Clustering Based Movie Recommendation System

Wojciech Nagórka, Kuba Czech, Vasyl Korzavatykh, and Andrii Chmutov

Poznan University of Technology Faculty of Computing and Telecommunications Piotrowo 3, 60-965 Poznan, Poland

June 9, 2024

Abstract

In this study, we present a movie recommendation system tasked with predicting film ratings based on the data stored in the MovieLens dataset. We create different models that include Genre-Based, Cluster-Based (using K-means), Movie-Based Regressors, and a hybrid approach, which is a combination of the aforementioned models. Our goal is to improve the prediction accuracy of simple methods.

1 Introduction

Movie recommendation systems have become an essential part of streaming services, providing users with personalized content based on their preferences. Some films which are well-rated on average might not be the best recommendation for everybody. That's why recommendation systems should take into account not only the film but also the user and their preferences. The aim of this approach is to create personalized recommendations and improve system performance in general.

2 Related Work and Types of Recommender Systems

The recommendation systems are the core of all streaming platforms. Services such as YouTube and Netflix take advantage of these methods to engage users for extended periods of time.

2.1 Collaborative Filtering Approach

Collaborative filtering approaches try to use the whole dataset in order to make a prediction. They take into account ratings of similar users [1].

2.2 Content-based Approach

Content-based approaches focus on each user's past preferences.These algorithms try to recommend objects that are just like those that a user favored in the past [2].

3 Dataset

We utilize the MovieLens dataset, a widely used benchmark in the recommender systems community. The dataset contains user ratings for a diverse collection of movies, along with metadata such as movie titles, genres, and release years. The dataset includes 33 million ratings given to 86 thousand films by over 330 thousand users. The data was collected between January 9, 1995, and July 20, 2023.

4 Algorithm

Our approach is centered around combining a couple of models in order to maximize prediction accuracy. We create three separate models, each one focusing on a different aspect of the dataset. Lastly, we combine them into one system in order to exchange the information gathered by each component. The proposed Hybrid Regression model consists of three separate models:

4.1 Methods

The **Genre-Based Regressor** predicts movie ratings based on user preferences for movie genres. It calculates the average rating given by a user for each genre and uses these averages for predictions.
The **Cluster-Based Regressor** uses the K-means clustering algorithm. It groups movies based on their genre, year of release and mean of user ratings. This model predicts ratings for unseen movies by utilizing the clusters obtained from K-means.
The **Movie-Based Regressor** is designed to predict movie ratings based on the average ratings of individual movies. This approach leverages the historical average ratings of each movie to make future predictions.
The **Hybrid Regressor** is obtained by combining the qualities of all the previous models. It takes the prediction results of three components: Genre-Based Regressor, Cluster-Based Regressor and Movie-Based Regressor, and returns a weighted average of them as a result.

Regressor	**Weight**
Genre-Based	0.35
Cluster-Based	0.45
Movie-Based	0.2

Table 1: Weights for Different Regressors

4.2 Parameters

According to our experiments, the number of ratings in the dataset for a given user heavily impacts the prediction quality for them. It is quite logical since the more information about the user we have, the more informed our prediction becomes.

Figure 1: The influence of the number of instances in the training set on performance

One of the components of the ensemble is the Cluster-Based Regressor, whose performance significantly relies on the number of clusters to be created. According to our experiments, the optimal number of clusters for our dataset is 5.

