Group Proposal - Data Mining Final Project

Aaron Gauthier, Pedro Uria, Zachary Buckley

The problem we have selected is to predict Netflix user ratings for movies based on their previous movie ratings and those of other users. We have selected this because we are aware that Netflix has provided anonymous customer ratings previously, and believe we can expand on a competition dataset that was provided between 2006 and 2009 to potentially get good prediction results using a combination of the data-mining techniques we've been learning in class.

Kaggle.com has the dataset Netflix provided for a competition aiming to improve their recommendation engine by at least 10% (https://www.kaggle.com/netflix-inc/netflix-prize-data). The data provided by Netflix does not seem to have a feature number problem, as once the data is combined into a single table, the vast majority of the data revolves around only 4 features (movie_id, user_id, rating, date). We will be doing data integration, pulling in metadata about specific movies from additional sources (currently looking at imdb and similar sites). It is likely, based on prior experience with genre information, that getting genre features into a usable form will require some cardinality reduction, particularly by 'combining two of more categories into one', as mentioned in the text.

Instance Selection on the existing Netflix data, we plan to use Cluster Sampling (p. 157) based on movie genre to start with, and may expand to other techniques like Data Clustering (p. 159) as our analysis progresses (potentially K-means on the weighted features). After clustering the users, we will use these subsets of the data to train supervised learning models in order to predict user rating. A different model would be trained for each cluster.

We'll be using the python libraries scikit-learn, NumPy, and pandas to implement our analysis, and preprocessing code. We will base the performance of our results on comparing our predictions with the provided test dataset from the kaggle site, using RMSE (which was used for the netflix prize competition).

Rough Project Schedule follows on the next page:

Date	Milestone	Description
3/31/19	Proposal Draft	Complete Draft Proposal and Topic Selection
4/2/2019	Finalize Topic Choice	Discuss topic with Amir, and Priyanka
4/6/2019	Group Proposal Due	Due Date for Group Proposal Submission
4/7/2019	Integration Complete	Finish code for loading in netflix and imdb data
4/14/2019	Analysis Complete	Complete Data preprocessing and analysis
4/20/2019	Group Paper/Presentation	Complete Group Paper and Presentation
4/21/2019	Individual Papers/Cleanup Project Complete (Submitted)	Complete individual papers/general cleanup
4/22/2019	Final Due Date	Due Date for Group Final Report and Presentation Submission