A Movie Recommendation System Using Collaborative Filtering

Anand Sahoo

Kshitiz Parashar

Ruxin Zhang

Wei Yu

Yuchen Zeng

Table of Contents

1. Executive Summary

This report analyzes data from the results of a movie survey filled out by 96 UC Davis MSBA students. The dataset contains each user's rating of 47 different movies. The objective is to use Collaborative Filtering to build a recommendation system based on the survey results. The designed model predicts rating scores for existing users, three new movies, and three new customers. We recommend extending the model's use case to businesses in e-commerce and mobile commerce. For example, a clothing retailer can reduce overproduction by adopting the Collaborative Filtering algorithm. A better model has also been recommended which takes Item-based collaborative filtering into account. This new approach requires the addition of new attributes to the dataset such as Movie genre, language, origin, etc.

2. Introduction

For streaming services, such as Netflix, one of their focuses is their recommendation systems. The more accurate their recommendation system is, the better customer experience from the streaming service. If the recommendation system provides accurate personalized suggestions to customers, customers can save time from looking for content that they are interested in. Moreover, the recommendation system can provide the company with a better understanding of what its customers are interested in. This helps them generate more profit by avoiding offering content that is not attractive to customers and propagating resources to what is attractive. Therefore, analyzing collected data from customers to improve the recommendation algorithms continually is of paramount importance. In this report, we will build our own recommendation system based on the data collected from the movie surveys.

3. Problem Formulation

Our analysis analyzes a database from a movie survey filled out by UC Davis MSBA students. To start the analysis, our first objective is to build a model to predict three movies'

scores of our group members. Then, we will utilize our model to predict three assigned movies and three new customers to further discuss our models' accuracy.

We implemented a user-based collaborative filtering algorithm to build our model. A collaborative filtering algorithm allows us to predict a user's score for a particular movie based on users who have similar movie tastes with the targeted user. To predict ratings for new movies and new customers, we will use average ratings while taking user bias into account.

4. Data Description

In this report, we use 96 users' ratings of 50 different movies to build the recommendation system. The rating is from 0 to 5 with 0 representing the user hasn't seen this movie before. Among all the users, one user gives the same ratings for all 50 movies, which leads to zero standard deviation. Thus, we remove this user's rating from the dataset.

5. Model Development

5.1 Predicting ratings for existing users and existing movies

We applied a user-based collaborative filtering algorithm in this report to build the recommendation system. The idea of user-based collaborative filtering is to find similar users according to their ratings of different movies, and predict a certain user's rating of movies according to his similar users' ratings.

First, we normalized the ratings of each user by calculating the user-user z-score. This is to prevent rating inflation or deflation of certain users. After data normalization, we calculate the Cosine Similarity between target users (the users that we are going to predict their ratings. They are user23, user 56, user 78, user 84, and user 87) and each other users. The Cosine Similarity represents the similarity between different users, the higher the value, the more similar the users are. Finally, we used the normalized Cosine Similarity as the similarity weight of each

user and calculated the weighted z-score for each target user. We can further get each target user's predictive ratings by using the weighted z-score.

5.2 Predicting ratings for existing users and non-existing movies

Since there is no data on the movies from other users, cosine similarity is not a good metric for the rating prediction. The other two metrics which make much more sense were mode and average ratings of the user. Taking the average of the total ratings the user has given till date can be assumed to be their rating for any new movies that appear in the database.

5.3 Predicting ratings for non-existing users and existing movies

The better metric of prediction for new users was to take the average of overall ratings for a particular movie. This basically assumes the new user to rate a particular movie similar to the historical average rating of the movie.

5.4 Predicting ratings of existing movies for new customers (after they have rated some movies)

When the new user adds their own ratings to the movie database, the problem essentially becomes similar to the first problem (Section 5.1). A user-based collaborative filtering algorithm is the basis of this recommendation model. After rating normalization, cosine similarities are calculated for the users which prove to be a good enough metric for predictions on this **limited** dataset. The term limited is put in bold to signify that there is scope to improve this recommendation system by introducing new and better variables, which is discussed in section 7 of this report.

