Association Mining via Co-clustering of Sparse Matrices

Brian Thompson, Rutgers University (bthom@cs.rutgers.edu) – Short Talk Proposal
Joint work with Linda Ness (ACS), David Shallcross (ACS), and Devasis Bassu (ACS)

Consider a matrix M representing a relation between two sets of objects, where matrix entry $M_{i,j}$ indicates the existence or strength of an association between objects i and j. Identifying *biclusters* (the intersection of a subset of the rows with a subset of the columns) with many strong pairwise associations has applications to ontology discovery in relational databases, analysis of gene expression data, text mining, collaborative filtering, and many other data mining tasks. In particular, we are interested in applications where the matrix is extremely sparse, and the goal is to efficiently find dense biclusters, if they exist.

There has been a significant amount of related work in both the computer science and bioinformatics literature. However, most previous approaches either are not effective for very sparse matrices, do not scale well, or require parameters that in practice may not be known a priori. We propose a new approach that attempts to address these concerns.

Co-clustering is the data mining task of simultaneously clustering the rows and columns of a matrix. Our approach consists of two main components: (1) Define a quality metric for co-clustered matrices; and (2) Find a co-clustering that maximizes the value of the metric. This poses a computational challenge because the number of possible co-clusterings is exponential in the size of the matrix.

We first suggest several metrics that are designed to reward co-clusterings with large, dense biclusters. We then present the CC-MACS (Co-Clustering via Maximal Anti-Chain Search) algorithm, which leverages a property of the proposed metrics in order to efficiently search the space of possible co-clusterings. We demonstrate the effectiveness of the CC-MACS algorithm by comparing it with related work and baseline algorithms, using both synthetic matrices and real-world data taken from the domains of finite element modeling and quantum chemistry.