Exploratory Data Analysis for Microarrays

Jörg Rahnenführer
Universität Dortmund, Fachbereich Statistik
Email: rahnenfuehrer@statistik.uni-dortmund.de

NGFN – Courses in Practical DNA Microarray Analysis
Regensburg, May 8, 2007

Overview
• Classification tasks for microarrays
• Cluster analysis
• Time series example
• Distance measures
• Cluster algorithms
• Comparisons and recommendations
• Gene selection
• Comparative study for tumor classification
• Assessment of cluster validity
• Estimating the number of clusters
• Classification tasks for microarrays

Classification Tasks for Microarrays

• Discriminant analysis: Classes known
• Cluster analysis: Classes not known

Classification

Interative Exploratory Data Analysis

You are all genes...
Cluster Analysis – Distance Measures

**Manhattan distance**

The distance between two vectors is the sum of the absolute differences over all coordinates.

\[ d_M(x, y) = \sum_{i=1}^{n} |x_i - y_i| \]

**Euclidean distance**

The distance between two vectors is the square root of the sum of the squared differences over all coordinates.

\[ d_E(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \]

**Correlation distance**

Distance between two vectors is 1.0 (or 1.0), where p

\[ d_C(x, y) = 1 - r \]

**Distance Measures – Time Series Example**

- **Bicluster**
  - Measurements of gene expression levels across samples
  - Clustering genes with similar temporal expression patterns

- **Time series example**
  - Clustering columns
    - Genes with similar patterns
  - Clustering rows
    - Samples with similar expression trajectories

- **Cluster Analysis – Distance Measures**
  - **Distance measure**
    - Euclidean distance
  - **Cluster algorithm**
    - A procedure to organize observations into clusters
  - **Correlation distance**
    - A measure of similarity or dissimilarity between two objects
Distance Measures - Time Series Example

Summary

• Euclidean distance measures average difference across coordinates.
• Manhattan distance measures average difference across coordinates, in a robust way.
• Correlation distance measures difference with respect to trends.

Standardization

- Data points (e.g. genes) are normalized with respect to mean and variance.
- Standardization makes sense if one is not interested in the magnitude of the effects, but in the effect itself. Results from standardization may be more comparable across studies.
- Standardization makes sense if one is not interested in the magnitude of the effects, but in the effect itself. Results from standardization may be more comparable across studies.

Cluster Algorithms

Appropriate choice of distance measure depends on your intention.

MESSAGE 2

Hierarchical Clustering

• Hierarchical clustering was the first algorithm used in microarray research to cluster genes (Eisen et al. (1998)).

1. First, each object is assigned to its own cluster.
2. Calculation of distance between clusters of (H) between clusters G and H

Hierarchical Clustering

• Calculation of distance between clusters G and H is based on object dissimilarity between the objects from the two clusters:

• Alternative to agglomerative clustering: Divisive clustering:
  - Iteratively, best possible splits are calculated:

Cluster Algorithms

- Correlation distance measures difference with respect to trends.
- Manhattan distance measures average difference across coordinates.
- Euclidean distance measures average difference across coordinates.

Summary

3. Step 2 is repeated until only one single cluster remains.
   - The smallest one is merged with this new node.
   - The two most similar clusters are joined, representing a new node of the clustering tree. The node is computed as average of all objects of the joined clusters.

.getMessage(2)
Hierarchical Clustering

- Visualization of hierarchical clustering with dendrogram:
  - Clusters that are joined are combined by a line.
  - Height of line is average distance between clusters.
  - Cluster with smaller variation typically plotted on left side.

- Procedure provides a hierarchy of clusterings, with the number of clusters ranging from 1 to the number of objects.

Hierarchical clustering does not show the full picture!

Hierarchical Clustering

- Visualization with heat map and dendrogram

Interest in specific genes: If you search for genes that are co-regulated with a specific gene of your choice, do so!

Don't do clustering, but generate a list of genes close to your gene with respect to a distance of your choice.
Cluster Algorithms – K-means

• K-means is a partitioning algorithm with a prefixed number of clusters. It tries to minimize the sum of within-cluster-variances.

  1. The algorithm chooses a random sample of different objects as initial cluster midpoints. Then it alternates between two steps until convergence:
    - Assign each object to its closest of the different midpoints with respect to Euclidean distance.
    - Calculate new midpoints as the averages of all points assigned to the old midpoints, respectively.

• K-means is a randomized algorithm, two runs usually produce different results. Thus, it has to be applied several times to the same data set and the result with minimal sum of within-cluster-variances should be chosen.

Cluster Algorithms – Self-Organizing Maps

• SOM’s are similar to k-means, but with additional constraints.
  - Mapping from input space onto one or two-dimensional array of total nodes.

  - Iteration steps (20,000-50,000):
    - Pick data point \( P \) at random
    - Move all nodes in direction of \( P \), the closest node in network topology most, the further a node is in network topology, the less.
    - Decrease amount of movement with iteration steps.

