Exploratory data analysis: clustering

Benedikt Brors

Dept. Intelligent Bioinformatics Systems
German Cancer Research Center
This talk is based on material obtained courtesy of Anja v. Heydebreck, MPI-MG.
Exploratory data analysis/unsupervised learning

- “Look at the data”; identify structures in the data and visualize them.
- Can we see biological/experimental parameters; are there outliers?
- Find groups of genes and/or samples sharing similarity.
- Unsupervised learning: The analysis makes no use of gene/sample annotations.
Clustering

Aim: Group objects according to their similarity.

![Clustering Diagram]
Clustering gene expression data

- Clustering can be applied to rows (genes) and/or columns (samples/arrays) of an expression data matrix.
- Clustering may allow for reordering of the rows/columns of an expression data matrix which is appropriate for visualization (heat map).
Clustering genes

Aims:

- identify groups of co-regulated genes
- identify typical spatial or temporal expression patterns (e.g. yeast cell cycle data)
- arrange a set of genes in a linear order which is at least not totally meaningless
Clustering samples

Aims:

- detect experimental artifacts/bad hybridizations (quality control)
- check whether samples are grouped according to known categories (meaning that these are clearly visible in terms of gene expression)
- identify new classes of biological samples (e.g. tumor subtypes)
Clustering: Distance measures

- Aim: Group objects according to their similarity.

- Clustering requires a definition of distance between the objects, quantifying a notion of (dis)similarity. After this has been specified, a clustering algorithm may be applied.

- The result of a cluster analysis may strongly depend on the chosen distance measure.
A **metric** $d$ is a function satisfying:

1. **non-negativity**: $d(a, b) \geq 0$;
2. **symmetry**: $d(a, b) = d(b, a)$;
3. $d(a, a) = 0$.
4. **definiteness**: $d(a, b) = 0$ if and only if $a = b$;
5. **triangle inequality**: $d(a, b) + d(b, c) \geq d(a, c)$.

A function only satisfying 1.-3. is called a **distance**.
Distance measures: Examples

Vectors $x = (x_1, \ldots, x_n), y = (y_1, \ldots, y_n)$

- Euclidean distance: $d_E(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$

- Manhattan distance: $d_M(x, y) = \sum_{i=1}^{n} |x_i - y_i|$

- One minus Pearson correlation:

$$d_C(x, y) = 1 - \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{(\sum_{i=1}^{n} (x_i - \bar{x})^2)^{1/2}(\sum_{i=1}^{n} (x_i - \bar{x})^2)^{1/2}}$$
Distance measures/standardization

• The correlation distance is invariant wrt shifting and scaling of its arguments:

\[ d_C(x, y) = d_C(x, ay + b), \quad a > 0. \]

• One may apply standardization to observations or variables:

\[ x \rightarrow \frac{x - \bar{x}}{\sigma(x)}, \]

where \( \sigma(x) \) is the standard deviation of \( x \).
The correlation distance and the Euclidean distance between standardized vectors are closely related:

\[ d_E(x, y) = \sqrt{2n d_C(x, y)}. \]
Distances between clusters

Extend a distance measure $d$ to a measure of distance between clusters.

- **Single linkage** The distance between two clusters is the minimal distance between two objects, one from each cluster.

- **Average linkage** The distance between two clusters is the average of the pairwise distance between members of the two clusters.
• **Complete linkage** The distance between two clusters is the maximum of the distances between two objects, one from each cluster.

• **Centroid linkage** The distance between two clusters is the distance between their centroids.
Hierarchical clustering

- Build a cluster tree/dendrogram, starting from the individual objects as clusters.
- In each step, merge the two clusters with the minimum distance between them - using one of the above linkage principles.
- Continue until everything is in one cluster.
- If you want a partition of the set of objects, cut the tree at a certain height.
- R function `hclust` in package `mva`.
Hierarchical clustering, example

Golub data, 150 genes with highest variance
Example: Clustering of rows and columns

**k-means clustering**

- User specifies the number $k$ of desired clusters. Input: Objects given as vectors in $n$-dimensional space (Euclidean distance is used).
- For an initial choice of $k$ cluster centers, each object is assigned to the closest of the centers.
- The centroids of the obtained clusters are taken as new cluster centers.
- This procedure is iterated until convergence.
Self-organizing maps (SOMs), or Kohonen networks, are a special variant of neural networks.

- They may be used for clustering.
- A predetermined number of clusters and a network topology must be chosen. Networks are always rectangular (e.g., 3×4).
- Initially, the network is randomly mapped into data space.
SOM continued

- Training requires a high number of iterations. In each step, one of the data points is chosen randomly.

- This point attracts the nearest node of the net by a certain force. Nodes connected to the nearest node are also attracted, but by smaller forces.

- There is a certain elasticity of the net, exerting a reset-force. Think of the edges of the net as made of rubber.

- With higher iterations, the net freezes and becomes more stiff. Nodes tend to stay near their final positions.
SOM continued

- In the end, the nodes should ideally mark cluster centers. Data points may be assigned to clusters by a number of methods, e.g. Voronoi tessellation, fixed-radius hyperspheres, or non-exclusive methods (which allow points to be in different clusters).
SOM: Initial mapping (2D data space)
SOM: training
SOM: final state
The concept of fuzzy clustering abandons the idea of fixed cluster membership. Instead, there is a certain probability for any object to belong to one of the clusters.

This allows to judge how reliable an assignment to a cluster is.

There are certain variants of fuzzy clustering, mainly build on k-means or c-means (fuzzy c-means, Gath-Geva-Algorithm). Some of them are available from R packages e1071 and cluster (routine fanny).

Fuzzy c-means or k-means assumes spherical clusters of same size, while more advanced algorithms allow for ellipsoidal clusters of differing sizes.
Many methods require the user to specify the number of clusters. Generally it is not clear which number is appropriate for the data at hand.

Several authors have proposed criteria for determining the number of clusters, see Dudoit and Fridlyand 2002.

Sometimes there may not be a clear answer to this question - there may be a hierarchy of clusters.
Which scale, which distance measure to use for clustering?

- Data should be normalized and transformed to an appropriate scale before clustering (\(\log\) or the generalized \(\log\) resulting from variance stabilization (R package \texttt{vsn})).

- Clustering genes: Standardization of gene vectors or the use of the correlation distance is useful when looking for patterns of relative changes - independent of their magnitude.

- Clustering samples: Standardizing genes gives relatively smaller weight for genes with high variance across the samples - not generally clear whether this is desirable.
- Gene filtering (based on intensity/variability) may be reasonable - also for computational reasons.
Some remarks on clustering

- A clustering algorithm will always yield clusters, whether the data are organized in clusters or not.

- The bootstrap may be used to assess the variability of a clustering (Kerr/Churchill 2001, Pollard/van der Laan 2002).

- If a class distinction is not visible in cluster analysis, it may still be accessible for supervised methods (e.g. classification).
References


