CSE 527 Notes Lecture 6, 10/15/03, Eric Herbig

Cluster Validation For Gene Expression Data

-leave out one cross validation

-cluster using all genes but leave out one condition

-use left out condition to check cohesiveness of clusters

-repeat for each condition

-can be used to compare algorithms

Figure of Merit (FOM)

-general indicator of cluster quality

-small variance of left out condition within clusters---good method

-as the number of clusters increases, FOM decreases (in general)

-used to determine (ballpark figure) optimal number of clusters

-example in class where FOM levels out at 5 clusters-->use 5 clusters

Clustering Case Studies

-Yeast data

-can clustering recreate 5 phases of cell cycle determined experimentally?

-single link performed very poorly, barely better than random

-other methods performed much better, all fairly similar (average link, complete link, k-means, centroid link)

-Ovarian cancer model

-single link- performed poorly

-complete link, cast, and k-means performed well

-average link performed ok

-centroid link performed poorly

-Rat CNS data

-single link performed alright (in contrast to yeast and cancer studies)

-Barrett's data

-lowest FOM correctly grouped/separated some known marker genes

-Conclusions

- -cast and K-means produce higher quality clustering than hierarchical
- -single link worst
- -FOM methodology allows comparison of any clustering algorithms on any data set.

-Adjusted Rand

- -measurement of agreement between 2 parititions of data-->must know truth to use
- -look at pairs of genes in algorithm and the true solution-->are they the same in both?
- -best score=1, worse<1, expected score for random partition = 0
- -as adjusted rand gets better so does adjusted FOM

-Hurbert Score

- -alternative to adjusted rand

Principle Component (PC) Analysis for Clustering

- -reduce dimensionality, while retaining as much variation as possible
- -linear transformation of original variable
- -PC uncorrelated and ordered
- -1st PC-->a line on a graph that gives most variation to the data
- -2nd PC-->line perpendicular to 1st PC giving most remaining variation

-PC Analysis of Yeast Sporulation Data

- -1st PC measured total average expression
- -2 PC's accounted for 90% variability in the data
- -3 PC's accounter for 95% variability in the data
  - -determine how many PC's needed to represent given % of the data

-Pictorial of graph with 2 PC's--->delineating clusters derived from known functions

- -clusters not very obvious
- -3 PC's--->better cluster separation

-Using first few PC's to define each point approximately preserves Euclidian distances

- -but with correlation coefficient-->no general relationship before and after PCA

-largest eigen values for PC's does not always lead to best clustering

-Ovary Data

-14 PC's needed to account for 90% variation

-Synthetic data

1. compute mean vector and covariance matrix for real data; generated random data sets by sampling from Gaussian with this mean/covariance structure
2. re-sample real data independently in each column

Method 1 preserves covariance structure, but not the empirical distribution of values; method 2 does the reverse.

Study Approach

-cluster original

-cluster using PCA

-cluster using PCA with higher Rand indices

(Continued next lecture)