Lecture 7

Clustering Algorithms

Bioinformatics Data Analysis and Tools

bvhoute@few.vu.nl
The challenge

• Biologists are estimated to produce 25,000,000,000,000,000 bytes of data each year (± 35 billion CD-roms).

• How do we learn something from this data?

• First step: Find patterns/structure in the data. ⇒ Use cluster analysis
Cluster analysis

- **Definition**: Clustering is the process of grouping several objects into a number of groups, or clusters.

- **Goal**: Objects in the same cluster are more similar to one another than they are to objects in other clusters.
## Simple example

<table>
<thead>
<tr>
<th>Species</th>
<th>Fat (%)</th>
<th>Proteins (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Horse</td>
<td>1.0</td>
<td>2.6</td>
</tr>
<tr>
<td>Donkey</td>
<td>1.4</td>
<td>1.7</td>
</tr>
<tr>
<td>Mule</td>
<td>1.8</td>
<td>2.0</td>
</tr>
<tr>
<td>Camel</td>
<td>3.4</td>
<td>3.5</td>
</tr>
<tr>
<td>Llama</td>
<td>3.2</td>
<td>3.9</td>
</tr>
<tr>
<td>Zebra</td>
<td>4.8</td>
<td>3.0</td>
</tr>
<tr>
<td>Sheep</td>
<td>6.4</td>
<td>5.6</td>
</tr>
<tr>
<td>Buffalo</td>
<td>7.9</td>
<td>5.9</td>
</tr>
<tr>
<td>Fox</td>
<td>5.9</td>
<td>7.4</td>
</tr>
<tr>
<td>Pig</td>
<td>5.1</td>
<td>6.6</td>
</tr>
<tr>
<td>Rabbit</td>
<td>13.1</td>
<td>7.1</td>
</tr>
<tr>
<td>Rat</td>
<td>12.6</td>
<td>12.3</td>
</tr>
<tr>
<td>Deer</td>
<td>19.7</td>
<td>9.2</td>
</tr>
<tr>
<td>Reindeer</td>
<td>20.3</td>
<td>10.4</td>
</tr>
<tr>
<td>Whale</td>
<td>21.2</td>
<td>11.1</td>
</tr>
</tbody>
</table>
Simple example

Composition of mammalian milk

![Diagram showing the composition of mammalian milk with clusters of data points indicating different categories. The x-axis represents fat (%) and the y-axis represents proteins (%). The clusters are labeled as Clustering.]
Clustering vs. Classification

- **Clustering (unsupervised learning):**
  We have objects, but we do not know to which class (group/cluster) they belong. The task here is trying to find a ‘natural’ grouping of the objects.

- **Classification (supervised learning):**
  We have labeled objects (we know to which class they belong). The task here is to build a classifier based on the labeled objects; a model which is able to predict the class of a new object. This can be done with Support Vector Machines (SVM), neural networks, etc.
Back to cluster analysis

The six steps of cluster analysis:
1. Noise/outlier filtering
2. Choose distance measure
3. Choose clustering criterion
4. Choose clustering algorithm
5. Validation of the results
6. Interpretation of the results
Step 1 – Noise/outlier filtering

• **Noise and outliers** are ‘bad measurements’ in data, caused by:
  – Human errors
  – Measurement errors (i.e. defective thermometer)

• Noise and outliers can negatively influence your data analysis, so it is better to filter them out.

• But be careful! Outliers could be correct, but just rare data.

• A lot of filtering methods exist, but we won’t go into that.
Step 2 – Choose a distance measure

- The **distance measure** defines the dissimilarity between the objects that we are clustering.
- **Examples.** What might be a good distance measure for comparing:
  - Body sizes of persons? \( d = |\text{length}_1 - \text{length}_2| \)
  - Protein sequences? \( d = \text{number of positions with different aa, or difference in molecular weight, or…} \)
  - Coordinates in n-dimensional space? Euclidean distance:
  \[
  d = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}
  \]
We just defined the distances between the object we want to cluster. But how to decide whether two objects should be in the same cluster?

The clustering criterion defines how the objects should be clustered

Possible clustering criterions:
- A cost function to be minimized. This is optimization-based clustering.
- A (set of) rule(s). This is hierarchical clustering.
Step 3 – Choose clustering criterion

- Different types of hierarchical clustering, based on the different clustering criterions:
  - Single linkage / Nearest neighbour
  - Complete linkage / Furthest neighbour
  - Group averaging / UPGMA

(see also the lecture on machine learning)
Step 4 – Choose clustering algorithm

- Choose an algorithm which performs the clustering, based on the chosen distance measure and clustering criterion.
Step 5 – Validation of the results

- Look at the resulting clusters!
- Are the object indeed grouped together as expected?
- Experts mostly know what to expect. For example, if a membrane protein is assigned to a cluster containing mostly DNA-binding proteins, there might be something wrong. (or not!)
Important to know!

• The results depend totally on the choices made. Other distance measures, clustering criterions and clustering algorithms lead to different results!

