MINING SUBSPACE AND BOOLEAN PATTERNS FROM DATA

By

Lizhuang Zhao

An Abstract of a Thesis Submitted to the Graduate

Faculty of Rensselaer Polytechnic Institute

in Partial Fulfillment of the

Requirements for the Degree of

DOCTOR OF PHILOSOPHY

Major Subject: Computer Science

The original of the complete thesis is on file
in the Rensselaer Polytechnic Institute Library

Examining Committee:

Mohammed Javeed Zaki, Thesis Adviser

Chris Bystroff, Member

David Musser, Member

Mukkai Krishnamoorthy, Member

Malik Magdon-Ismail, Member

Rensselaer Polytechnic Institute
Troy, New York

September 2006
(For Graduation December 2006)

ABSTRACT

In this thesis, we set up a systematic framework for mining subspace and boolean patterns from data. Subspace patterns are extracted from real-valued datasets, whereas boolean patterns are mined from binary-valued datasets. For real-valued datasets, we mainly consider scaling and shifting relationships, whereas for binary-valued datasets, we mine arbitrary boolean expressions (OR, AND, CNF, DNF). Boolean patterns are also used to mine redescriptions, i.e., to describe the same group of objects in multiple "orthogonal ways".

The subspace pattern mining has been tailored to gene microarray data clustering to find biclusters and triclusters. We present two novel deterministic clustering algorithms: MICROCLUSTER and TRICLUSTER (the first 3D microarray subspace clustering), which can mine arbitrarily positioned and overlapping biclusters/triclusters. Depending on different parameter values, they can mine different types of clusters, including those with constant or similar row/column values, as well as scaling and shifting expression patterns. Optionally, the two algorithms merge/prune some clusters having large overlaps. We also give a useful set of metrics to evaluate the clustering quality, and show their effectiveness on real microarray expression data.

We also present a comprehensive framework, BLOSOM, for mining boolean patterns from binary-valued datasets. This framework forms the algorithmic core of a redescription mining solution. We organize the space of boolean expressions into four categories; pure conjunctions, pure disjunctions, conjunction of disjunctions, and disjunction of conjunctions. For each category, we propose a *closure* operator that naturally leads to the concept of a *closed* boolean expression. Further, the closed generators form a lossless representation of all possible boolean expressions and, hence, all possible redescriptions. BLOSOM efficiently mines several forms of frequent boolean expressions by utilizing a number of methodical pruning techniques.

The future work is to extend subspace pattern mining to more general notions of linear relationships, combining scaling and shifting patterns. For boolean pattern mining, we plan to apply them to the gene ontology (GO) to discover cross relationships between different GO hierarchies.