Project Proposal

Title

Movie Recommendation System Considering Bias in Ratings

Member

Jingci Wang, Junmei Luo

Problem

Recommendation System seeks to predict the "rating" or "preference" that a user might give to an item. It has been widely used in movies and TV shows services like Netflix and YouTube. We want to build a movie recommendation system to recommend new movies to users based on history records. We intend to do a comparative study on the different algorithms used to do recommendation popularly and build a hybrid model out of them. Furthermore, we will analyze the bias potentially within the datasets and try to refine our model accordingly.

Key Components

1. **Recommendation System**

1) **Popularity-based Recommender** makes the same recommendation to every user, based on the popularity of movies. One possible solution to the cold-start problem.

2) **Collaborative Filter Recommender** focuses on the relationship between users and items. Mainly used for users with rating history. We would use both Pearson correlation and Cosine Similarity to determine the similarity between users and movies.

- User-based Collaborative Filter
- Item-based Collaborative Filter
- Matrix-Factorization
- Neural Collaborative Filtering: This method comes from a literature published in recent years that we found quite interesting. It mainly built a collaborative filter framework by introducing the powerful deep neural network. We would like to try to reimplement the frame of neural collaborative filtering and compare the results with other methods we would use.

2. **Find bias in movies rating**

There can be some bias in movie ratings. To be more specifically, some movies are universally loved or hated but some users may have niche interests, so that the ratings cannot reasonably represent people's opinions of movies; some users are pickier than others which will result in high or low scores for all the movies rated by one person

regardless of preferences because of the lack of uniform standards for ratings. To tackle these problems, we plan to do some data exploration first and find the pattern of bias underlying the datasets. Then we would try to make use of machine learning knowledge such as introducing bias term to find if we could better reduce the effect of bias rating.

3. **Dataset**

Our datasets come from Kaggle Competition. The datasets describe ratings and free-text tagging activities from MovieLens, a movie recommendation service. It contains 20000263 ratings and 465564 tag applications across 27278 movies. These data were created by 138493 users between January 09, 1995 and March 31, 2015. This dataset was generated on October 17, 2016. Users were selected at random for inclusion. All selected users had rated at least 20 movies.

The information in the data includes UserId, Movies information (MovieId, title, genres), Tags (Tags applied to movies by users, TagId, Movie-Tag relevance data, Tag descriptions), Ratings of movies by users, Identifiers that can be used to link to other sources.

Validation

We will use Root Mean Squared Error (RMSE) to measure how our recommendation system performs. RMSE can measure the degree how each user is interested in each movie and represent the accuracy of recommendation. The less the RMSE is, the better our system performs.

We also care about how many movies recommended by our system are really the right ones for users, so we set a standard of ‘like a movie’ for whose rating is above 3. We can draw ROC Curve and the area under the ROC Curve presents the percentage of movies that be recommended belong to ‘like a movie’. The larger the area is, the better our system performs.

References

[1] Dataset: Kaggle Competition
https://www.kaggle.com/grouplens/movielens-20m-dataset#genome_tags.csv
[2] Neural Collaborative Filtering:
https://www.comp.nus.edu.sg/~xiangnan/papers/ncf.pdf