DISSERTATION PROPOSAL

Amin Hosseininasab

"Mining Knowledge from Sequential Data: Algorithms and Applications"

Monday, November 26, 2018
3:00 pm
Tepper Quad 2118

Modern business management is increasingly more reliant on data. This has led to an exponential growth in data gathering for almost all organizations and disciplines, ranging from fields of marketing and finance to manufacturing, operations and healthcare. With the emergence of *big data*, comes the task of data analysis and knowledge discovery. The process of knowledge discovery involves selection, cleaning, and modeling of the database, using data mining to find relevant patterns, and interpreting the mined patterns for descriptive, predictive and prescriptive analytics. My research focuses on the last three steps of the knowledge discovery process, namely modeling data, developing data mining algorithms, and pattern interpretation.

In the first chapter, we investigate the Multiple Sequence Alignment (MSA) problem. The aim is to identify patterns of similarity between sequences of data. In the first step, we use Multivalued Decision Diagrams (MDD) for a novel representation of all pairwise sequence alignments (PSA). PSA MDDs are then synchronized using side constraints to model the MSA problem as a Mixed-Integer Program (MIP). A logic-based Benders decomposition algorithm is developed to solve the resulting MIP. Numerical results on benchmark instances show that the algorithm either solves considered instances to optimality, or significantly improves the accuracy of alignment compared to state-of-the-art heuristic MSA solvers, and the best exact approach in the literature.

In the second chapter, we consider the Constrained Sequential Pattern Mining (CSPM) problem. CSPM aims at identifying frequent patterns on a sequential database of items while observing constraints defined over the item attributes. We introduce novel techniques for CSPM that rely on an MDD representation of the database. Specifically, our representation can accommodate multiple item attributes and various constraint types, including a number of complex constraints for the first time. Results show that our approach is competitive with or superior to existing methods in terms of scalability and efficiency.

The third chapter is motivated by the methodology developed in Chapter two, which is used to mine knowledge from applications in marketing, finance and healthcare. We first tailor the mining algorithm to model data from these applications, and plan to develop more sophisticated criteria to mine relevant and insightful patterns. We aim to show the relevancy of pattern mining in different applications, and how current prescriptive algorithms may benefit from the mined insights.

In the last chapter, we aim to study the problem of choosing in-network physicians for a major insurance company. The problem involves using data mining and optimization techniques to determine the best physicians for the insurance network with the aim of reducing cost, while considering competition, customer retention, and various performance, demand, and capacity constraints.