Cluster Analysis

Jake Sauter

Background

- Cluster analysis is the most frequently used multivariate technique for analyzing gene sequence expression data
- Clustering is appropriate where there is no a priori knowledge about the data (unsupervised technique)
- In this situation, the only possible approach is to study the similarity between different samples or experiments
- Clustering has become so popular in this field that most authors presenting results obtained with microarrays feel the need to include some type of clustering diagram in their papers

Background

- Clustering is the process of grouping together similar entities

  input: n-dimensional vector

  argument: measure of similarity / distance / metric

  output: many different types, but mostly groups of similar inputs

- The input space of the problem is a n-dimensional space, where n can be the number of samples or the number of feautures

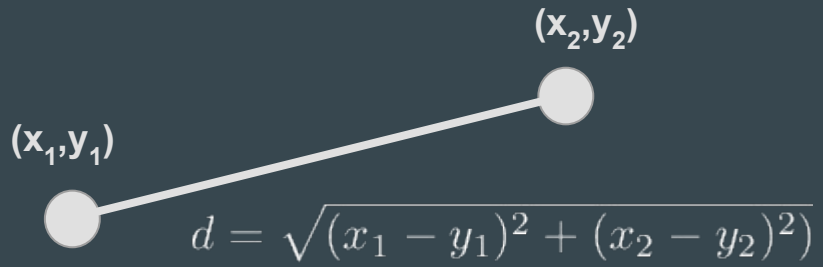
Distances

- A distance metric *d* is a function that takes as arguments two points x and y in an n-dimensional space R^n and has the following properties:
  - Symmetry: The distance should be symmetric such that d(x,y) = d(y,x)
  - Positivity: The distance between any two points should be a real number greater than or equal to 0
  - Triangle Inequality: The distance between two points x and y should be shorter than or equal to the sum of the distances from x to a third point z and y to z (the distance should be the shortest path between two points).

Distances

- There are many different valid distance measures, we will go over a few of them here
- Euclidean Distance: What is thought of normally as "distance", a very intuitive distance metric

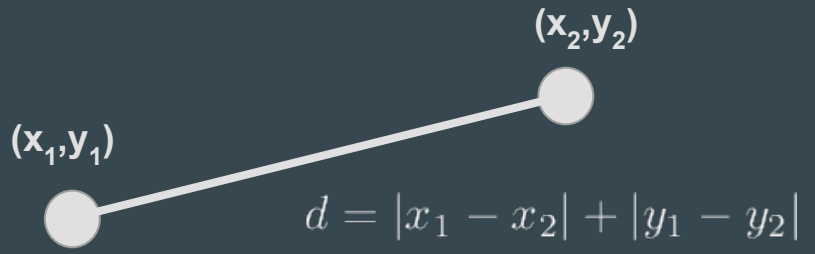
$$d_E(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2}$$
$$= \sum_{i=1}^{n} (x_i - y_i)^2$$



Distances

- Manhattan Distance: Can easily be thought of as how many city blocks must be walked to get from point a to point b on a blocked city design. Only movements along axis directions are allowed
  - This distance measure slightly emphasized outliers as a change of one unit in one coordinate direction leads to a 14% change with respect to Euclidean distance

$$d_m(x,y) = |x_1 - y_1| + |x_2 - y_2| + \cdots + |x_n - y_n|$$
$$= \sum_{i=1}^{n} |x_i - y_i|$$



Distances

- Pearson Correlation distance: Will be proportional to the covariance of two coordinates.
  - This is effective when the points are experiments and dimensions are genes, allowing experiments with very highly correlated genes to be close together
  - This can be used for testing the reliability of equipment or experimental conditions

$$d_R(x, y) = 1 - r_{xy} \quad \text{where}$$

$$r_{xy} = \frac{s_{xy}}{\sqrt{s_x}\sqrt{x_y}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Distances

- The Pearson distance can be a very bad measure if a gene is incorrectly measured!
  - The jackknife correlation aims to solve this issue with leaving out one dimension each iteration
  - The selected value is then the minimum correlation value

$$d_j(x,y) = min\{d_R^1(x,y), d_R^2(x,y) \cdots d_R^n(x,y)\}$$

where d^k_R is d_R with the k-th element removed

- However, this measure is only robust to 1 outlier, so it is still not a great measure

Clustering

- The results of clustering algorithms differs, though it is usually a form of a set of clusters
- Clustering is not necessarily deterministic, the same clustering algorithm applied to the same data may produce different results
    - Some clustering algorithms start with a random choice of clusters
- Membership of a pattern to a cluster should be taken with a grain of salt and further analysed
- The fact that two patterns belong to the same cluster does not necessarily mean that are close to one another

Warnings

- ANYTHING can be clustered
- Given enough patterns, they will always cluster
- There is no scientific value in that there are genes that behave in a similar way, given the amount of genes in the genome and common sample sizes
    - The Scientific value should come from what can be said about the genes that fall in the same cluster and what can be done with said genes
- In most cases, clustering is highly dependent on the distance metric used
    - Changing the distance metric may dramatically affect the number and membership of the clusters, as well as the relationship between them