For example, try hierarchical clustering with different cluster criteria: http://www.elet.polimi.it/upload/matteucc/Clustering/tutorial_html/AppletH.html
Popular clustering algorithms

Some popular clustering algorithms are:

- K-means
- Hierarchical clustering
- Self organizing maps (SOM)
- Fuzzy clustering
K-means

Steps:
1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids (centroid = cluster center = average of all objects in that cluster).

2. Assign each object to the group that has the closest centroid.

3. When all objects have been assigned, recalculate the positions of the K centroids.

4. Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.
K-means
K-means

– **Advantages:**
  - Fast
  - Easy to interpret

– **But:**
  - K-means has to be told how many groups ($K$) to find
  - Easily affected by noise and outliers
  - No measure is provided of how well a data point fits in a given cluster
  - No guarantee to find global optimum: the quality of the outcome depends on the initialisation. => You should run the algorithm several times with different initialisations.
Hierarchical clustering

Dendrogram

Venn Diagram of Clustered Data

Hierarchical clustering

- **Agglomerative HC**: starts with singleton and merge clusters.
- **Divisive HC**: starts with one sample and split clusters.

**Different types:**
- Single linkage
- Complete linkage
- Average linkage
- Average group linkage
Hierarchical clustering

Steps in agglomerative clustering:

1. **Initialization**: each object forms one cluster.
2. Find the two most similar clusters and group them together to form a new cluster.
3. Calculate the similarity between the new cluster and all remaining clusters.

---

**Diagram:**
- Initialization: each object forms one cluster.
- Find the two most similar clusters and group them together to form a new cluster.
- Calculate the similarity between the new cluster and all remaining clusters.
Hierarchical clustering

The distance of two objects is obvious; it is defined by the distance measure of your choice. But what about the distance between two clusters? The distance of two objects is obvious; it is defined by the distance measure of your choice. But what about the distance between two clusters?

From: Introduction to Hierarchical Clustering Analysis, Pengyu Hong
Hierarchical clustering

Single Linkage (= nearest neighbor)

Dissimilarity between two clusters = Minimum dissimilarity between the members of two clusters

Tend to generate “long chains”

From: Introduction to Hierarchical Clustering Analysis, Pengyu Hong
Hierarchical clustering

Complete Linkage

Dissimilarity between two clusters = Maximum dissimilarity between the members of two clusters

Tend to generate “clumps”

From: Introduction to Hierarchical Clustering Analysis, Pengyu Hong
Hierarchical clustering

Average Linkage (≈ UPGMA)

Dissimilarity between two clusters = Averaged distances of all pairs of objects (one from each cluster).

From: Introduction to Hierarchical Clustering Analysis, Pengyu Hong
Hierarchical clustering

Average Group Linkage

Dissimilarity between two clusters = Distance between two cluster means.

From: Introduction to Hierarchical Clustering Analysis, Pengyu Hong
Hierarchical clustering

Example: Nearest neighbor Algorithm

- Nearest Neighbor Algorithm is an agglomerative approach (bottom-up).
- Starts with \( n \) nodes (\( n \) is the size of our sample), merges the 2 most similar nodes at each step, and stops when the desired number of clusters is reached.

Nearest Neighbor, Level 2, $k = 7$ clusters.

Nearest Neighbor, Level 3, $k = 6$ clusters.
Nearest Neighbor, Level 4, $k = 5$ clusters.
Nearest Neighbor, Level 5, $k = 4$ clusters.
Nearest Neighbor, Level 6, k = 3 clusters.
Nearest Neighbor, Level 7, $k = 2$ clusters.
Nearest Neighbor, Level 8, $k = 1$ cluster.
Clustering in Bioinformatics

- Microarray data quality checking
  - Does replicates cluster together?
  - Does similar conditions, time points, tissue types cluster together?
- Cluster genes  Prediction of functions of unknown genes by known ones
- Cluster samples  Discover clinical characteristics (e.g. survival, marker status) shared by samples.
- Promoter analysis of commonly regulated genes

From: Introduction to Hierarchical Clustering Analysis, Pengyu Hong
Functional significant gene clusters

Two-way clustering

Gene clusters

Sample clusters

Functional category: extracellular (all: 395/5021) (highlighted: 14/49, PValue=3.66e-008)
Bhattacharjee et al. (2001)
Human lung carcinomas mRNA expression profiling reveals distinct adenocarcinoma subclasses.
Fuzzy clustering

- **Hard clustering**: Each object is assigned to exactly *one* cluster.

- **Fuzzy clustering**: Each object is assigned to *all* clusters, but to some extent. The task of the clustering algorithm in this case is to assign a membership degree $u_{i,j}$ for each object $j$ to each cluster $i$.

$$\sum_{i=0}^{C} u_{i,j} = 1$$
Fuzzy clustering of gene expression data

• Experimental observation: Genes that are expressed in a similar way (co-expressed) are under common regulatory control, and there products (proteins) share a biological function\(^1\).

• So it would be interesting to know which genes are expressed in a similar way. ⇒ Clustering!

Fuzzy clustering of gene expression data

- Many proteins have multiple roles in the cell, and act with distinct sets of cooperating proteins to fulfill each role. Their genes are therefore co-expressed with different groups of genes.

- The above means that one gene might belong to different groups/clusters.

⇒ We need ‘fuzzy’ clustering.