Building a Course Recommendation System (using data mining techniques)

Adhish Adhikari

Introduction

- Data mining techniques have been used by companies like Amazon and Netflix to provide recommendations for what to buy or what TV shows to watch.
- Huge corpus of consumer data is analyzed in order to generate patterns that allow these systems to recommend products.
- Recently, studies have been dedicated to applying data mining techniques in other fields, like education or health, but are still largely unexplored.
- I propose building a course recommendation system that uses unsupervised data mining techniques to recommend courses to students based on courses taken by past students in a similar field of study.

Background

Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics and database systems.

Two common data mining techniques have been explored in my study:

- Association Rule Learning (ARL) can be used to discover interesting relations between variables in large databases. Apriori algorithm is the most widely used Association Rule Learning algorithms. This algorithm identifies characteristics such as support (popularity of item-set), confidence (likelihood of association) and lift (likelihood of association taking support as constant) from the training data set to create association rules which can be then used for predictions.

- Sequence Mining (SM) is similar to ARL in that it is also used to discover a set of patterns shared among objects. However, this technique only analyzes objects which have between them a specific order. The SPADE algorithm, a SM implementation, finds frequent sequences efficiently using intersections on id-lists on a vertical database.

Data/Methods

The classifier, which is an implementation of Apriori algorithm and SPADE algorithm, gets training data from a database that is then processed to extract prominent patterns. I initially trained the model with 130 training item-sets (divided into two sets for each semester) of courses (90% of total dataset) taken by students at the Georgia Tech University. The rules are then filtered to get rid of redundancies and bias (preferences of students that deviate from a 'normal' course path due to biases about the professor, timing of classes, etc.). A final set of rules are then created and sent to a server that compares the rules with the user input data. Based on the semester and major of the student, courses are then recommended to students based on similar patterns identified in the rules. The diagram below summarizes the process.

Results

- Accuracy of recommendations varies from 53% - 73% depending on the number of courses entered. Accuracy increases as the number of input courses increases.
- There is also a slight increase in overall accuracy with rules generated by incorporating bias as opposed to without bias.

Discussion & Further research

- The results suggest modest accuracy but the method of measuring accuracy far from the best and very lenient.
- One of the major improvements should be a larger set of data. The data used for this study was simply too small to get any significant results.
- The biases play a huge role in changing the directions of identified patterns. More in-depth analysis of biases should be done in order to get any reasonable results.

References


Acknowledgements

This project was conducted as part of the Computer Science senior capstone, Spring 2018 under the supervision of Charlie Peck, David Barbella and Ajit Chavan [Computer Science department of Earlham College].