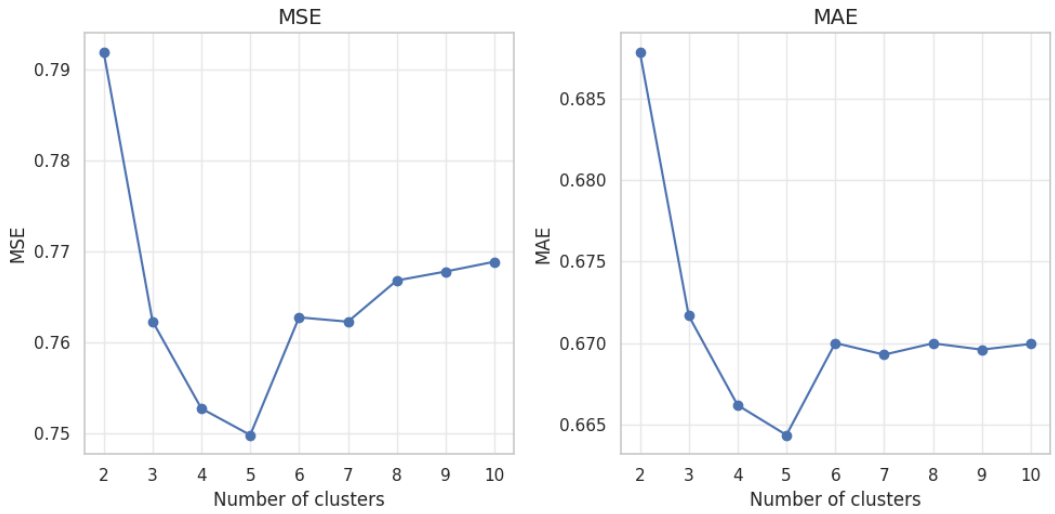

Figure 2: The influence of the number of clusters on MSE and MAE

5 Results

We evaluate the performance of our hybrid regression model using three metrics, namely mean squared error (MSE), mean absolute error (MAE), and accuracy (the number of correctly predicted ratings on the training set with a tolerated error of ± 1). We compare them with the naive approaches:

1. **Naive baseline** - we predict a rating of 2.5 for every pair of user and movie.
2. **Mean of training set** - we predict a rating equal to the mean rating in the training set for every pair of user and movie.

It is clear that the implemented regressors outperform the naive approaches. The MSE and MAE metrics dropped noticeably while the accuracy increased.

Model	**MSE**	**MAE**	**Accuracy**
Genre-Based	0.856	0.711	0.751
Cluster-Based	0.802	0.669	0.788
Movie-Based	0.932	0.745	0.727
Hybrid	0.75	0.664	0.78

Table 2: MSE, MAE, Accuracy of different models

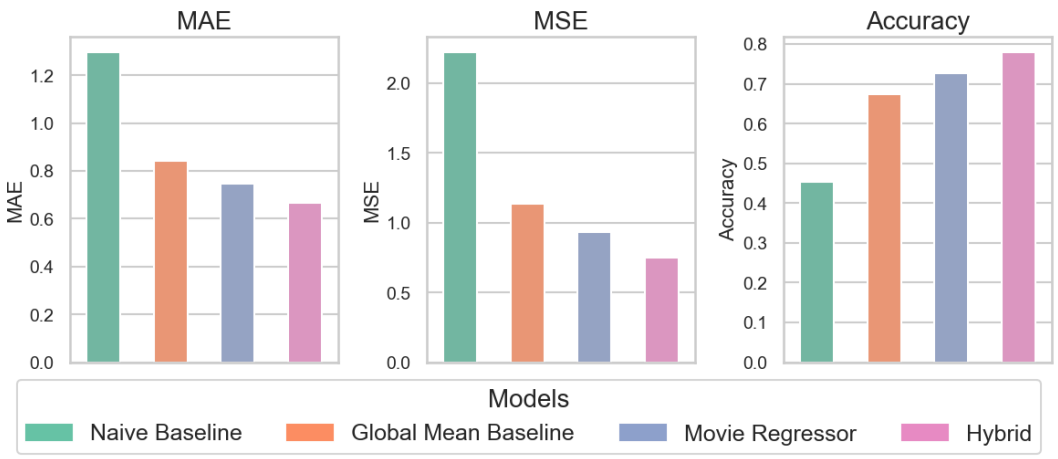

Figure 3: Comparison of Movie-Based Regressor and Hybrid Regressor with naive approaches

6 Conclusions

In this study, we introduced a hybrid regression model for movie recommendation systems that integrates Genre-Based, Cluster-Based, and Movie-Based Regressors.

The results of our experiment using the MovieLens dataset show that the hybrid model significantly improves prediction accuracy compared to naive methods. We were able to reduce the MSE to 0.75 and MAE to 0.66. The hybrid approach combines the advantages of each of its components, providing better predictions and resulting in a well-performing system. Some interesting findings:

- The amount of data we have regarding a particular user heavily impacts the prediction quality.
- The amount of clusters which produces the best results is 5. Increasing the number of clusters beyond 5 does not improve the prediction quality.

References

[1] Eda Kavlakoglu Jacob Murel Ph.D. What is collaborative filtering? 2024. URL: `https://www.ibm.com/topics/collaborative-filtering`.

[2] Uzma Javed, Khadija Shaukat, Imran A. Hameed, Faisal Iqbal, Tariq Mahboob Alam, and Shuhong Luo. A review of content-based and context-based recommendation systems. *International Journal of Emerging Technologies in Learning (iJET)*, 16(3):274–306, 2021. URL: `https://www.learntechlib.org/p/219036/`.