Ozan Gokdemir, Rain Kwan, Thomas Kane
Data Mining and Machine Learning Fall 2018
Assignment 1 - The Apriori Algorithm
September 19th, 2018

Mining for Associations Between Movie Preferences

In the rapidly growing field of machine learning, application of data mining algorithms for creating personalized product suggestions has been transforming numerous fields of business. Such algorithms are very well-integrated into the entertainment industry, powering the business models of renowned companies like Netflix, Hulu and Amazon Prime Instant Video. Particularly interested in the effect of machine learning in acute transformation of Netflix's business model, we choose to study the MovieLens dataset.

Our dataset contains 100 thousand rows, comprised of two columns: the user id and the movie id. The number of movies watched by each user ranges between 20 and 737. We wrote a script in R to group entries by user id. We then sorted this grouped list, again, using R. This cleaning and formatting process gave us a csv file that we could process with Python 3. Then, we wrote a script to generate a list of sets, in which each set contains all the movies that an individual user watched. Intuitively, each of these sets represent a “transaction”, the format of data that is suitable for the Apriori algorithm.

Although our initial implementation of the apriori algorithm generated valid associations derived from the provided test dataset, it turned out to be extremely inefficient on our movie database, demonstrating tremendous time and space complexities. Our attempt to run it on our movie database of 100 thousand entries took approximately 6 hours before the execution was interrupted with an “out of memory error”. By the time the execution was interrupted, it had overflowed from 12 gigabytes of RAM and had performed approximately 400 gigabytes of write operations on the swap space of the SSD!

In order to mitigate the complexity of the algorithm after our disastrous first attempt, we added in a hashmap cache for support values of itemsets. Additionally, we implemented the transaction reduction method that we learned about from a research paper*. Further testing revealed that by increasing our support threshold from 10 to over 160, and increasing our confidence threshold from 0.25 to 0.98, the algorithm executed quickly and generated 17 association rules. While the transaction reduction method tremendously ameliorated the performance, the support cache turned out to be redundantly increasing the space complexity(up to 9 gigabytes of RAM) and making no significant contribution to the time complexity. Therefore, we removed the cache.

The initial form of our association rules were mappings of numeric movie id values to the others. For example, frozenset({'95'})===>frozenset({'50', '174'}) with confidence = 0.95652. In order to interpret these findings, we wrote a Python script to process the movielisting.txt file into a hashmap that maps the id of a movie to its name. Then, we used this hashmap to replace the movie id in the rules with movie names. This provided us with a result that looked like this: frozenset({'Star Trek 3: The Search of Spock'})===>frozenset({'Star Trek: The Wrath of Khan'}) with confidence= 1.0.

Analysis of our findings reveal that the Star Trek series(particularly Search for Spock, The Undiscovered Country and The Wrath of Khan) is quite popular- satisfying 160 support, 0.98 confidence thresholds. We also found that people who watch the Star Wars series tend to watch Star Trek, too, and vice versa. Toy Story series, along with some other animated movies shares a common audience. Furthermore, this audience seems to be interested in Star Wars, as well. In general, we found that users' movie preferences vary within their preference of genre. Finally, our findings demonstrate strong associations between movies of the same series. That is to say, if a series is popular, future sequels of the series are also likely to be popular.

Since we found strong associations within series and between series of the same genre, we would recommend the distributors of these movies to present sequels of the same series and sequels of series of the same genre together. To clarify, if a user has watched a Mission Impossible movie, it would be clever to suggest them not only another Mission Impossible movie but also a Fast and Furious movie since these series are both comprised of action movies.

(*) Singh et. al, Improving Efficiency of Apriori Algorithm Using Transaction Reduction, 2013 can be found at: http://www.ijsrp.org/research-paper-1301/ijsrp-p1397.pdf