Class Notes (7): 20.October.2003
Divya Bhat
Clustering: PCA

Cluster Validation Wrap-up:

Cluster Validation: External Criteria
Methods:

- Compare w/ Gold Standard
  - often not available
- Uniformity of clusters with respect to external info
  - look at current categorization of the genes and see if the known ones (e.g. MIPS or Gene Ontology categories) make sense in the clusters you determined
- Quantifiable Methods: relative entropy, adjusted rand indices, etc.

Cluster Validation: Internal Criteria
validate based on compactness and separation of result

Methods:

- residual sum of squares to cluster centers vs. sum of squares between centers
- silhouette - average distance to points in same cluster vs. nearest other cluster

$$s(i) = (b(i) - a(i)) / \max(b(i) - a(i))$$

Problem: how useful is this if the original clustering algorithm chose the wrong metrics?

Cluster Validation: Model-based Validation
given statistical models, how well does the data fit?
Method: look at likelihood ratio that data could have been generated by one model vs. another

Our Methodology (in the slides presented in previous lecture):
Leave out one cross-validation: look at agreement between clusterings with all data except leaving out different conditions

Principal Component Analysis

PCA is a standard technique for statistical data analysis. The "principal components" are linear combinations of the original variables. The 1st PC is the combination that best "explains" the data, i.e. maximizes the variance of the data when projected onto that dimension; the 2nd PC is the linear combination orthogonal to the 1st PC that best explains the residual variance, etc.

Geometrically, if the data is viewed as a set of points in n-space, modeled as an ellipsoid (which would be appropriate if the data were independent sample points from an n-dimensional Gaussian distribution), then the PCs are exactly the axes of the ellipsoid; 1st PC is the longest axis of the ellipsoid, etc. It is commonly observed that the first few PCs account for a large percentage of the variation in a data set, and so representing the original points by their projections onto those axes gives a lower-dimensional, yet relatively accurate, approximation. Intuitively, the last few PCs are discarded as "noise".

It is natural to ask whether this dimension/noise reduction technique is a useful preprocessing step as a prelude to gene expression cluster analysis. The paper summarized here studies this question. Some prior work had looked at gene expression data after projection onto 1st 2 or 3 PCs and observed that clusters were not evident, but this could be because either there were no sharp clusters or because the projection was obscuring them (or a combination of both effects).

Since it's not clear how many PCs to retain, the study varied the number. The idea is to cluster with the first *m* Principal Components, with m varying from 1 to the dimensionality of the data set. For comparison, we also looked at clustering after projection onto:

- m randomly selected orthogonal axes. There is theoretical support that randomly projecting data into lower dimensions often makes structures more compact, which is good, although this turned out to be ineffective in this case.
- the "best" m PCs, not necessarily the first m. It's too expensive computationally to try all possible subsets of m PCs, so we used a greedy heuristic to search for good subsets. "Quality" of a subset was evaluated by adjusted Rand index. This is *not* a criterion that could be directly applied in practice, but the goal was to test the "conventional wisdom" that the first few PCs contained the most useful information.

Conclusions:

PCA sometimes helps but sometimes hurts. It is not true that the first few PCs are always the best to use. PCA generally was less beneficial with good clustering algorithms.

Bottom Line: Successes by other groups make it a technique worth considering,but it should not be applied blindly.

Linear Discrimination (Fischer Linear Discriminate Analysis)

supervised learning – train a classifier on examples with known, trusted classification, then use it to classifying additional unknown samples into one of two categories

method: draw a line (or hyper plane) between categories. The line is determined by finding the Gaussian model of each category (the same model for each category, except with different mean vectors), and determining the line that separates genes more likely belonging in one category than the other.

This method is good (provably optimal according to certain criteria) if the data satisfy the assumptions (the data is Gaussian in nature).

It is unclear whether the Gaussian model applies to gene expression (but we'll see some data on that in a week or so).

Other forms of Dimensionality Reduction:

Multidimensional Scaling
Independent Component Analysis