A hybrid approach recommendation: Combining content-based and collaborative filtering

Jinling Xing

Polytechnique Montral

November 29, 2017

Outline

Introduction

I introduced a hybrid approach to recommender systems with the following characteristics.

The system combining both content-based(CB) and collaborative filtering(CF) approaches.

The CB component of the system encompasses two matrices: the userDistance and the itemDistance matrices.

The CF component of the system relies on the typical user-user and item-item similarity matrices computed from the known, past user-item ratings.

Dataset

The dataset comprises 100,000 ratings from 943 users on 1,682 movies. The dataset can be downloaded from the Recommendation System course website.

1. u.data, which comprises basic information on users (gender, age, occupation, zip code).

2. u.item, which comprises all item information (title, release date, genre).

3. u.user, which comprises the ratings on 5-point scales for all user-item pairs.

Preprocess data

I delete some columns like the $'IMDb.URL'$,$'unknown'$ from the u.item and only save the information about the movie's genre and release data.

Then, I separate the year-month-date from the release data and only save the release year.

Finally, I got the "*voteitem*", which is the data I really want to use. It's a user-movie matrix(each user votes each movies)

Method principle

First, I achieved the following methods separately

- 1. Content-based systems.
- 2. Collaborative filtering systems.

Second, I used a linear regression to combine these two methods.

Third, I used the cross-validation to compute the Mean Absolute Error(MAE) and Mean Squared Error(MSE) to check this hybrid method.

Collaborative filtering method

The collaborative filtering component of the system used the "*Pearson Correlations*" to compute from the known, past user-item ratings, providing for a memory component of the recommender.

Finally, I got a User-User similarity matrix(943*943) and an Item-Item similarity matrix(1682*1682).

content based method

The content based component of the system computes the "*jaccard function*" based on the users and items. The *jaccardfunction* usually is used on a sparse Matrix.

Finally, I got a UserDistance matrix(943*943) and an ItemDistance matrix(1682*1682).

Combine the CF and CB

Adding weight to CF and CB. The vote by users or the item vote can be calculated as the following formula:

Theorem

$$Vij = W_1 * Vij(CF) + W_2 * Vij(CB)$$

Prediction & Cross validation

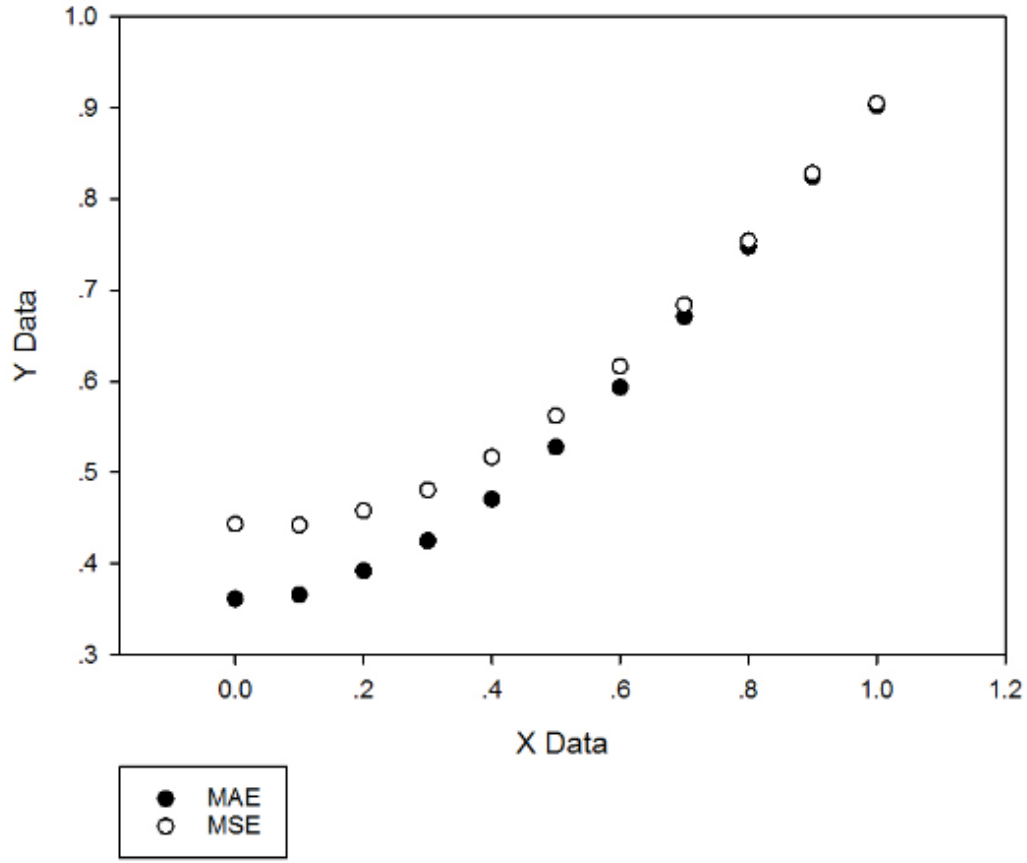
- Prediction: Compare cosine similarity and do User-User and Item-Item Prediction
- Cross validation: Compute prediction for each vote while keep other original vote and Mean absolute error and Mean square error for User-User and Item-Item separately.

