

A new method to identify significant clusters in gene expression data

Katherine S. Pollard & Mark J. van der Laan

Division of Biostatistics, U.C. Berkeley

`www.stat.berkeley.edu/~laan`

Motivation: Microarray Data

- We observe a matrix X whose columns are n copies of a p -dimensional vector of relative gene expression measurements.
- Each measurement is a ratio, calculated from the intensities of two fluorescently labeled mRNA samples cohybridized to an array spotted with known cDNA sequences.
- Data preprocessing may include background subtraction, normalization, log transformation.

Example: Tumor vs. healthy tissues of n cancer patients.

NOTE: Methodology also applies to gene chips, where each element of X is a quantitative expression level rather than a ratio.

Goals

1. Identifying interesting subsets of genes.
2. Clustering:
 - Genes.
 - Patients.
 - Genes and patients *simultaneously*.
3. Classification and prediction.
4. Defining statistical notions such as parameter, parameter estimate, consistency, and confidence.
5. Assessing the reliability of subsets, clusters, predictors.

Statistical issues are particularly crucial with the high dimensional data structures and relatively small samples of gene expression data.

Clustering

Algorithms map a $p \times p$ dissimilarity matrix \mathbf{D} into p cluster labels.

Approaches:

- Supervised (COBWEB, SVMs, CART, gene-shaving) vs. Unsupervised
- Model-based (AUTOCLASS, SNOB) vs. Nonparametric
- Partitioning (SOMs, PAM, MASLOC, KMEANS) vs. Hierarchical
 1. Agglomerative (single, complete, and average linkage CLUSTER, AGNES)
 2. Divisive (SOTA, DIANA, TSVQ)
- Graphical approaches (CAST)

Clustering

Definition: For any group of elements with cluster labels, Kaufman & Rousseeuw define the **silhouette** for an element as:

$$S_j = \frac{b_j - a_j}{\max(a_j, b_j)},$$

where a_j, b_j are the average dissimilarities of element j with the members of its own cluster and its neighboring cluster, respectively.

Nice Properties:

- General distance metric.
- Robust cluster profiles.
- Sensible ordering (hierarchical).
- Identify parameters of biological interest.

Clustering

Examples:

1. **Partitioning Around Medoids (PAM)**, Kaufman & Rousseeuw (1990).
 - Minimizes over the vector of K potential medoids $\sum_j d_1(x_j, M)$.
 - Each medoid identifies a cluster.
2. **PAMSIL**, van der Laan, Pollard & Bryan (2001).
 - Replaces the PAM criteria function with average silhouette.
 - More “efficient”, identifies small clusters.
3. **Hierarchical Ordered Partitioning and Collapsing Hybrid (HOPACH)**, van der Laan & Pollard (2001).
 - Builds a tree of clusters.
 - At every level, each cluster is split into two or more smaller clusters.
 - Clusters are ordered deterministically based on \mathbf{D} .
 - Collapsing steps correct errors.

Choosing the Number of Clusters

- **Global Criteria:**

1. 30 methods reviewed by Milligan & Cooper, 1985.
2. Phase transitions in simulated annealing (Rose *et al.*, 1990).
3. Graph theory (e.g.: cliques in CAST) (Ben-Dor *et al.*, 1999).
4. Model-based methods (Scott & Symmons, 1971, Roeder, 1994).
5. Average silhouette (Kaufman & Rousseeuw, 1990).

- **Resampling Methods:**

1. Gap statistic (Tibshirani *et al.*, 2000).
2. WADP (Bittner *et al.*, 2000).
3. Clest (Dudoit & Fridlyand, 2001).
4. Bootstrap (van der Laan & Pollard, 2001).

Choosing the Number of Clusters

Problem: Existing criteria identify global structure only.

Approach: For each of a series of proposed clustering results, apply the clustering routine independently to each of the clusters and evaluate the global criteria to obtain a measure cluster heterogeneity. Average over clusters. The minimum indicates the clustering result with most homogeneous clusters.

- Any global criteria.
- Any clustering routine.

Illustration:

- Criteria=silhouette,
- Clustering=PAM, HOPACH.

Mean Split Silhouette

Given a clustering with K clusters, consider each cluster $k = 1, \dots, K$ separately.

1. Apply the clustering algorithm to the elements in cluster k .
2. Choose the number of child clusters that maximizes average silhouette.
3. Call this maximum the split silhouette, SS_k .

Define **Mean Split Silhouette** as:

$$MSS(K) = \frac{1}{k} \sum_{k=1}^K SS_k$$

MSS measures average cluster heterogeneity.

Method: Choose the number of clusters K which minimizes $MSS(K)$.

Simulations: Genes

Parameters

$$n = 30, p = 512$$

$$\mu = (-9, -8, -5, -4, 4, 5, 8, 9), \sigma \in (0.05, 0.5, 0.95, 1.5, 2, 5)$$

D = Euclidean

PAM ($k = 1, \dots, 32$), HOPACH ($l = 1, \dots, 6$)

Simulations: Patients

Parameters

$n = 360$, simulate mean over a group of genes

$\mu = (1, 2, 5, 6, 14, 15, 18, 19)$, $\sigma \in (\approx 0, 0.01, 0.05, 0.15, 0.25, 0.5, 0.95, 0.99)$

$D = |\mu_i - \mu_j|$

PAM ($k = 1, \dots, 32$), HOPACH ($l = 1, \dots, 6$)

Simulations: One Cluster?

- Most global criteria (e.g: silhouette) not defined for 1 cluster.
- Or resampling methods needed in order to estimate 1 cluster.
- MSS can be evaluated easily for 1 cluster.

Simulation with unimodal data:

$$n = 360, p = 1$$

$$\mu = 0, \sigma = 0.05$$

D = Euclidean

PAM ($k = 1, \dots, 20$), HOPACH ($l = 1, \dots, 6$)

Summary

MSS is an example of a general method:

- Maps a clustering algorithm, distance matrix (any metric) and a global criteria into optimal cluster labels.
- Identifies finer structure in gene expression data.
- Provides a measure of cluster heterogeneity.
- Computationally easy.
- Has applications in other contexts.