CCSU
DEPARTMENT OF MATHEMATICAL SCIENCES

COLLOQUIUM

Friday, December 5
9:30 – 10:00 AM
Davidson Hall, Room 207

IMPLEMENTATION OF STABILITY-BASED CLUSTER VALIDATION USING PREDICTION STRENGTH

ERIC FLORES-ACOSTA

(Data Mining MS Thesis Presentation)

CENTRAL CONNECTICUT STATE UNIVERSITY

Abstract: Clustering validation metrics are some of the important indicators of the data miner toolset. In this work we explore three different formulations from the family of stability-based clustering validation metrics. It is demonstrated that these three approaches share a structure of four stages, differing mainly in the way that these four stages are implemented. As a proof of concept, a new stability-based clustering validation framework is developed with a Python class mapping these four stages as four Python methods. This framework was tested by using it to implement Tibshirani & Walther (2005) Prediction Strength clustering validation metric and then challenged by comparing results obtained against those resulting from other existing implementations of the Prediction Strength metric. The new program was also used to explore the correct number of clusters on a data set with low dimensionality and well separated clusters and a data set with higher dimensionality and moderately separated clusters. The number of clusters indicated by the new program was consistent with the number of clusters suggested by the Silhouette coefficient and IBM PASW Modeler when using the BIRCH clustering algorithm. In addition, the clusters discovered where relevant as a way of characterizing the behavior of the two most important subgroups of the data set.

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