The hybrid method result for User-User

UserDistance	UserUserSim	Mean Absolute Error	Mean Square Error
0.0	1.0	0.3611645	0.4430539
0.1	0.9	0.3657854	0.4421633
0.2	0.8	0.3917284	0.4579008
0.3	0.7	0.4248891	0.4807310
0.4	0.6	0.4705987	0.5167096
0.5	0.5	0.5274834	0.5619436
0.6	0.4	0.5932980	0.6164540
0.7	0.3	0.6709412	0.6836926
0.8	0.2	0.7475969	0.7540629
0.9	0.1	0.8246948	0.8282960
1.0	0.0	0.9022308	0.9048868

Table: The hybrid method result for User-User

The hybrid method result for User-User



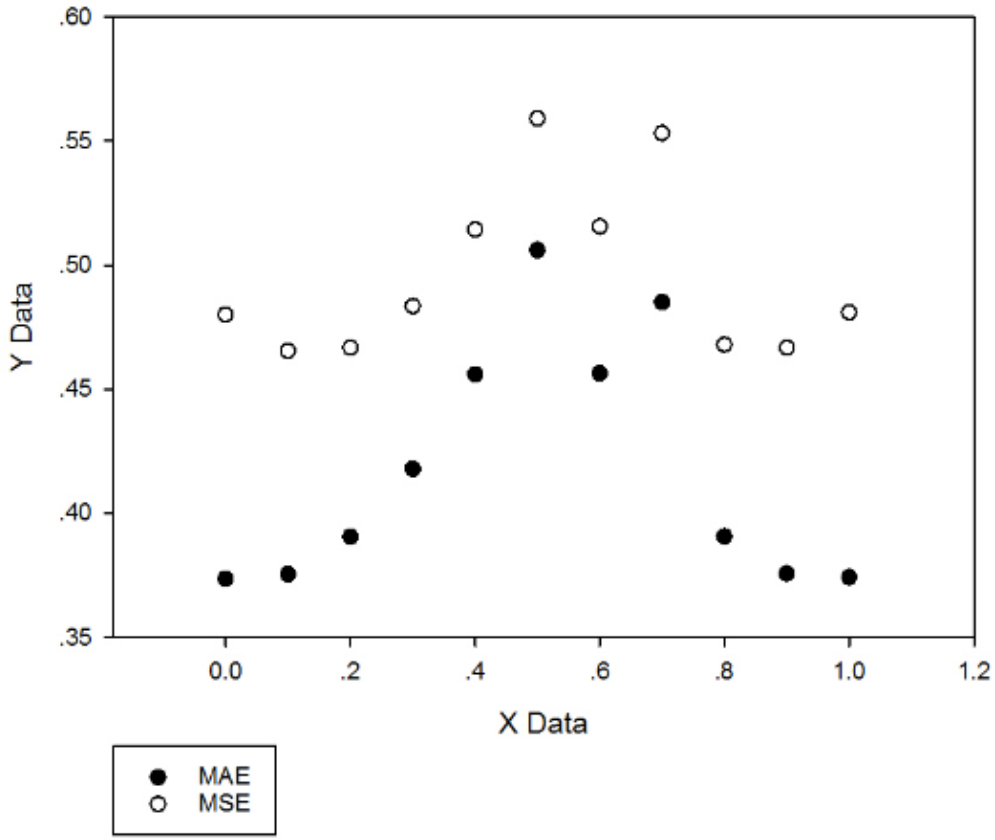
The X-axis represents the range of UserDistance is from 0.0 to 1.0 and corresponding the range of UserUserSim is from 1.0 to 0.0.

The hybrid method result for Item-Item

ItemDistance	ItemItemSim	Mean Absolute Error	Mean Square Error
0.0	1.0	0.3735659	0.4800009
0.1	0.9	0.3753491	0.4654963
0.2	0.8	0.3904772	0.4667025
0.3	0.7	0.4177423	0.4835021
0.4	0.6	0.4559735	0.5143693
0.5	0.5	0.5060185	0.5591446
0.6	0.4	0.4563085	0.5155504
0.7	0.3	0.4850656	0.5531617
0.8	0.2	0.3905734	0.4679713
0.9	0.1	0.3756668	0.4666804
1.0	0.0	0.3741724	0.4810178

Table: The hybrid method result for Item-Item

The hybrid method result for Item-Item



The X-axis represents the range of MoviesDistance is from 0.0 to 1.0 and corresponding the range of ItemItemSim is from 1.0 to 0.0.

Conclusion

The collaborative filtering performed well on User-User analysis, when the weight of content-based increased, the performance decreased. The collaborative filtering and content-based approached both performed well separately on Item-Item analysis.

But combining the CF and CB together, the User-User and Item-Item both performed worse.

The value of MSE in both two cases is higher than MAE, because of the square is a big number.

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} \qquad MSE = \frac{\sum_{i=1}^{n} (y_i - x_i)^2}{n}$$

We can conclude that this hybrid method combining CF and CB does not suit the sparse matrix.

Questions?