Data Mining and Machine Learning in Education with Focus in Undergraduate CS Student Success

Ph.D. Candidate: William Gregory Johnson, Advisor: Dr. Anu G. Bourgeois
Department of Computer Science, Georgia State University, Atlanta, GA 30303

Goals and Methods

- Analyze CS student data that follows from a prerequisite of either MATH2420 or CSC2510 and:
  - Compare effective results on remaining core CS courses (CS grades, course failures, population behaviors, transfer versus native)
  - Compare performance and job placement resulting from fulfillment of the prerequisite course (IT vs CS)
  - Currently under research with IRB-H18480

- Represent the CS student in/with graphical model(s) and use social network analysis of the data (clickstream) from our LMS and perhaps GSU network to discover relationships based on graph data mining and theory

- Compare transfer vs native CS student data in CS course loads per semester within course levels and student classifications

- Develop a position paper on transfer CS students addressing:
  - CS student’s desire for lower cost of core courses at a 2-year college
  - Conflict of 2-year college’s desire for their students to complete an AS degree, resulting in more classes at 2-year, leaving mostly CS courses with a 4-year university
  - CS student's desire to finish a BS degree within three years, post AS degree, with CS course overloading (>= 3 per semester) resulting in negative performance impacts

Motivations

- Preliminary data analysis shows an increase in population of transfer CS students and a trend to enroll in more CS courses per semester than native
- Analysis also shows transfer student performance is consistently different, namely higher CS course fail rates and lower CS GPA scores
- Use graph data mining to build CS student communities based on relationships of behavioral outcome strategies between student and faculty, cognitive and non-cognitive gains, and LMS and GSU network analysis
- Develop experimental models of predictive course sequencing, adaptive and targeted tutoring, and a student facing learning analytics reporting system to better understand the CS student perceptions and perceived effects of the models

Contributions

- A framework model for a student and faculty facing learning analytics reporting system to deliver advanced analysis of evidence data, LMS interactions, and student performance outcomes
- Community feature predictors of CS students through analysis of LMS graph data with a goal to improve and deepen the usage of our LMS in the undergraduate CS curriculum
- A common coding scheme or grammar for evidence based education qualitative data for building new analysis and measurement instruments for use in other undergraduate STEM programs for education science research

Thesis / Problem

- Higher enrollment in colleges and universities in computer science (CS) programs result from an increasing demand of CS graduates
- Knowledge discovery using EDM from datasets of CS student learning and behavior, demographics, academic advising, CS courses, and social network analysis
- Use data mining and machine learning to produce evidence based data, where student prediction models can be made, tested, and improved through empirical research
- Identify communities within the evidence based data to reveal predictors of achievement, identification of skill deficits, and realize targeted intervention modalities