Clustering and dimensionality reduction

Clustering and clustering algorithms

Clustering is the method of establishing structure within a collection of unlabeled data. Clusters are organized according to points that are similar to each other as well as different from points in other clusters. Measuring similarity entails either plotting gene expression of each experiment or measuring correlation under different parametric (Pearson or Euclidean) or non-parametric (Spearman or Kendall) conditions. The main goal of the clustering algorithm is to, therefore, create clusters that are internally coherent while simultaneously different from each other. Commonly used clustering analysis methods include hierarchical clustering and flat clustering. Hierarchical clustering creates an informative hierarchy of clusters whereas flat clustering creates a flat group of clusters that does not possess any real structure that would relate clusters to one another. Hierarchical clustering consists of algorithms, such as single-link, complete-link, group-average, and centroid similarity, each of which differ in their similarity measures. In contrast, flat clustering comprises algorithms such as K-means in which the main purpose is to minimize the average squared Euclidean distance of certain points from their cluster centers.

Data mining in addition to statistical data analysis, such as machine learning, pattern recognition, bioinformatics, and information retrieval, are a few fields that primarily use clustering as well as its different algorithms. Transcriptomics uses clustering and clustering algorithms to specifically build groups of genes that convey coexpression in regards to related proteins for certain pathways or co-regulated genes; however, clustering and clustering algorithms have various applications in other fields as well, such as medicine, business and marketing, and computer science.

Similarity measure (in clustering)

Similarity measure quantifies the similarity between data points. In biological data sets, similarity is calculated as correlation or association, i.e. correlation between two genes in gene expression data. Similarity measures can be calculated using parametric or non-parametric tests. Common parametric tests are Pearson correlation and Euclidean distance, and common non-parametric tests are Spearman's rank correlation and Kendall rank correlation.

Similarity measure is an essential component to cluster analysis, which groups objects by their similarity or dissimilarity. For example, Javanovic et al. used the R package “seriation” to cluster gene expression into heat maps, and the package uses Euclidean distances to calculate the similarity measure to sort objects into clusters.

PCA - Principle Component Analysis

PCA is a technique used to identify strong patterns and to elucidate the underlying structure in a given dataset (usually two-dimensional, e.g. genes vs. experiments). It can also be defined as a way to cluster data. PCA is often needed when there are many variables (e.g. genes vs. experiments) in a dataset and the data cannot be analyzed easily. PCA is useful because it reduces the dimensionality of the data set but attempts to retain account for the variation in the data. It does this by identifying directions, known as principle components or eigenvectors, which, in their totality, define the variation in the data. The vectors (or principle components) are sorted according to decreasing importance with respect to how much variation they explain in the data. Therefore, by only examining the first eigenvectors, one can reduce the dimensionality of the data. The original data matrix can be unambiguously reconstructed by multiplying the eigenvector by its weight by the projection of each gene onto the eigenvector.

A good reference paper can be found here: http://www.nature.com/nbt/journal/v26/n3/full/nbt0308-303.html