Proposal of EECS 349 Machine Learning

Yaliang Wang(ywo130) Sisi Chen(sci963) Haomin Zeng(hzy075)

Our initial task is to accomplish streaming media recommendation system. In other word, given a user with incomplete information, our recommendation system can imply to which extent this user like the music of a given artist. Furthermore, the recommendation system can list out the priority of artists based on our prediction for a user.

This task is of both high interestingness and great significance. On the one hand, this recommendation system benefits users of streaming media a lot. Usually, given users of streaming media might have a same taste of music, the recommendation system can reach out to new music which satisfy their personal music taste. On the other hand, the recommendation system also benefits for provider of streaming media. The providers can design personalized music album for specific users to cater to users' music favor and make more profits.

In the HetRec 2011, a dataset of Last.FM with 92,800 artist listening records from 1892 users, generated by the Information Retrieval Group at Universidad Autónoma de Madrid Datasets, has been released(avaliable online http://grouplens.org/datasets/hetrec-2011/). The dataset includes bi-directional user friend relations, user-listened artist relations(include the listening count), and the user's created tag for a specific artist.

These three types of information will be used as the features for recommendation. The tag information of artist can be used to generate the "genes" of a particular artist and the listening count, as well as the user friend relations, can be employed to create a user's preference. The general idea of the recommender system is to build user's preference and make a best match with the artist's "genes" in order to recommend artist for different users.

The structure of data is well designed and easy to use. Besides, the ML techniques we plan to try first is k-nearest neighbor algorithm. If treating the tag information of the artist and the listening history of user as the features, we can generate a high dimensional feature space and then project them to a low dimensional space with t-SNE algorithm. The groups of users and artists in the low dimensional feature space can be assign into k classes based on the k-nn algorithm.

In order to evaluate the success of our system, we will divide the users of dataset into two parts - training and testing parts. In testing, we give information about bi-directional user-friend relationship and partial tag and listening information of testing user as an input, the recommendation system will compute out the list of artists, who are not reached before, ordered according to favor priority. We will compare this list with an established list of listening count of the testing user to evaluate our system accuracy.