CCSU

DEPARTMENT OF MATHEMATICAL SCIENCES

COLLOQUIUM

Friday, December 4
9:30 – 10:15 AM
Davidson Hall, Room 204

PERFORMANCE OF VARIOUS BASE AND ENSEMBLE CLUSTERING APPROACHES

TAREK ABDEL-AZIM

(Data Mining MS Thesis Presentation)

CENTRAL CONNECTICUT STATE UNIVERSITY

Abstract: Standard partitioning and hierarchical clustering methods have strengths and weaknesses in their ability to find structure in data. For example, the popular K-means algorithm suffers from a number of shortcomings: it is sensitive to data order, initialization points, and noise. K-means also performs poorly when clusters are neither convex nor well-separated.

In this thesis, the performance of two alternative base or individual partitioning clustering algorithms, DBSCAN and Mixtures Models, is assessed alongside K-means on 17 difficult or challenging datasets.

In addition, three ensemble clustering methods – (a) ensemble clustering after Leisch (1999), (b) ensemble clustering after Dudoit and Fridlyand (2002), and (c) a Voting Merging Algorithm after Dimitriadou, Weingessel, and Hornik (2002) – are evaluated in terms of their ability to improve the performance of the K-means algorithm on the same 17 datasets. All base and ensemble algorithms are assessed with both the Rand and Jaccard indices.

Key findings include:

1. Implementation of multiple starts (multiple random initialization points) with K-means notably improves performance vs. the more typical implementation with a single, random set of initialization points.
2. DBSCAN was the method of choice for datasets that are not convex; however, DBSCAN requires two a priori user inputs to which the algorithm is very sensitive.
3. Mixture Models performed competitively (i.e., was either the top performer or close to it) on 14 out of 17 datasets.
4. Of the ensemble methods, only Leisch's bagged clustering materially improves the performance of K-means on the datasets evaluated.

No one clustering methodology performs the best on all types of datasets. Variable selection, data visualization, and heuristics support the selection of the right clustering method for a given dataset.