problems with high misclassification risk

Conclusions

Supervised algorithms identify gene clusters . . .
- whose average expression makes the discrimination of several different tissue types as simple as it can be
- which are reasonably stable and more than just random noise artifacts

Supervised clusters are (potentially) useful in . . .
- medical diagnostics, because they identify groups of interacting genes with excellent predictive potential, also known as tumor markers
- functional genomics, as they give a clue on pathways, gene interaction and gene regulation

Outlook & Extensions of Pelora:
- additional clinical variables can be part of the clustering process to refine it
- extension to continuous response variables is possible in the same framework

Availability: Software for Wilma/Pelora is available as R-package, contact

dettling@stat.math.ethz.ch