Supervised Clustering of Genes

Marcel Dettling

Seminar für Statistik, ETH Zürich

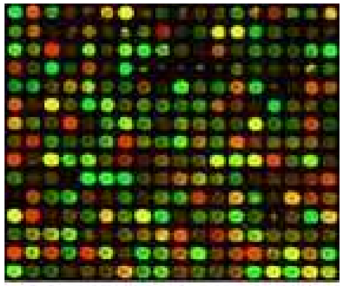
August 25, 2002

dettling@stat.math.ethz.ch

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

Microarray Gene Expression Data

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



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

- 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 $(y_j) = (y_1, y_2, \dots, y_n)$

Assumption: Only a few marker components of genes determine a tissue's type.

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

Identification of Markers *or* Gene Clustering

Very important for:

- Understanding how the genome works and gaining insight to biological processes
- obtaining accurate predictors for medical diagnostics

Previous work:

- *Hierarchical Clustering, Self-Organizing Maps, Principal Component Analysis, . . .*

$\rightarrow$ Unsupervised, clustering genes only by similarity (correlation) among each other

$\rightarrow$ Don't reveal groups of genes with special interest for tissue discrimination

Supervised Clustering of Genes:

Grouping of variables (genes), controlled by information about the Y variables, i.e. the supervised outcome of microarray experiments

- *Tree Harvesting* (Hastie et al., 2000): an indirect 2-step approach
- *Partial Least Squares* (Nguyen & Rocke, 2002): difficult to interpret
- *Supervised Clustering of Genes* (Dettling & Bühlmann, 2002): to be presented . . .

Our Algorithm for Supervised Clustering of Genes ...

...tries to find marker groups of genes, whose average expression profile has great potential for explaining the response Y, i.e. for tissue discrimination

The criterion for clustering

The strength of tissue discrimination by the expression profile of a single gene or a group average is measured by an empirical objective function consisting of:

a) *The Score function*: Wilcoxon's test statistic for two unpaired samples

b) *The Margin function*: A measure for the size of the gap between tissue classes

The clustering procedure

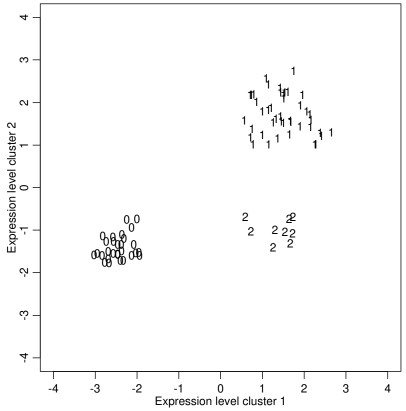
1) Start with the best single gene, according to the criterion

2) Repeat averaging the current cluster with single genes, optimizing the criterion

3) Cleaning stage: Repeat rooting out single genes, optimizing the criterion

4) Repeat forward search and cleaning stages until the cluster stabilizes

Typical Output and Evaluation of the Algorithm



Typical output on more than 10 datasets

●Cluster size is mostly between 3-9 genes

●Very clear separation of the tissue types

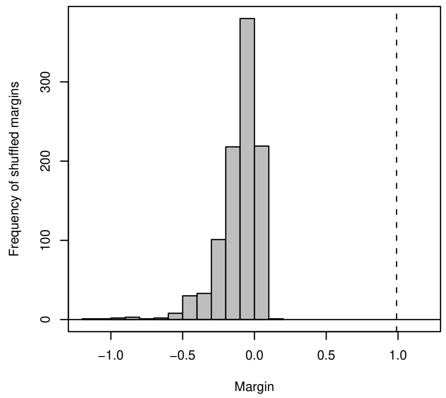
●Error-free classification of training data

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

CV-error	Leukemia	Prostate	Colon	SRBCT	NCI 60
Gene Clusters	1.39%	4.90%	16.13%	0.00%	19.67%
Literature	1.39%	$\approx$ 10%	$\approx$ 12%	0.00%	$\approx$ 35%

Facts about the predictive potential from more than 10 microarray datasets:

- Very competitive, often better than sophisticated state-of-the-art classifiers

- Works well with simple classifiers, due to low dimensionality

- Better with group averages than with the same, non-averaged single genes

Discussion

Why does it work so well:

Microarray data usually contain genes with differential expression, masked by a substantial noise component. Identifying and averaging these differentially expressed genes reduces the noise component and makes discrimination easier.

Conclusion:

It is a sensible, easy to understand clustering heuristic which reveals reasonably stable marker gene groups that are more than just a noise artifacts and that show good predictive potential.

Availability:

Software is freely available as an R-Package called “SupClust”, please contact:

dettling@stat.math.ethz.ch