Wong Man Yiu 1155160210

BMEG3105 Data analytics for personalized genomics and precision medicine

Lecture 7 Clustering

Lecturer: Yu LI Wednesday, 27 September, 20223

Outline of lecture:

1.Clustering

- Why clustering?
- What is clustering?
- How to do clustering

2.Similarity and dissimilarity measurement

3. Hierarchical clustering

Why clustering?

- Help cluster items for better organization and faster searching.

E.g., online shopping: we would see different categories, such that we could find our targets more quickly

- Cluster people
  - For treatment: As different patients require different treatments, so there are different divisions in hospital or specified hospital, e.g., cancer treatment, dental, children, elderly
  - For sale: Different groups of customers have different needs. By identifying them, the product can be optimized based on the need of the targeting group, so as to cater their interests. Noted that it is not necessarily to group customers by age or gender

In biology:

- Cluster gene
  - To identify co-expressed genes which may involve in the same pathway, which means those genes together would fulfill some specific functions within the cell

- To identify differentially expressed genes which related to diseases. Some of the genes may be highly expressed in some tumors, but less expressed in normal cells, and we may regard this cluster of gene maybe related to cancer

- Cluster samples/cells
  - May help identify new disease sub-types. E.g., there are different stages for tumor
  - May help identify new cell types

What is clustering analysis?

- Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups

- In other words, we want the intra-cluster difference between objects are small, but the inter-cluster different are large

- This brings to the problem of how we define "differences" and its size

Uses of clustering analysis

- Understanding
  - As a stand-alone tool to get insight into data distribution
  - As a pre-processing step for other algorithms
  - E.g., group related documents for browsing, group genes and proteins that have similar functionality, or group stocks with similar price fluctuations, discover new groups (cell types)

- Summarization
  - Reduce the size of large data sets, for example cells/genes in the same group would have similar function, we can use a single cell/gene to represent the whole group

  - Preserve privacy, e.g., in medical data, we don't need to identify the patient

Wong Man Yiu 1155160210

Elements needed to do clustering

- Data to be clustered
- Similarity measurement
- Clustering algorithm (the executive procedure)

Similarity and dissimilarity

- Similarity
  - Numerical measure of how alike two data objects are
  - Higher when objects are more alike
  - Often falls in the range [0,1]
- Dissimilarity (distance)
  - Numerical measure of how different two data objects are
  - Lower when objects are more alike
  - Minimum dissimilarity is often 0
  - Upper limit varies (may not have upper limit)

Similarity measurement:

- Cosine similarity
  - Calculate the cosine of the angle forming by 2 vectors
  - It is computed by the dot product between 2 vectors over the product of their norms
  - Formula:

$$cos(d_1, d_2) = \frac{d_1 \cdot d_2}{(|d_1| * |d_2|)}$$

  - If the result is 0, that means the angle is 90°
- Correlation
  - Measure the linear relationship between objects
  - It is computed by the covariance of 2 variables over product of their standard deviations
  - Formula

$$\rho_{X,Y} = \operatorname{corr}(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\mathrm{E}[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

Wong Man Yiu 1155160210

- Euclidean distance
  - Calculate the straight-line distance between 2 points
  - Formula:

$$Ed(\boldsymbol{p},\boldsymbol{q})=\sqrt{\sum_{k=1}^{m}(p_k-q_k)^2}$$

Where m is the number of dimensions (attributes) and p_k and q_k are, respectively, the k-th attributes (components) or data objects p and q.

  - Normalization is necessary if scales of different dimension differ

- Minkowski distance
  - Minkowski Distance is a generalization of Euclidean Distance
  - Formula:

$$dist(\boldsymbol{p},\boldsymbol{q})=(\sum_{k=1}^{m}|p_k-q_k|^r)^{\frac{1}{r}}$$

Where r is a parameter, m is the number of dimensions (attributes), and p_k and q_k are, respectively, the k-th attributes (components) or data objects p and q.

Consider 3 cases:

1. r = 1. City block (Manhattan, taxicab, L_1 norm) distance.
   - A common example of this is the Hamming distance, which is just the number of bits that are different between two binary vectors (summation of difference along x- and y- axis)

2. r = 2. Euclidean distance
   Example:

(r = 1.: red line; r = 2: blue line)

3. $r \to \infty$. "supremum" (L_{max} norm, L_{∞} norm) distance
   - This is the maximum differences between any component of the vectors

- Mahalanobis distance
  - Calculating distance considering the data distribution
  - Formula:

$$mahalanobis(\boldsymbol{p},\boldsymbol{q}) = (\boldsymbol{p}-\boldsymbol{q})^T \Sigma^{-1} (\boldsymbol{p}-\boldsymbol{q})$$

