Recommendation system based on K-means and rule based association.

Maria Musiał 156062
Wiktoria Szarzyńska 156058
Joanna Szczublińska 156070
Lidia Wiśniewska 156063

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

Abstract. *This paper presents a movie recommendation system that uses K-Means clustering and rule-based association learning with the MovieLens dataset. It groups users based on their movie interactions and preferences for different genres. Within each group, it identifies common patterns in genre preferences. The system creates rules to capture these patterns. When predicting a rating, it considers the user's past ratings, the movie's genres, and the user's group to find relevant rules and predict how much the user might like the movie. This method predicts user's movie rating based on user preferences and learned genre relationships.*

Keywords: *recommender systems, data mining, association rules*

1. Introduction

Recommender systems are crucial in suggesting relevant movies to users, especially with the big amount of content available on streaming platforms. The idea is to predict rating for a movie based on user's rating history. There we extract information what are the features of the movies a user likes. Later go through all clusters to make association rules. This hybrid approach will guarantee us satisfactory results when measured using mean absolute error.

We present a movie recommender system that utilizes K-Means clustering to group movie ratings based on the genres. Within each cluster, association rule learning is employed to identify patterns in how users co-watch movies from different genres. These rules capture relationships between genres, allowing the system to recommend movies based not only on a user's past ratings but also on genre combinations they might find satisfactory.

2. Related Work

Several studies have used K-Means clustering in recommender systems to group users or items based on their ratings or features (e.g., Zhang et al., 2010) [1]. This approach allows for personalized recommendations by focusing on users within similar clusters. Association rule learning has also been explored in recommender systems to identify relationships between items and recommend items that users tend to purchase or rate together (e.g., Sarwar & Karypis, 2002)[2]. Our work combines these techniques to leverage both user-movie interactions and movie genre information for recommendations, particularly rating prediction.

3. Dataset

The proposed recommender system is evaluated using the MovieLens dataset, a publicly available collection of movie ratings and user information. Different versions of MovieLens exist, containing varying numbers of users, movies, ratings, and potentially additional information like timestamps or genres. This work utilizes the MovieLens latest dataset, which contains 330,975 users, 86,000 movies, and 33,000,000 ratings. The specific features used in the analysis will be detailed in the Data Preprocessing section.

4. Algorithm

The proposed recommender system employs a two-step approach: movie clustering using K-Means algorithm and rule-based recommendation based on genre co-occurrence patterns within the clusters.

4.1. Data Preprocessing

The data preprocessing step prepares the MovieLens data for clustering and rule learning. We hypothesize that the top 2000 users who rated most movies have the most important data for our model. Then we change how genre is encoded using one-hot encoding. We're also dropping redundant column timestamp. We're going to work with cross validation on 5 splits to ensure best results.

4.2. K-Means Clustering

K-Means clustering is used to segment movies into groups based on their genres. The optimal number of clusters (k) can be determined using techniques like the elbow method, which visually identifies the point where increasing the number of clusters no longer significantly reduces the within-cluster sum of squares (WCSS) for the MovieLens data. In our case its 5.

4.3. Association Rule Learning

Within each movie cluster identified by K-Means frequent itemset mining is performed to identify sets of genres that occure in similar films in each cluster. These frequent itemsets capture patterns in genre co-occurrences within a specific cluster of similar genres. Association rules are then generated from the frequent itemsets.

4.4. Rule-based prediction

To predict a rating of a movie to a particular user, the system considers the user's rated movies and the genres of the movie being recommended. The rules associated with the movie's predicted cluster are then filtered based on the user's past ratings and the movie's genres within the data. These filtered rules represent genre co-occurrence patterns relevant to the user's rated movies. Rating is predicted by taking the median of confidence for rules we got out of filtering times rating of already given rating to similar movies.

5. Results

The findings from our study highlight significant impact of using the K-means clustering algorithm and association rules in improving the accuracy of our recommendation system. The considerable reduction in mean absolute error following the integration of association rules underscores their efficiency in refining recommendation outcomes.

However, there's room for further improvement to enhance the system's effectiveness and user experience. Future work can explore:

Advanced Clustering Techniques: Investigate alternative clustering algorithms like hierarchical or density-based clustering to potentially capture more intricate relations.

Feature Engineering: Capturing more information about things like a year of production of a movie from the title. There could be also some useful information in timestamp, but that would require further analysis and probably some domain knowledge.

6. Conclusions

The findings from our study highlight significant impact of using the K-means clustering algorithm and association rules in improving the accuracy of our recommendation system. The considerable reduction in mean absolute error following the integration of association rules underscores their efficiency in refining recommendation outcomes.

However, there's room for further improvement to enhance the system's effectiveness and user experience. Future work can explore:

Advanced Clustering Techniques: Investigate alternative clustering algorithms like hierarchical or density-based clustering to potentially capture more intricate relations.

Feature Engineering: Capturing more information about things like a year of production of a movie from the title. There could be also some useful information in timestamp, but that would require further analysis and probably some domain knowledge.

References

[1] Zhang, Y., Zhou, M., Liu, Y., and Ouyang, C. A k-means clustering approach to collaborative filtering for movie recommendations.

[2] Sarwar, B. M. and Karypis, G. Item-based collaborative filtering recommendation algorithms.