f

Center for Cybercrime Studies

presents

A Spectral-based Clustering Algorithm for Categorical Data Using Data Summaries with an application in criminal justice data: NIBRS

Dr. Eman Abdu

Mathematics & Computer Science Department
John Jay College of Criminal Justice

Date: Tuesday, April 25th, 2017
Time: 10:00 – 11:00 am
Math & CS Conference Room (6.63.37), 6th Floor, New Building

Abstract:

Abstract: In this talk, I present a novel spectral-based algorithm for clustering categorical data that combines attribute relationship and dimension reduction techniques found in Principal Component Analysis (PCA) and Latent Semantic Indexing (LSI). The algorithm uses data summaries that consist of attribute occurrence and co-occurrence frequencies to create a set of vectors each of which represents a cluster. We refer to these vectors as "candidate cluster representatives." The algorithm also uses spectral decomposition of the data summaries matrix to project and cluster the data objects in a reduced space. I refer to the algorithm as SCCADDS (Spectral-based Clustering algorithm for Categorical Data using Data Summaries). SCCADDS differs from other spectral clustering algorithms in several key respects. First, the algorithm uses the attribute categories similarity matrix instead of the data object similarity matrix (as is the case with most spectral algorithms that find the normalized cut of a graph of nodes of data objects). SCCADDS scales well for large datasets since in most categorical clustering applications the number of attribute categories is small relative to the number of data objects. Second, non-recursive spectral-based clustering algorithms typically require K-means or some other iterative clustering method after the data objects have been projected into a reduced space. SCCADDS clusters the data objects directly by comparing them to candidate cluster representatives without the need for an iterative clustering method. Third, unlike standard spectral-based algorithms, the complexity of SCCADDS is linear in terms of the number of data objects. Results on datasets widely used to test categorical clustering algorithms show that SCCADDS produces clusters that are consistent with those produced by existing

algorithms, while avoiding the computation of the spectra of large matrices and problems inherent in methods that employ the K-means type algorithms. The algorithm can also be used to cluster data associated with criminal incidents (NIBRS) to discover victim profiles or crime patterns