K-Means Clustering Based Movie Recommender System

Julia Lorenz, Jakub Kubiak, Mateusz Nowicki, Krzysztof Skrobała, Wojciech Bogacz

Poznan University of Technology
Marii Skłodowskiej-Curie 5, 60-965 Poznan, Poland

Abstract. *The exponential growth of entertainment services has granted unprecedented access to a vast array of content. In the era where that enormous volume of available movies can overwhelm users, movie recommendation systems are crucial to effectively navigate them to the areas of their preferences. This paper introduces a movie recommendation system utilizing K-means clustering based algorithm to predict the rating a user would assign to the movie from the dataset. Our approach demonstrates how clustering can effectively group users based on similar tastes and preferences, facilitating more personalized and accurate recommendations. By utilizing K-means clustering, we aim to enhance the efficiency of the recommendation process, particularly in the context of large and diverse datasets.*
Keywords: *recommender systems, data mining, association rules, clustering, k-means*

1. Introduction

The exponential growth of entertainment services has granted unprecedented access to a vast array of content. In the era where that enormous volume of available movies can overwhelm users, movie recommendation systems are crucial to effectively navigate them to the areas of their preferences. This paper introduces a movie recommendation system utilizing K-means clustering based algorithm to predict the rating a user would assign to the movie from the dataset. Our approach demonstrates how clustering can effectively group users based on similar tastes and preferences, facilitating more personalized and accurate recommendations. By utilizing K-means clustering, we aim to enhance the efficiency of the recommendation process, particularly in the context of large and diverse datasets.

2. Related Work

2.1. Recommender Systems

The article *Movie Recommender Systems: Concepts, Methods, Challenges, and Future Directions* [1] provides overview of various filtering techniques, algorithms and metaheuristic approaches used in recommender systems. Besides that, it also discusses the challenges such systems face and highlights advancements in this field of study and emphasizes the strength of collaborative approaches that combine multiple methods to create a recommender system.

2.2. Collaborative Filtering

Collaborative Filtering [2] is a process of evaluating entities based on assumption that preferences of others who have shown similar behaviours in the past would reflect in the predicted preferences. It has found its application in many web services, due to its flexibility, effectiveness and ability to overcome problems such as sparsity and loss of information.

2.3. Clustering

With aim to explore data multiple clustering methods can be utilized [3]. Pattern recognition is a common application of clustering, widely used in recommender systems to group entities with the same characteristics/preferences.

3. Dataset

In our study we used the dataset collected by the GroupLens Research. This data described ratings of movies on a five star scale from a movie recommendation service MovieLens [4]. It contains approximately 34 million ratings with about 2 million tags across almost 90,000 movies. This is the latest version of the dataset published by this group. We use the full version of this dataset for out calculations.

4. Algorithm

4.1. First Method

Our approach involves several preprocessing steps to ensure meaningful and efficient predictions. Initially, we filter out movies with fewer than five reviews to eliminate statistically insignificant data. We then focus exclusively on positive reviews, selecting those with scores greater than four.

For data organization, we employ a mini-batch strategy, dividing the dataset into 49 batches. This data is represented in a binary matrix where rows correspond to users and columns to movies, indicating whether a user has positively reviewed considered movie.

We utilize k-means clustering to assign users clusters based on their review patterns. Due to the complexity of this operation, we perform an additional looping process instead of simple batching to manage memory constraints effectively.

4.2. Second Method

In the improved approach we did not encode the data into binary table. By grouping the reviews on the movie type, the processed data became much smaller, therefore we could avoid using mini-batch strategy.

At this point rows correspond to specific users and columns to mean rating values assigned to movie category by the user. After all, we have set number of cluster to 14, as this value resulted in the best experimental results.

5. Results

5.1. First Method Results

The results have not been satisfactory. The clustering approach gave worse MSE than simply making the prediction equal to the average review value of considered movie.

The reason, most probably, lays in the usage of K-means. Since the data have been encoded into binary table of reviews, the Euclidean distance used by the algorithm reduces to counting the number of variables on which two cases disagree. This then leads to plenty of situations where the ties occur, and the ties are resolved arbitrarily

5.2. Second Method Results

Our second approach yielded substantially better results. Basing the approach on movie types grouping resulted in vast improvements that, furthermore, grew with the increase of the number of clusters. Landing at the lowest mean squared error value with 14 clusters. This strategy resulted in MSE equal to approximately 0.781 which largely outperformed the previous method and our baseline method that simply assigned a score equal to the movie's average rating.

Approach	**Mean Squared Error**
Average Movie Review	0.933
First Method	1.744
Second Method	0.781

Table 1. Mean Squared Error for Different Approaches

6. Conclusions

Our research demonstrates that a Collaborative Filtering system utilizing K-means clustering and association rules outperforms baseline solutions significantly. But there is always room for improvement. Integrating Content-based Filtering methods or Deep Learning approaches in the context of recommendation systems can lead to significant advancement in performance. Furthermore, the ever-growing volume of data in today's world creates the need for continuous improvement in recommender systems. As user expectations evolve, so too must these systems. By actively exploring cutting-edge techniques, we can ensure that recommendation systems remain relevant and valuable tools in the age of information overload.

References

[1] Jayalakshmi, S., Ganesh, N., Čep, R., and Senthil Murugan, J. Movie recommender systems: Concepts, methods, challenges, and future directions, 2022. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9269752/.

[2] Schafer, J. B., Frankowski, D., Herlocker, J., and Sen, S. Collaborative filtering recommender systems, 1970. URL https://link.springer.com/chapter/10.1007/978-3-540-72079-9_9.

[3] Bindra, K. and Mishra, A. A detailed study of clustering algorithms. In *2017 6th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, pages 371–376. 2017. doi:10.1109/ICRITO.2017.8342454.

[4] Harper, F. M. and Konstan, J. A. The movielens datasets: History and context, 2015. URL https://doi.org/10.1145/2827872.