K-Means Clustering

- K-Means is one of the simplest, fastest and most widely used clustering algorithms
- K-Means takes the number of desired clusters (k) as an input argument
- K Means clustering algorithm :
  - Randomly assign k points as the centers of the clusters
  - Calculate the distance from every point to every cluster center
  - Assign each point to a cluster
  - Reassign cluster centers as the mean of each cluster
  - Recalculate the centroid of each cluster
  - Repeat this process until no pattern moves from one cluster to another

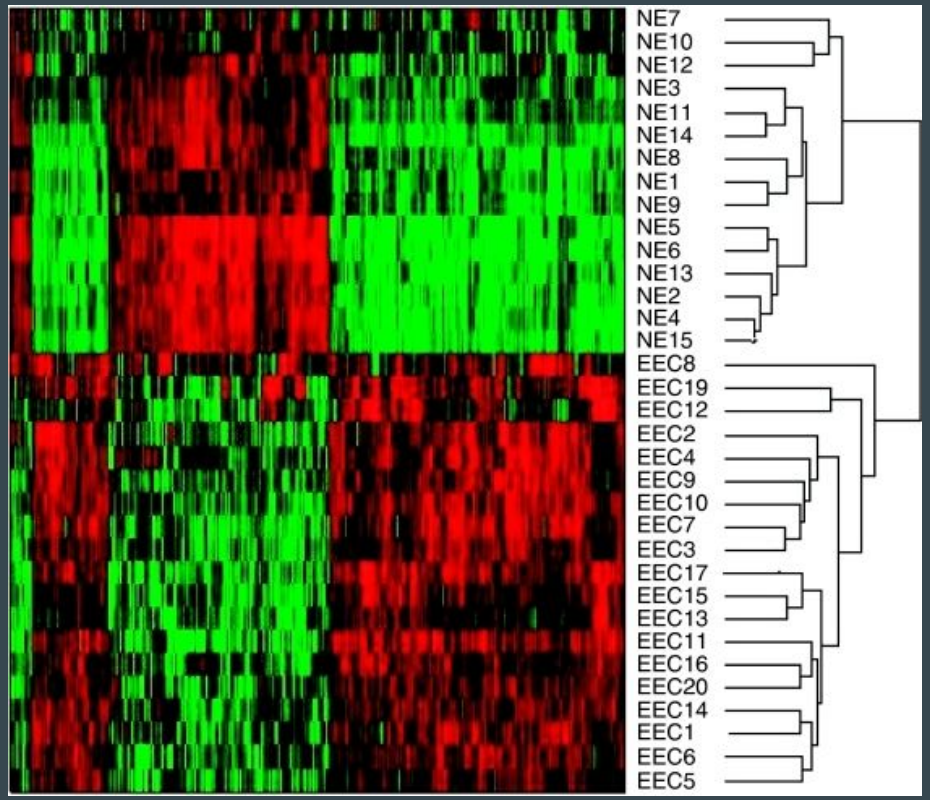
K-Means Clustering

- With K-Means clustering, care should be taken in centroid initialization so that a cluster is not initialized far away from all points, leaving an empty cluster
- A common practice is to initialize centroids as k points chosen randomly from the existing patterns
- This ensures that
  - The starting cluster centers are in an area populated by data
  - Each cluster will have at least one pattern

Hierarchical Clustering

- Hierarchical clustering has been used since the very beginning of the microarray field
- This method aims at the more ambitious task of providing the definitive clustering that characterized a set of patterns the context of a given distance metric
- The result of hierarchical clustering is a complete tree with individual patterns as leaves and the root as the convergent point of all branches, called a dendrogram
- This dendrogram represents a hierarchy of categories based on the degrees of similarity

Hierarchical Clustering



[1]

Hierarchical Clustering

- This method is deterministic and can be applied in a bottom-up (agglomerative) or top-down (divisive) method
- Bottom Up Hierarchical Clustering :
  - Assign n clusters, each containing one pattern
  - Compute the distance from each cluster to every other cluster
  - Merge the two most similar clusters
  - Repeat distance calculation and merging until only one cluster remains

Hierarchical Clustering

- Top Down Hierarchical Clustering
  - Consider the whole set of patterns to be clustered, and use any of a large number of non-hierarchical clustering algorithm to divide the set into two clusters
    - K-Means with k=2 is a possible choice
  - Recursively repeat this process on each of the smaller clusters as they are obtained
  - Terminate when all small clusters contain a single pattern

Partitioning Around Medoids (PAM)

- PAM Clustering starts with computing a dissimilarity matrix from the original data structure with a distance measure of choice
- After this dissimilarity matrix is computed, the resulting distance matrix is mapped into a specified number of clusters
- This algorithm is almost the same as K-Means, with a difference being **cluster centers must be a present data point (medoid)**
- The medoids are representations of the cluster centers, which are robust with respect to outliers (in the same way median is robust to outliers)
    - This is particularly important in the common situation in which many elements do not have a clear-cut membership to any specific cluster