Data point to Node (cluster prototypes)

Tamayo et al. (1999): First use of SOM’s for gene clustering from microarrays

Cluster Algorithms – PAM

• PAM (Partitioning around medoids, Kaufman and Rousseeuw (1990)) is a partitioning algorithm, a generalization of k-means.

  • For an arbitrary dissimilarity matrix \( d \) it tries to minimize the sum (over all objects) of distances to the closest of \( k \) prototypes.

  • Objective function:
    - \( (d) \): Manhattan, Correlation, ...

  • BUILD phase: Initial ‘medoids’.
  • SWAP phase: Repeat until convergence:
    - Consider all pairs of objects (i,j), where i is a medoid and j not, and make the i-j swap (if any) which decreases the objective function most.

Estimating the Number of Clusters

• Internal indices
  - Statistics based on within- and between-clusters matrices of sums-of-squares and on cross-products (Milligan & Cooper (1985): exhaustive comparison of 30 indices)
  - Estimate is number of clusters \( K \) that minimizes/maximizes an internal index

• Model-based methods

• Gap statistic
  - Resampling method, for each \( K \) compare an observed internal index to its expected value under a reference distribution and look for \( K \) which maximizes the difference (Tibshirani et al., 2001)

  Caution: Does not work in high dimensions (e.g. large numbers of genes)

• Average silhouette width (Kaufman & Rousseeuw, 1990)

  - Heuristic approach: Average silhouette width
  - For each observation’s average silhouette width \( s(i) \) is defined:
    - \( a(i) := \text{average dissimilarity between } i \text{ and all other points of its cluster.} \)
    - For all other clusters \( C \), let \( d(i,C) := \text{average dissimilarity of } i \text{ to all observations of } C \).
    - Define \( b(i) := \min_C d(i,C) \).
    - Define silhouette width:
      \[ s(i) := \frac{b(i) - a(i)}{\max(a(i),b(i))} \]

  - Maximal average silhouette width over all observations can be used to select the number of clusters.
  - Observations with \( s(i) \) close to 1 can be considered well-clustered, observations with \( s(i)<0 \) are misclassified.

  - The optimal number of clusters cannot be determined in general, as the quality of a clustering result depends on the concept of a cluster.
**Cluster Validity**

- **Silhouette plots for clustering Leukemia patients (Chiaretti et al., 2004)**
  - $K=2$ clusters
  - Green: Well separated cluster
  - Red: No clear cluster structure

- **If true class labels are known, the validity of the clustering can be verified by comparing true class labels and clustering labels with external cluster indices.**
- **Number of misclassifications $n_{ij} = \#$ objects in class $i$ and cluster $j$**
- **Iteratively match best fitting class and cluster, and sum up numbers of remaining observations.**
- **Rand index**
  - Probability of randomly drawing 'consistent' pair of observations.

**Cluster Validity - Comparative Study**

- **Comparative study for tumor classification with microarrays:** Comparison of hierarchical clustering, k-means, PAM and SOM's
- **Data sets:**
  - Ross et al.: NCI60 cancer dataset, http://genome-www.stanford.edu/nci60, 9 cancer classes: 9 breast, 6 central nervous system, 7 colon, 8 leukemia, 8 melanoma, 9 lung, 6 ovarian, 2 prostate, 8 renal, cDNA microarray
- **Superiority of k-means with repeated runs (Similar for discriminant analysis: FLDA best, Dudoit et al. 2001)**
- **Superiority of PAM with Manhattan distance especially for noisy data**
- **Differences depend on the specific dataset**

**Gene Selection**

- **Preselection of genes**
  - Various approaches for gene selection, especially in supervised learning.
  - For clustering samples, a practical procedure is to choose the top 100-200 genes with respect to variance (across samples). This decreases noise and computation time.

**Classification**

- **Simple cluster algorithms work better in case of little model knowledge!**
  - (But: More sophisticated methods might be more appropriate with more a priori knowledge)

**Classification**

- **MESSAGE 3**
  - Simple cluster algorithms work better in case of little model knowledge!
  - (But: More sophisticated methods might be more appropriate with more a priori knowledge)
Gene Selection

- Preselection of genes

Various approaches for gene selection, especially in supervised learning. For clustering samples, a practical procedure is to choose the top 100-200 genes with respect to variance (across samples). This decreases noise and computation time.

Distance matrices for clustering Leukemia patients (Chiaretti et al., 2004)

Clustering after supervised feature selection

NO!

Do not first select genes based on the outcome of some covariable (e.g., tumor type) and then look at the clustering. You will ALWAYS find difference w.r.t. your covariable, since this is how you selected the genes!

R commands and libraries

- library(mva)
- Hierarchical clustering:
  hclust()
- Kmeans:
  kmeans()
- Principal components:
  princomp()
- library(cluster)
- PAM:
  pam()
- Silhouette information:
  silhouette()
- library(cclust)
- library(mclust)

SUMMARY

MESSAGE 1:
Discriminant analysis: CLASSES KNOWN
Cluster analysis: CLASSES NOT KNOWN

MESSAGE 2:
Appropriate choice of distance measure depends on your intention!

MESSAGE 3:
Simple cluster algorithms work better in case of little model knowledge!

Literature

Literature


