Building a Course Recommendation System (using data mining techniques)

Adhish Adhikari

Earlham COLLEGE

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 (likeliness of association) and lift (likeliness of association taking support as constant) from the training data set to create association rules which can be then used for predictions.

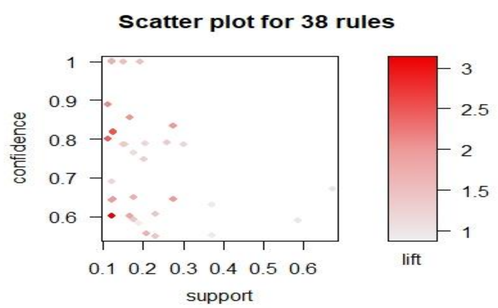

Figure 1: Scatterplots are used to determine the minimum support and confidence for the training set

- 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.

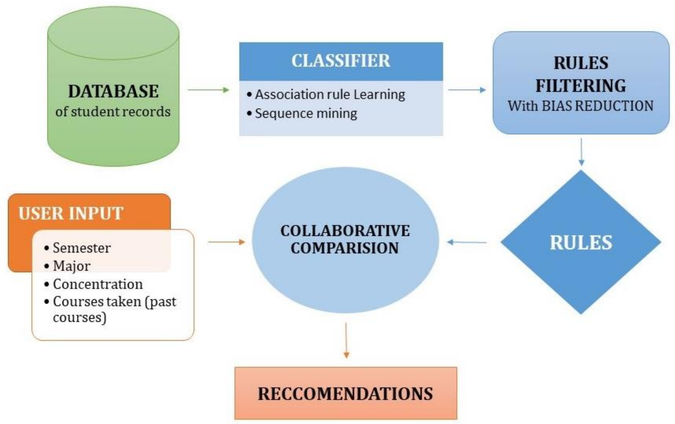

Figure 2: Flowchart describing the Course recommender architecture

- The actual recommendations made are based on the Apriori implementation (using R programming language) while sequences generated from the SPADE implementation are only used to make background tests.
- The user interface is designed to be a web app that prompts the user to enter input information (given in the flowchart above) and outputs a list of recommended courses as well as a set of all possible courses to take.

Testing - The testing is done using the 10% test data that we extracted out of the overall dataset. The test is designed as follows:

- Step 1 – The user inputs 3 of the 4 total courses in a single test item-set (for one particular student) and checks if they are recommended the fourth course or not.
- Step 2 – The user proceeds to then change the input to 2 out of 4 (for a recommendation of the other 2 courses), and 1 out of 4 courses (for a recommendation of the other 3 courses)

Results

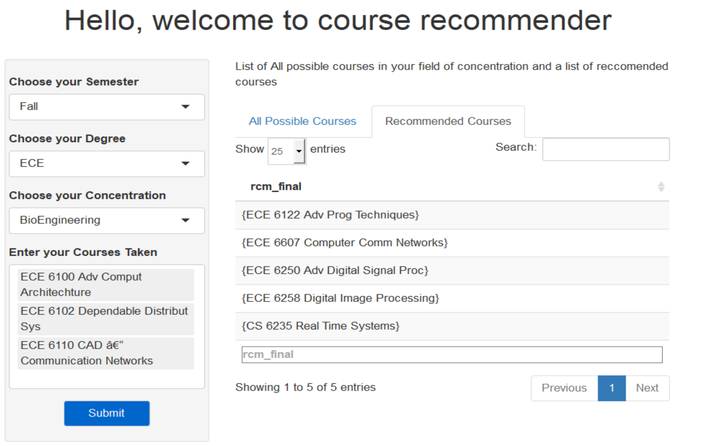

Figure 2: A screenshot of the course recommender web app in action

- 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 improvement 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

1. *Hahsler, Michael, Bettina Grün, and Kurt Hornik. "Introduction to arules–mining association rules and frequent item sets." SIGKDD Explor 2.4 (2007): 1-28.*
2. *Ds, Grewal, and Kaur K. "Developing an Intelligent Recommendation System for Course Selection by Students for Graduate Courses." Business and Economics Journal, vol. 7, no. 2, 2015, doi:10.4172/2151-6219.1000209*

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].