6. Results

5.1 The following table shows the true and predicted ratings of 5 target users on 3 movies.

	23-Yuchen Zeng		56-Anand Sahoo		78-Kshitiz Parashar		84-Ruxin Zhang		87-Wei Yu	
	TRUE	Estimated	TRUE	Estimated	TRUE	Estimated	TRUE	Estimated	TRUE	Estimated
The Social Network	4	3.76	4	4.02	4	4.07	0	2.65	0	4.03
A Prophet	3	3.58	0	3.81	0	3.97	0	2.67	0	4.15
Amour	3	3.56	0	3.80	0	4.00	0	2.61	4	4.12

5.2 The following table shows the predicted ratings of the same 5 target users on 3 new movies.

	23-Yuchen Zeng	56-Anand Sahoo	78-Kshitiz Parashar	84-Ruxin Zhang	87-Wei Yu
Winter's Bone	3.56	3.81	4	2.67	4.16
A Serious Man	3.56	3.81	4	2.67	4.16
Son of Saul	3.56	3.81	4	2.67	4.16

5.3 The following table shows the predicted ratings of 3 new target users on 3 movies.

	Camille	Shachi	Amy
Avatar	4.08	4.08	4.08
The Wolf of Wall Street	4.59	4.59	4.59
Inception	4.14	4.14	4.14

5.4 The following table shows the predicted ratings of the same 3 target users on the same 3 movies as the prior part, but this time we have some ratings from these users.

	Camille	Shachi	Amy
Avatar	2.63	4.11	3.18
The Wolf of Wall Street	2.70	4.16	2.74
Inception	2.62	4.19	2.87

7. Recommendations and Managerial Implications

From our result, Kshitiz has a true rating of 4 for *The Social Network*, while the estimated score is 4.067. The estimated rating scores indicate that the user-based collaborative algorithm is effective to predict a targeted customer's taste by analyzing a customer who has a similar taste with this targeted customer. Therefore, by implementing the user-based collaborative algorithm to e-commerce and mobile commerce, retailers can provide customers personalized and more accurate suggestions to improve their customer experience. Moreover, the recommendation system can help retailers to know their customers' tastes better, to reduce unnecessary production in advance. For example, if a clothing retailer implements a recommendation system and notices their customers are not interested in a particular design of skirts, then they can reduce the production of this design or similar design of skirts to avoid unnecessary waste.

But the model works only when the prediction pertains to existing users or existing items (here movies). For other cases, we had to take the average of user/movie ratings for our predictions. This is not an accurate depiction of the real-world scenario. Another metric to predict new movies for existing users can be mode. This metric keeps a check on the user's bias in rating the movies. For example, if a user generally rates most of the movies they watch as 4, then it is wise to predict that a new movie they watch will be rated as 4.

However, this dataset only has one feature, movie ratings. This limitation on the dataset limits the model to be only based on User-based Collaborative filtering. Several user biases affect the prediction accuracy. User's item taste changes temporally and that is hard to account for in the above model. A better alternative would be Item-based Collaborative Filtering. This approach finds similar items (movies) instead of similar users and then recommends similar movies that the user has had in their past preferences. For this, the genre of the movie would be an appropriate feature addition to the dataset. Other important features would be: the demographic origin of the movie, language of the movie, etc. Item-based Collaborative Filtering is advantageous over the User-based approach because the items don't change over time. This reduces the impact of user bias on the model and hence results in better recommendations.

8. Conclusion

The Collaborative Filtering algorithm is effective when there is an existing database of customers' information (i.e., an existing customer's taste toward an existing product). This algorithm had a good performance. For example, the actual rating score that our group member gave to a particular movie is 4 and our predicted rating is 4.06. However, it is not effective to provide predictions for new customers and new products. In our movie recommendation system, we used the average rating scores to predict ratings for new customers and new movies, and we also take users' bias into account. Transitioning from User-based Collaborative filtering to Item-based Collaborative Filtering is recognized to be a better modeling approach.