Homework 5
Researching Applications of Machine Learning

CMU 10-601: Machine Learning (Spring 2017)
https://piazza.com/cmu/spring2017/10601
OUT: March 08, 2017
SUMMARY DUE: March 22, 2017 11:59 PM
PEER REVIEW DUE: April 5, 2017 11:59 PM
TAs: Rui Sun, Prakruthi Prabhakar, Simon Du, Sarah Schultz

The title of the paper here

1 Data Description

I will use the "MovieLens 1M dataset", which comes from University of Minnesota researchers started from 1997(GroupLen research project). The MovieLens dataset is composed from 3 different .dat file which represents generic data file. I think the challenges on this dataset is user bias. For user bias part, the MovieLens data set only collect data from users with at least 20 ratings, however, for people who didn't like to rate movies, those people in the MovieLens dataset maybe no reference value for them.

2 Task Description

The author introduced two implementations of Neighborhood-based collaborative filtering algorithm. One is user-based, another is item-based. The author detailed described how to choose the top-k users/items for user-based/item-based collaborative filtering, and compare the advantages/disadvantages as well as their efficiency. For me, I think the most challenging thing is space demand. Although this neighborhood-based algorithm leverage the offline preprocessed data to enhance the online efficiency, however, offline phase costs lots of space and time. The most interesting topic for me is how to integrate user-based and item-based algorithm to create more efficient system online.

3 Method

The Neighborhood-based algorithm calculated the k nearest user/item in advance by cosine similarity/Pearson similarity. Then the predict function will use the weighted average of these neighbors to get the predict value. The most important part to let this algorithm efficient is preprocessing the similarity of the data, which let online prediction can get the result very quickly.

4 Results

The Neighborhood-based algorithm is simple, easy to implement and debug as well as interpreted. Generally, item-based method contains higher accuracy because the predicted result is based on the past results of the same user. However, user-based method use prediction of other similar users, which may still exist difference for non-overlapping parts. For me, I don't prefer either method specifically. I like different kinds of movies, if I always get recommandation from item-based, I lose the chance to get suggestion from other users. On the other hand, if always using user-based, I may get movies I don't have interested sometimes.

5 Questions about the Paper

In the 2.3.4, the author told us that item based need less frequently do the offline computing than user-based since the increasing rate of item is much lower than users. However, I wonder know what's the threshold for recomputing the similarity? And how do you decide the threshold value?

References

[1] Charu C. Aggarwal. *Recommender Systems.* Springer International Publishing, Switzerland, 2016.