CCSU
DEPARTMENT OF MATHEMATICAL SCIENCES

COLLOQUIUM

Friday, April 25
11:15 – 11:45 AM
Davidson Hall, Room 207

ASSESSMENT OF SIMILARITY-BASED CLUSTER EVALUATION METHODS

JILL WILLIE

(Data Mining MS Thesis Presentation)

CENTRAL CONNECTICUT STATE UNIVERSITY

Abstract: We employ a method of repeatedly sampling the available data, fitting a cluster algorithm, and observing the similarity of the resulting cluster schemes. We apply two clustering algorithms, k-means and hierarchical clustering. As a measure of how well we might expect the clusters to hold over new data we measure the similarity of the cluster results over random samples of the full dataset to each other. While in practice it is rare to have a known, ideal cluster structure, we have synthetic data. Therefore we also measure how accurately the resulting clusters represent a known ideal structure in the data to gain insight. We focus on discrete, binary data as one might encounter when modeling keywords from text.

To measure cluster scheme similarity we employ the Rand Index, the Jaccard Index, and Adjusted Rand Index. The computation of each of these indexes is described in detail. The applicability and interpretation of each index is discussed. Results are examined using both synthetic data and real data and compared to two internal measures of cluster quality, the Silhouette and pseud-F statistic. We manufacture various data with defined cluster structures, and compare the performance of each index on these data as well as on data without any cluster structure. We also apply this technique against text data to illustrate one potential application.

We conclude that examining the patterns in similarity among repeated sample outcomes provides valuable and complementary information to that gained by examining cluster quality in terms of compactness and dispersion.

For further information:
gotchevi@ccsu.edu 860-832-2839
http://www.math.ccsu.edu/gotchev/colloquium/