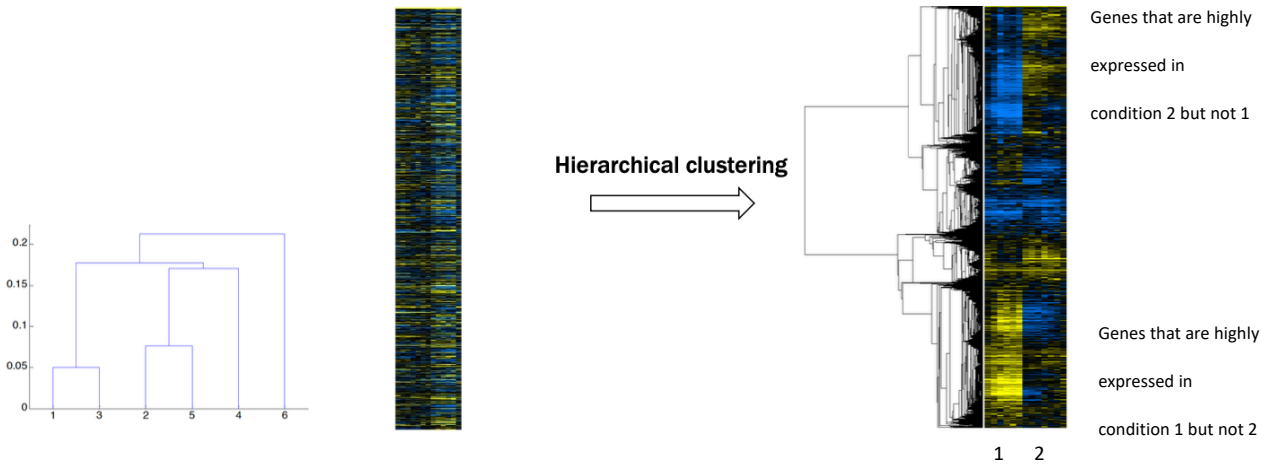
Where $\sum$ is the covariance matrix

Hierarchical clustering

- Produces a set of nested clusters organized as a hierarchical tree

- Can be visualized as a dendrogram
  - A tree like diagram that records the sequences of merges

- They may correspond to meaningful taxonomies
  - Gene clusters, phylogeny reconstruction, animal kingdom...



Wong Man Yiu 1155160210

Steps of Hierarchical clustering

1. Compute the Similarity or Distance matrix
2. Let each data point be a cluster
3. Merge the two closest clusters
4. Update the similarity or distance matrix
5. Repeat steps 2-4 until only a single cluster remains

Methods to update the distance matrix after merging

- Min
- Max
- Group average
- Distance between centroids

➢ Running example

Data matrix:

Gene	wt	mutant_1	mutant_2	mutant_3
At4g35770	1.5	3	3	1.5
At1g30720	4	7.5	7.5	5
At4g27450	1.5	1	1	1.5
At2g34930	10	25	23	15
At2g05540	1	1	2	1

Here we choose to compute the distance matrix with **linear correlation** and update the distance matrix with **minimum distance**.

1. Compute the Similarity or Distance matrix with linear correlation.

	At4g35770	At1g30720	At4g27450	At2g34930	At2g05540
At4g35770	1				
At1g30720	0.9733	1			
At4g27450	-1	-0.9733	1		
At2g34930	0.9493	0.9909	-0.9493	1	
At2g05540	0.5774	0.562	-0.5774	0.4528	1

Values in the table represent the correlation coefficient of 2 genes

2. Let each gene be a cluster. As the correlation coefficient between gene 2 and 4 is the largest, which means they are the closest clusters, so we would merge them.

	At4g35770	At1g30720	At4g27450	At2g34930	At2g05540
At4g35770					
At1g30720	0.9733				
At4g27450	-1	-0.9733			
At2g34930	0.9493	0.9909	-0.9493		
At2g05540	0.5774	0.562	-0.5774	0.4528	

3. Update the similarity or distance matrix with minimum distance (largest coefficient). For example, the distance between gene1 and gene2 is 0.9733, while distance between gene1 and gene 4 is 0.9493. So the distance between gene1 and the cluster gene 2 and 4 would be 0.9733.

	At4g35770	At1g30720	At4g27450	At2g34930	At2g05540
At4g35770					
At1g30720	0.9733				
At4g27450	-1	-0.9733 ->-0.9493			
At2g34930	0.9493 ->0.9733		-0.9493		
At2g05540	0.5774	0.562	-0.5774	0.4528 ->0.562	

4. After updating the matrix, we would see the correlation coefficient between gene 1 and cluster gene2 and 4 is the largest, so we would merge them.

	At4g35770	At1g30720	At4g27450	At2g34930	At2g05540
At4g35770					
At1g30720	0.9733				
At4g27450	-1	-0.9493			
At2g34930	0.9733		-0.9493		
At2g05540	0.5774	0.562	-0.5774	0.562	

5. Update the similarity or distance matrix with minimum distance.

	At4g35770	At1g30720	At4g27450	At2g34930	At2g05540
At4g35770					
At1g30720					
At4g27450	-1 ->-0.9493	-0.9493			
At2g34930			-0.9493		
At2g05540	0.5774	0.562 ->0.5774	-0.5774	0.562 ->0.5774	

6. As the correlation coefficient between gene 5 and the cluster gene1, 2 and4 is the largest, so we would merge them.

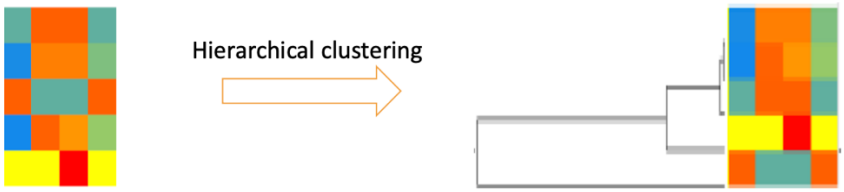
	At4g35770	At1g30720	At4g27450	At2g34930	At2g05540
At4g35770					
At1g30720					
At4g27450	-0.9493	-0.9493			
At2g34930			-0.9493		
At2g05540	0.5774	0.5774	-0.5774	0.5774	

7. Lastly, as there is only gene 3 left, so merge it with the cluster gene1, 2, 4 and 5.

	At4g35770	At1g30720	At4g27450	At2g34930	At2g05540
At4g35770					
At1g30720					
At4g27450	-0.5774	-0.5774			
At2g34930			-0.5774		
At2g05540			-0.5774		

Node2
Node1
Node3

- After clustering, we can shuffle the rows to put those who are similar together for better visualization.



Noted whether we need the original data matrix to update the distance matrix depends on our methods (i.e., we need it if we update the distance matrix with distance between centroid).