Clustering and Combinatorial Chemistry

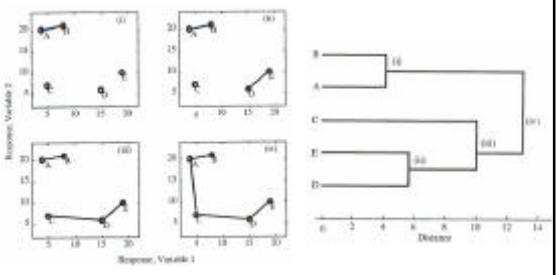
Pattern Recognition Types

- Unsupervised
  - Class memberships are not imposed, but are explored through the pattern recognition process
  - HCA, PCA
- Supervised
  - Class memberships are imposed based on knowledge about the system
  - A training set with known class memberships is required
  - KNN, SIMCA

Hierarchical Cluster Analysis

- Pattern recognition methods (and clustering) require some measure of similarity or distance
- Example
  - Euclidean distance = $\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2}$
    (can be extended to cover any number of descriptors)
- HCA
  - Starts with single-sample clusters
  - Joins nearest clusters iteratively

Hierarchical Cluster Analysis



Linkage Methods

- Single
  - Distance between clusters is the distance between the two closest members of the clusters
- Complete
  - Distance between clusters is the distance between the two farthest members of the clusters
- Average
  - Distance between clusters is the average over all pairs of members in the clusters

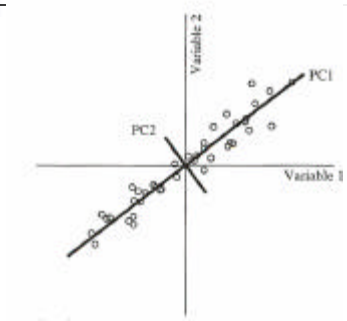
HCA Strengths/Weaknesses

- Strengths
  - All variation in the data is represented
  - Good for reducing data of high dimensionality
  - Presentation is simple and standardized
- Weaknesses
  - Variables contributing most to the clustering are not apparent
  - Some of the distance data is lost
  - Doesn't assist in human interpretation

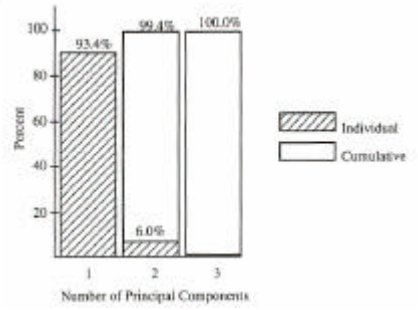
Principal Components Analysis

- Principal components analysis is a variable reduction method (an alteration of the coordinate system) – allowing visual analysis of multi-dimensional data in fewer dimensions
- The first principal component explains the maximum amount of variation possible in the data set in one direction – the % of variation explained can be precisely calculated

PCA – Example



How Many Components (Rank)?



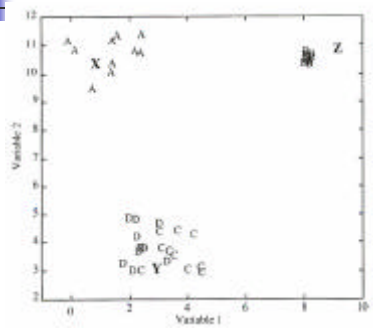
PCA Strengths/Weaknesses

- Strengths
  - Displays highly dimensional data with relatively few plots
  - Can filter noise from data sets
  - Can determine amount of variation contained in each descriptor (loading)
- Weaknesses
  - Inherent dimensionality (rank) must be determined
  - If the dimensionality is greater than three, visualization is still difficult

K-Nearest Neighbor

- A supervised pattern recognition technique
- Class membership of unknowns is assigned on the basis of the K nearest samples
- An optimal value of K is determined through cross-validation

KNN Example



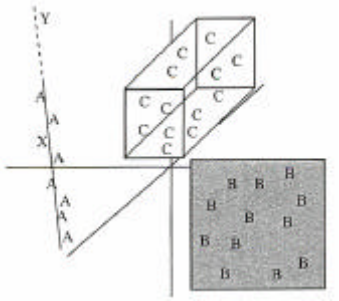
KNN Strengths/Weaknesses

- Strengths
  - Simple to implement
  - Can be used with few examples per class
  - Additions to training set are easily accomodated
- Weaknesses
  - No outlier detection (a classification is always made)
  - Class shape information is not utilized

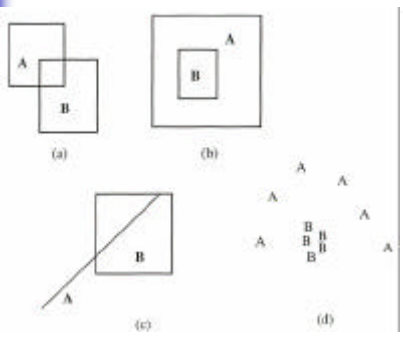
SIMCA

- Soft Independent Modeling of Class Analogies (SIMCA)
- Applies PCA to model shape and position of the object formed by samples for class definition
- Unknown objects are predicted on the basis of their inclusion in a classes object

SIMCA Example



Class Overlap



SIMCA Strengths/Weaknesses

- Strengths
  - Can detect if a sample does not have membership in any class of the training set
  - Both shape and position of the class object are used (not required to be linear)
  - Signal averaging can occur as PCA is used
- Weaknesses
  - Many samples per class are required for accurate object shape and size
  - Statistical classification criteria aren't always applicable

Why Cluster?

- Conformational analysis of relatively large structures provides many conformations, many of these fall in families having similar shape
- De novo design produces a wide variety of structures, might want to synthesize one representative from each cluster
- Combinatorial library design – can select representative examples from different clusters for synthesis

Combinatorial Library Design

- Consideration 1: Library Purpose
  - Focused: working with 1 biomolecular target
    - Probably will use existing leads to help design library
  - General: working with many biomolecular targets
    - Probably want a very diverse group of structures
- Consideration 2: Synthetic Strategies
  - Library design must take into account the chemistry that can be used to enumerate the library as well as the available starting materials

What is diversity?

- Should be measured by the same criteria that determine biological activity:
  - Pharmacodynamic considerations
    - Shape
    - Size
    - Electron Distribution
  - Pharmacokinetic considerations
    - Hydrophobicity
    - Size
    - Metabolism
- Many of these are reflected in standard QSAR descriptors

Class Exercise

- Consider how some of your modeling results obtained so far could benefit from clustering
- Cluster your data as you determined above

OR

- Perform a conformational search on a relatively large compound
- Calculate descriptors that characterize shape and/or electronics of the conformations
- Cluster your conformational search results

Class Exercise, cont'd

- What can you learn from your clustering?
  - How do the number of clusters compare to the number of data points?
  - Do members of the cluster seem to have anything in common to you?

Reading

- First Edition: Section 8.11-8.12
- Second Edition: 9.13-9.14, 12.14