Biclustering

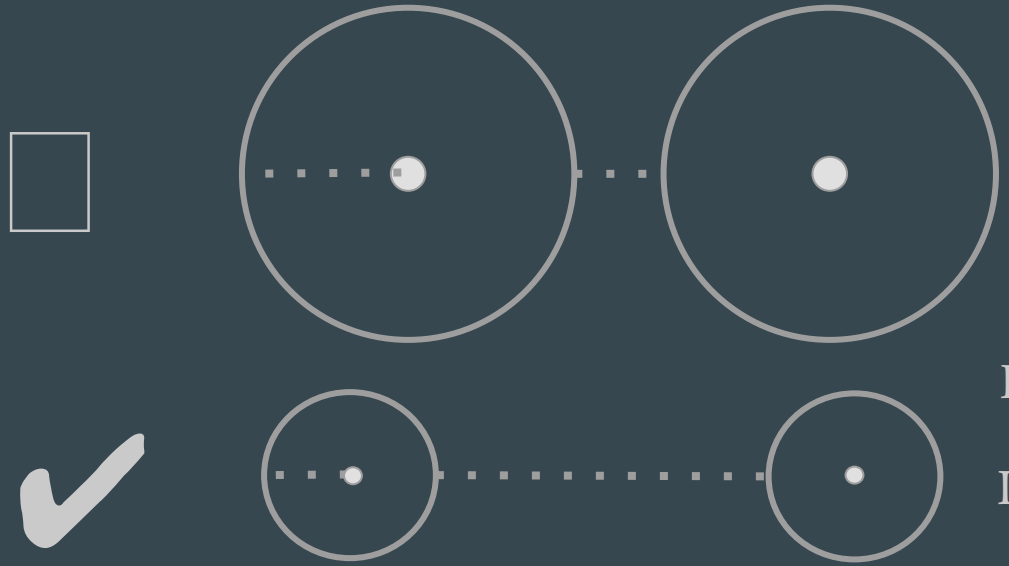
- One can observe that in microarray data, the activities of genes are not independent of each other.
  - It is important to study groups of genes and not single genes
- K-Means and Hierarchical Clustering assume that related geneses should have similar expression profiles across all samples
  - Though this assumption does not hold in all experiments
- Biclustering was proposed to overcome these limitations

Biclustering

- A Bicluster can be defined as a subset of genes that are correlated under a subset of samples
- Biclustering refers to simultaneously clustering both rows and columns of a given matrix of patterns
    - This helps in discovering local patterns that cannot be identified by the standard one-way clustering algorithms
- Biclustering has been used in several applications such as clustering microarray data, protein interactions, collaborative filtering and text mining.

Assessing "Goodness" of Clusters

- One way to assess the goodness of fit of a given clustering is to compare the size of the clusters vs. the distance to the nearest cluster



Assessing "Goodness" of Clusters

- Another possible quality indicator is the average of the distances between the members of a cluster and the center, very similar to before but slightly more robust
    - This is normally done by summing the square of the distances from every point to the center, called Total Sum of Squares
    - The total sum of squares can be taken between clusters and within clusters, and the proportion of this value can be assessed
- The diameter of the smallest sphere including all members of a given cluster may also be used as a quality assessment,
    - Though this is a sensitive measure

Confidence in Cluster Assignment

- How confident can we be that the pattern falls in a given cluster?
    - We can follow a gene through several clusterings to ensure it belongs with its group
- This can also be addressed using a bootstrapping approach, where a goodness of fit measure is based on many repeats of the same experiment on slightly different data sets all constructed on from the available data
- Essentially clustering many times and the confidence of a pattern belonging to a cluster is inversely proportional to the amount of times is moves to a different cluster

Results

Training data cluster, all features

	ALL	AML
1	0.1481481	0.7272727
2	0.8518519	0.2727273

```
> train_cluster$withinss / train_cluster$betweenss
[1] 4.390243 3.480078
```

Testing data cluster, all features

	ALL	AML
1	0.65	0.6428571
2	0.35	0.3571429

```
test_cluster$withinss / test_cluster$betweenss
] 0.9207035 1.1718051
```

Training data cluster, PCs

	ALL	AML
1	0.7037037	0.09090909
2	0.2962963	0.90909091

```
> train_cluster$withinss / train_cluster$betweenss
[1] 3.480078 4.390243
```

Testing data cluster, PCs

	ALL	AML
1	0.2	0.5714286
2	0.8	0.4285714

```
> test_cluster$withinss / test_cluster$betweenss
[1] 1.1718051 0.9207035
```

References

[1] Bignotti, E et al. “Trefoil Factor 3: A Novel Serum Marker Identified by Gene Expression Profiling in High-Grade Endometrial Carcinomas.” British Journal of Cancer 99.5 (2008): 768–773. PMC. Web. 4 Oct. 2018.

[2] Drăghici Sorin. Statistics and Data Analysis for Microarrays: Using R and Bioconductor. Chapman and Hall, 2012.