Hybrid Movie Recommendation Systems: Enhancing Collaborative Filtering with Fuzzy C-Means Clustering and Singular Value Decomposition

Matías Figueroa Contreras
Departamento de Ingeniería Informática
Universidad de Santiago de Chile
Santiago, Chile
matias.figueroa.c@usach.cl

***Abstract*—Recommender systems play a critical role in filtering large volumes of information and delivering personalized suggestions to users. However, challenges such as data sparsity and computational scalability remain significant obstacles. This paper proposes a hybrid recommendation system that integrates Fuzzy C-Means (FCM) clustering with Alternating Least Squares (ALS) based Singular Value Decomposition (SVD) and Pearson correlation. FCM is employed to group users based on their rating patterns, enhancing neighbor selection and reducing computational costs. SVD with ALS is utilized to address data sparsity by reconstructing missing ratings and uncovering latent user-item interactions. Pearson correlation further refines predictions by identifying the most similar neighbors within clusters. Experimental evaluations conducted on the MovieLens ml-100k dataset demonstrate the proposed method's effectiveness, achieving competitive performance in terms of Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) compared to traditional methods. However, the results also highlight the limitations of the proposed approach when compared to more recent methods based on deep learning, which achieve significantly better accuracy by leveraging neural networks to model complex user-item interactions. This work underscores the potential of combining clustering, dimensionality reduction, and similarity measures to enhance recommendation systems while identifying opportunities for future integration of advanced machine learning techniques.**

***Index Terms*—Recommender systems, Collaborative filtering, Hybrid recommendation, Fuzzy C-Means clustering, Singular Value Decomposition (SVD), Alternating Least Squares (ALS), Pearson correlation**

I. Introduction

Recommender systems play a crucial role in helping users navigate large volumes of information by suggesting items that align with their preferences. These systems are widely used in domains such as e-commerce, streaming services, and online learning platforms. Among various recommendation techniques, collaborative filtering has emerged as one of the most effective methods due to its ability to leverage user interactions to generate personalized recommendations [1].

Despite its success, collaborative filtering faces significant challenges. One primary issue is the sparsity problem, where most users rate only a small subset of items, leading to difficulties in accurately identifying similar users or items. Another critical challenge is ensuring scalability and computational efficiency, especially in large datasets. Additionally, improving recommendation accuracy, measured through metrics such as Mean Absolute Error (MAE) and Root Mean Square Error (RMSE), remains a constant goal for researchers.

To address these challenges, hybrid approaches appear combining clustering and dimensionality reduction. Clustering methods, such as Fuzzy C-Means (FCM), aim to group users with similar preferences, reducing the computational burden and enhancing the quality of neighbor selection. Simultaneously, dimensionality reduction techniques like Singular Value Decomposition (SVD) with Alternating Least Squares (ALS) tackle data sparsity by uncovering latent factors that explain user-item interactions. Unlike traditional SVD, ALS iteratively optimizes user and item matrices while incorporating regularization, ensuring robustness against sparse and noisy data [2].

This paper proposes a hybrid recommendation system that combines Fuzzy C-Means clustering with regularized Singular Value Decomposition (SVD) and Pearson correlation to improve movie recommendations. The FCM algorithm partitions users into clusters based on their rating patterns, ensuring that similar users are grouped together. Within each cluster, Pearson correlation identifies the most similar neighbors, enhancing the accuracy of predictions. Meanwhile, SVD with ALS addresses data sparsity by iteratively reconstructing missing ratings and uncovering latent features, ensuring convergence and stability in the factorization process. The proposed approach is evaluated using the MovieLens ml-100k dataset, a benchmark dataset widely used in recommender system research due to its structured format and comprehensive user-item interactions.

The main contributions of this work are as follows. It introduces a novel hybrid methodology that integrates Fuzzy C-Means (FCM) clustering with Singular Value Decomposition (SVD) and Pearson correlation, effectively addressing data sparsity and improving neighbor selection. The approach

enhances movie recommendation accuracy, achieving better MAE and RMSE metrics compared to traditional collaborative filtering methods. Additionally, by applying Pearson correlation within FCM clusters, the computational cost of similarity calculations is significantly reduced. Comprehensive experiments on the MovieLens ml-100k dataset validate the system's effectiveness in generating accurate and efficient recommendations.

The remainder of this paper is organized as follows: Section 2 reviews related work on collaborative filtering, clustering, and dimensionality reduction techniques. Section 3 describes the proposed methodology, including the integration of Fuzzy C-Means, SVD, and Pearson correlation. Section 4 presents the experimental setup, results, and analysis. