CS6220 Data Mining Techniques

Group Project Proposal

Group Centaurus
Lu Liu, Ke Yang, Mengshi Li, Jing Shen, Siyu Chen

Project Overview

Nowadays, people have multiple ways to entertain, and TV shows are one of the most popular things that people would like to do after work or school. There are a lot of online streaming services and platforms, like Netflix, Disney+ and HBO, all providing tons of movies or shows to the audience. However, different people have different tastes, so now the platforms also need to provide intelligent recommendation systems to the user so they can quickly find out what they like much sooner. For example, anime now is very popular not only with kids or teenagers but also liked by a lot of adult audiences. So if there is a more detailed and specific recommendation system that can help the users find out what they want to watch in tons of animations, it can be very practical in approaching more users from the platforms' perspective.

In our final project, we will propose the use of a recommendation system using data mining techniques to recommend anime. There are several steps for our team to conduct in order to accomplish the project. Firstly, we will research online to get a suitable database for the project. Secondly, our team will get a suitable machine learning model for it. Thirdly, we will process features and make the features work for the model. Then, we will train the model and tune the parameters for better performance. Finally, metrics and visualization will be used to help demonstrate the project results.

Data Collection

The dataset for this project has been found online. We will use MyAnimeList Database 2020 from Kaggle. Link: https://www.kaggle.com/hernan4444/anime-recommendation-database-2020
In the snapshot of the dataset, we can see it has approximately 17000 training examples. Each example has 35 columns to explore, providing sufficient analysis potential. We also considered other datasets for comparison and picked the most suitable one considering time updated, usability rating, examples count, and features count. The dataset also contains subsets for both user-based and item-based ratings, which will be helpful when applying different machine learning models.

Detail Compact Column 10 of 35 columns

MAL_ID	Name	Score	Genres	English na...	Japanese ...	Type
1	Cowboy Bebop	8.78	Action, Adventure, Comedy, Drama, Sci-Fi, Space	Cowboy Bebop	カウボーイビバップ	TV
5	Cowboy Bebop: Tengoku no Tobira	8.39	Action, Drama, Mystery, Sci-Fi, Space	Cowboy Bebop:The Movie	カウボーイビバップ 天国の扉	Movie
6	Trigun	8.24	Action, Sci-Fi, Adventure, Comedy, Drama, Shounen	Trigun	トライガン	TV
7	Witch Hunter Robin	7.27	Action, Mystery, Police, Supernatural, Drama, Magic	Witch Hunter Robin	Witch Hunter ROBIN (ウイッチハンターロビン)	TV

Figure 1. MyAnimeList Dataset Snapshot

Data Cleaning and Processing

The dataset preparation will be done using various data cleaning and processing techniques, including missing value and unbalanced data handling. The dataset also contains both categorical and numerical data, which we can apply the hot encoding technique to map categorical features for comparison. We will also perform data normalization to ensure each data entry can be utilized the same way across the dataset.

Feature Selection

The main dataset we are using has 35 columns, among which 10 columns are the number of users who scored the anime by 1 to 10. Due to the fact that our feature space was so large, we will reduce the feature space and simplify the complexity of the high-dimensional dataset while retaining trends and patterns by using PCA for dimension reduction. We will first evaluate the relationship between each feature and the target result and then transform the columns into several principal components that can explain the majority of the information or the data variance of the feature set.

Modeling

Unsupervised

Unsupervised Nearest Neighbors is a preferable model for user-based collaboration filtering. We will use KNN as a baseline to recommend animes based on similar users. If two users gave similar rates on a certain amount of animes, we will consider them as similar users.

Content based approach utilizes distinct features of an instance to make recommendation for additional instances that have similar properties. To implement the content based

approach, we will apply the **TF-IDF(TfidfVectorizer)**. This will recommend animes purely based on the title a user has already rated. So we will first filter the animes that this user gives high rates, then based on these animes, we will recommend new animes to them.

Supervised

We think that the rate a user may give to an anime seems to have a relationship with how they rate each genre(Action, Adventure, Comedy etc). And based on the rates, we can determine what anime we will recommend to users. So we will calculate the average scores that every user gave to each genre. Then we will use these scores as input, and use the scores these users gave to an anime as target value. We will test around 100 animes in total to get the average performance for each model.
We will test **Linear Regression**, **Support Vector Machines** and **Naive Bayes models**. Then we will evaluate the performance of these models and find the best one.

Model Evaluation

We are using the underlying assumption of the collaborative filtering approach for unsupervised models, and we planned to implement the content-based recommender based on supervised model. In order to evaluate with different methods our models, we need to test the statistical accuracy metrics like MAE, RMSE. Also we will test two important decision support accuracy metrics: precision and recall, including MAP@K and MAR@K, the former helps evaluate how relevant the items in the recommendation system are, the latter gives insight into how well our recommendation system is able to recall all items the user gives at high rates. To avoid recommending the same anime again and again, we need to maximize the precision and recall at K.

Project Deliverables

The final deliverable will consist of a final report, a Jupyter Notebook, and a presentation to demonstrate how we complete the project. We will include detailed data processing, model training, and results analysis in the report to answer exciting questions we raised for the dataset. In addition, a comparison among different algorithms and models using metrics and stats reports will also be discussed in the final deliverables.