The Cluster Sensitivity Index (CSI)

A Qualifier for Peer Groupings

November 2006 @ CAIR

Willard Hom, Director
Research & Planning, Chancellor’s Office,
California Community Colleges

Preface

- Comments about this topic are welcome so that we can improve our work.
- A paper for publication, on this topic, is forthcoming.

Objectives of This Talk

- Propose the CSI as a diagnostic tool for cluster analyses, esp. in peer grouping.
- Propose the weighted peer group mean as a remedy for certain problems that occasionally arise in cluster analyses.

CSI

- The Cluster Sensitivity Index is a proposed measure to help analysts understand the usability of cluster analyses for decision-making.
- This is a “work in progress.”

The Need for the CSI

- Analysts use peer groups for evaluating institutional situations.
- Cluster analysis is often the tool of choice for defining a peer group.
- Cluster analysis has a “method bias” that can affect peer group definitions.
- We could use a tool to detect this method bias (or sensitivity to choice of computation).

Uses of Peer Grouping

- Higher education.
- California K-12 system.
- Medical care.
- Businesses involved in benchmarking

Sources of Method Bias in Cluster Analysis

- Proximity measure
  - Distance (i.e., Euclidean, etc.)
  - Similarity
- Clustering Algorithm (some examples below)
  - Single Linkage
  - Average Linkage
  - Ward’s
  - Other algorithms

An Example

- The next slide shows an excerpt of a cluster analysis to find the peer group for a specific college (Palomar, by chance).
- We ran three different cluster analyses and found three different peer group definitions for Palomar. The methods were (1)Avg Linkage w/Euclidean distance; (2) Ward's w/ Euclidean distance; and Ward's w/Minkowski distance.

Institution	AverageLinkage Method	Ward's Method	Ward's Method II
Palomar	X	X	X
American River	X	X	X
Sacramento City	X	X	
Santa Rosa	X	X	X
Diablo Valley	X	X	X
San Francisco	X	X	X
De Anza	X	X	X
Moorpark	X		
El Camino	X	X	
East L.A.	X	X	
Pasadena	X	X	X
Santa Monica	X	X	X
Long Beach	X	X	
Mt. San Antonio	X	X	X
Saddleback	X	X	X
Riverside	X	X	X
Count per method	16	15	11

Doing the CSI

- Find the smallest peer group for Palomar

  This is from Ward’s Method II.
- Find the number of additional institutions that the other two methods defined as peers to Palomar.

  These are Long Beach, East L.A., El Camino, Sacramento, and Moorpark (5 in count).

Doing the CSI, part 2

- Find the number of colleges that the alternate methods (Avg.Linkage & Ward's) could have defined as peers.
  108 – 11 = 97
- Divide the count of "newly" found colleges by the count of potential peers or 5/97.
  The CSI for Palomar = .052

What Does This CSI Mean?

- The peers defined for Palomar are relatively stable, regardless of which clustering method the analyst may use.
- The mean of this peer group could be a frame of reference for Palomar, with some standard precautions.

Interpreting the CSI

- The CSI can range from zero to one.
- The higher the CSI, the more uncertainty there is for the definition of peer members based upon one clustering method.
- Personal levels of risk aversion and future empirical research would indicate what a given level of CSI indicates to the analyst.

What to Do With a High CSI

- Check your data and data processing/clustering process for anomalies.
- Warn audiences that the cluster results for a given institution are tenuous.
- Produce a summary statistic for the institution’s peer group that adjusts for the “fuzziness” of its cluster results.

Weighted Peer Group Mean

- Adjusts the peer group mean for the partial “membership” (fuzzy membership) of some institutions.
- Accounts for the frequency that an institution is defined as a peer.

Example of Weighted Peer Group Mean

- For the Palomar peer group example, let’s compute this figure for the variable of college age (years since the college was started).

Institution	Years of Age	AverageLinkage Method	Ward’s Method	Ward’s Method II
Palomar	60	X	X	X
American River	51	X	X	X
Sacramento City	90	X	X	
Santa Rosa	88	X	X	X
Diablo Valley	57	X	X	X
San Francisco	71	X	X	X
De Anza	39	X	X	X
Moorpark	39	X		
El Camino	60	X	X	
East L.A.	61	X	X	
Pasadena	82	X	X	X
Santa Monica	77	X	X	X
Long Beach	79	X	X	
Mt. San Antonio	60	X	X	X
Saddleback	38	X	X	X
Riverside	90	X	X	X
Mean Age by Method	65.1	65.1	66.9	64.8

Institution	Years of Age	Weight*	AverageLinkage Method	Ward's Method	Ward's Method II
Palomar	60	3	X	X	X
American River	51	3	X	X	X
Sacramento City	90	2	X	X	
Santa Rosa	88	3	X	X	X
Diablo Valley	57	3	X	X	X
San Francisco	71	3	X	X	X
De Anza	39	3	X	X	X
Moorpark	39	1	X		
El Camino	60	2	X	X	
East L.A.	61	2	X	X	
Pasadena	82	3	X	X	X
Santa Monica	77	3	X	X	X
Long Beach	79	2	X	X	
Mt. San Antonio	60	3	X	X	X
Saddleback	38	3	X	X	X
Riverside	90	3	X	X	X
Mean Age by Method	65.1		65.1	66.9	64.8

* Weight is the number of times that the institution was defined as a peer for Palomar.

Institution	Years of	Weight*	Yrs x Wt
Palomar	60	3	180
American River	51	3	153
Sacramento City	90	2	180
Santa Rosa	88	3	264
Diablo Valley	57	3	171
San Francisco	71	3	213
De Anza	39	3	117
Moorpark	39	1	39
El Camino	60	2	120
East L.A.	61	2	122
Pasadena	82	3	246
Santa Monica	77	3	231
Long Beach	79	2	158
Mt. San Antonio	60	3	180
Saddleback	38	3	114
Riverside	90	3	270
WPGM =	65.7	42	2758

Applications of CSI

- Use when a college needs to know if a peer grouping from a cluster analysis is sensitive to the method used (i.e., “method bias”).
- Use if a college has access to the data to run alternate clusterings with different cluster methods. (Or have the data owners provide the alternate outcomes.)

Some Major Assumptions of CSI

- Cluster analysis (and many classification methods) will find different peer institutions for a college if we vary the methods used.
- Peer membership can be a “fuzzy” state.
- The analyst lacks information about the true clusters in the set of institutions.
- The variables used in the cluster analysis are relevant to the objective and contain valid and reliable data.

More Major Assumptions

- The different methods of clustering or classification provide equally valid peer results. (But a random selection of methods could help in the use of the CSI.)
- The population to be peer grouped is relatively small.
- The primary objective is the variability of peer grouping for a specific college, not the validation of all peer groups.

Summary

- The CSI is a tool for evaluating method bias in the classification of a given set of data (about institutions or any entities).
- If the CSI causes you concern, you can use the weighted peer group mean as one remedy.
- The CSI can apply to any classification effort (not just cluster analysis) and to any kind of population (not just institutions).

Contact Info

- Willard Hom, Director
  Research & Planning Unit
  Chancellor's Office,
  California Community Colleges

  whom@cccco.edu
  (916) 327-5887