Clustering based movie recommender system

Vasyl Korzavatykh Andrii Chmutov Kuba Czech Wojciech Nagórka

Poznań University of Technology, Faculty of Computing and Telecommunications

POLITECHNIKA POZNAŃSKA

Introduction

When people try to decide on a movie to watch at the very moment, they often consider a variety of factors, for example mood, genre preferences, length of movie, actor, director etc. However, in recent years with development of streaming services such as Netflix, Amazon Prime Video or Disney+ people are offered personalized movie recommendations. On a wave of rising popularity of such services, we came up with an idea to implement such a system by ourselves.

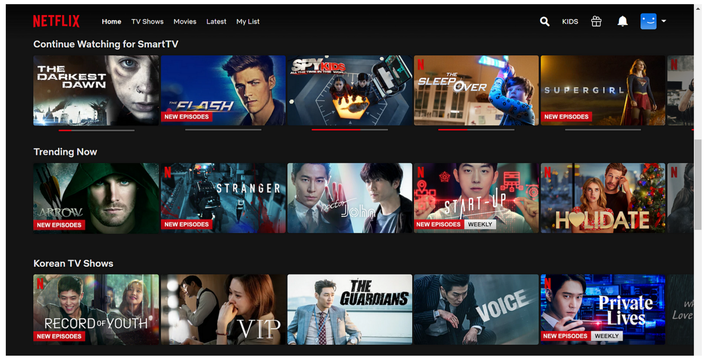

Figure 1. Netflix movie recommendation system

Objectives

During our research we asked ourselves many questions related to implementation and results.

- Is it possible to predict rating that a movie will be granted just from database of watched movies and user ratings?
- How accurate can be such a model? How much will the opinion of the user be different than the one calculated by our algorithm?
- Does more mean better – will we get better results for more clusters and more data in the training set?

We will try to answer above questions and more in next sections!

Methods

- **Genre-Based Regressor**: *GenreBasedRegressor* class predicts movie rating based on user preferences for different genres
- **Cluster-Based Regressor**: *ClusterBasedRegressor* class predicts movie rating by clustering (K-means) movies based on their genre, year of release and mean of user ratings
- **Movie-Based Regressor**: *MovieBasedRegressor* class predicts movie rating based on average rating of watched movies for each user
- **Hybrid Regressor**: *HybridRegressor* class predicts movie rating based on weighted average of results from previous regressors. Weights can be manually set and in our case they are equal to [0.35, 0.45, 0.2]

Results

Dataset

- **Training Data**: 90% of MovieLens dataset
- **Test Data**: 10% of MovieLens dataset
- **Features**: User ID, Movie ID, Genre information

Performance Metrics

- **MSE (Mean Squared Error)**: measures the average squared difference between predicted and true values
- **MAE (Mean Absolute Error)**: measures the average absolute value of difference between predicted and true value
- **Accuracy**: Proportion of predictions within specified tolerance (±1)

Results Summary

Model	MSE	MAE	Accuracy
Genre based	0.856	0.711	0.751
Clustering based	0.802	0.669	0.788
Movie based	0.932	0.745	0.727
Hybrid	0.75	0.664	0.78

Table 1. MSE, MAE, Accuracy of different models

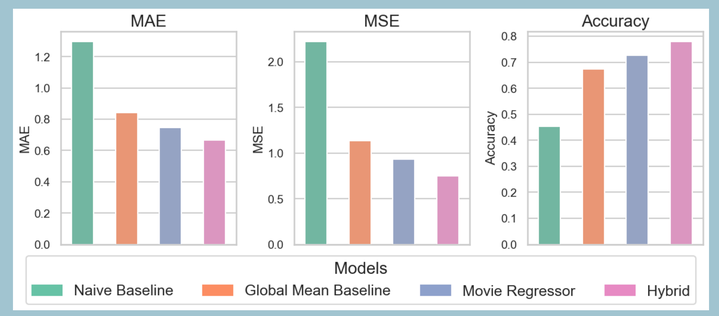

Figure 2. Comparison of performance between different models (Naïve Baseline assigns 2.5 for every movie, while Global Mean Baseline assigns mean of all training data to every movie)

Results description

- **Genre Based Regressor** showed solid performance with mediocre evaluation
- **Cluster Based Regressor** achieved highest accuracy among all models but MSE was not comparable to the best one
- **Movie Based Regressor** had higher MSE and lowest accuracy. Although simple, this model had worst performance
- **Hybrid Regressor** achieved the best performance with lowest MSE and MAE and almost highest Accuracy. This approach successfully highlighted the strengths of each individual model

Results

Number of records in training set vs performance of Hybrid model

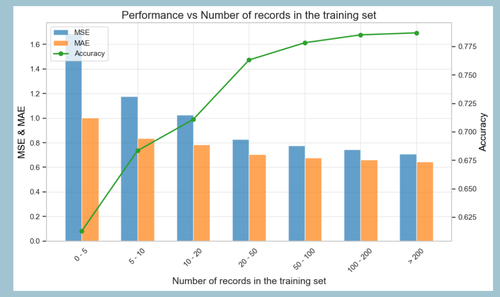

Figure 3. Comparison of evaluation of Hybrid model for different number of records in training set

It can be seen that, the more movies were watched and rated by the user, the better recommendations were obtained for *HybridRegressor* (lower MAE, MSE and higher accuracy).

Number of clusters vs performance of Hybrid model

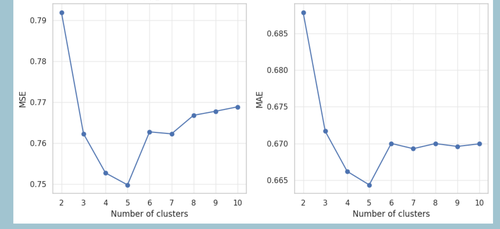

Figure 4. Comparison of MSE and MAE of Hybrid Model for different number of clusters

As we can see, number of clusters in Hybrid approach matters - for k smaller or equal to 5, MAE and MSE will decrease when we increase number of clusters. However, when we set k too high, our results will deteriorate and both MAE and MSE will increase (probably our model overfits then). The best performance was obtained when number of clusters k was set to 5.

Conclusions

During our research, we answered all previously posed questions

- It is possible to predict rating that movie will get from a user, based on his/her previous experience.
- Accuracy that we got for the best model was impressive - for Hybrid model accuracy was equal to 78%, MAE was 0.664 and MSE was 0.75.
- The more data there is in training set (i. e. more movies are watched and rated by user), the better is prediction by the model. Besides, best number of clusters k was derived - it was 5 and increasing it does not improve prediction even making it worse.