Cluster Validation

Psych 993
Methods for Clustering and Classification
Lecture 2
Today’s Lecture

• Topic Assignment.

• Validation of clustering results.

• Describing results.

• Discussion of Chapter 7 of Gordon (1999).
Topics to be Assigned

• Discriminant Analysis

• Hierarchical Clustering Methods
Topics to be Assigned

• K-means clustering algorithms

• Latent Class Analysis
Topics to be Assigned

- Latent Profile Analysis

- Finite Mixture Models
Topics to be Assigned

- Growth Mixture Models

- Cognitive Diagnosis Models
  TBA
Clustering Techniques

• Largely seen as exploratory analyses of data structure.

• *Post hoc* evaluations of relative must be taken as is…
  – Often times clustering results are inappropriate.
  – There is a great need to validate your final solutions.
Validation Methods

• Gordon discusses several general methods to validate cluster results.
  – Not all methods in his chapter will be applicable to all clustering methods.

• A general, flexible way to get a crude estimate of validity is to look at the stability of the result.
  – Dividing the data and running the same method (cross-validation).
  – Multiple analyses with differing clustering methods.
  – Changing the metric of the dissimilarity.
  – Changing the criteria of the clustering method.
Subsets of the Data

• General approach cited by Gordon:
  1. Divide data into two subsets: A and B.
  2. Apply clustering algorithm to A – get $c$ classes.
  3. Each object in B is assigned to “closest” class in A.
  4. Apply clustering algorithm to B – get $c$ classes.
  5. Compare partitions of B (based on #3 and #4).

  – If agreement is high, have high confidence in result.
General Pattern in Classification Studies

• In statistics, exploratory analyses are often used to formulate models.
  – Such models are then used for subsequent confirmatory studies.

• Part of the problem with clustering is that often research is not concerned with generality.
  – Only concerned with objects in sample.
Types of Validation

• Gordon cites Jain and Dubes (1988, Ch. 4) as defining three types of cluster validation:
  – External tests
    • Comparing classification with information not used to create classification.
  – Internal tests
    • Comparing parts of the classification with the original data.
  – Relative tests
    • Compare several different classifications of the same set of objects.
Types of Tests of Structure

• Gordon describes tests for:
  1. Complete absence of class structure.
  2. Validity of an individual cluster.
  3. Validity of a partition.
  4. Validity of hierarchical partition.
Null Models

- Poisson model
- Unimodal model
- Random permutation model
- Random dissimilarity matrix model
- Random labels model

- Note that the last three are permutation based methods – combinatorial data analysis techniques.
Tests of the Absence of Class Structure

• Such tests use null models as a comparison to final solution.
• Typically such tests are not used in research for reasons such as:
  – Confidence in data containing distinct classes.
  – Interest in solely obtaining a dissection of the data set
  – Intend subsequently to validate the classification that is obtained, and to realize that a two-stage testing procedure would complicate evaluation of the significance level of any test.
Assessing Individual Clusters

- One way has been to specify what an ideal “valid” cluster resembles.
- More widely applicable methods involves the definition of an index of cluster adequacy.
  - Provides likelihood of such index values under null model.
Cluster Validity Profiles

• Create probabilities for cluster membership based on hypergeometric distribution of objects.
• The hypergeometric distribution arises when a random selection (without repetition) is made among objects of two distinct types.
  – Here our two distinct types are
    • Objects within a similar cluster – the Between
    • Objects not in a similar cluster – the Within
• Complicated and difficult to use.
Monte Carlo Validation

- General approach is to simulate data under a null model hypothesis.
- Once data are simulated, clusters are formed.
- Again, this is more of a specific test of validity.
Assessing Partitions

• Questions in the assessment of partitions:
  1. Is there a close correspondence between two independently-derived partitions of the same set of objects?
  2. Which of a set of partitions agrees best with an externally-provided partition?
  3. Does a specified partition into \( c \) (say) clusters comprise compact and isolated clusters?
  4. When a clustering procedure provides partitions of data into \( c \) clusters for several different values of \( c \), which is the most appropriate partition?
  5. Does a partition into \( c \) clusters obtained from the output of a clustering procedure comprise compact and isolated clusters?
Cluster Validation Statistics

- Main cluster validation statistics involve thinking about the possible results of two “clustering procedures”
  - One procedure may be the “truth”
- One frequently used statistic is the Rand Statistic (1971).
  - This statistic has been modified by Hubert and Arabie (1985).
Assessing Hierarchical Classifications

Several questions posed by Gordon:

1. Is there a close correspondence between two independently-derived hierarchical classifications of the same set of objects?

2. Does a specified hierarchical classification provide an accurate summary of the relationships within a set of objects?

3. Does a hierarchical classification obtained from the application of a clustering procedure to a set of objects provide an accurate summary of the data?
Cluster Description

• Measures of dissimilarity can be used to describe clusters.
  – I feel these do not adequately tell the picture to the substantive researcher.

• The chapter describes very specific methods that are not applicable in all clustering situations.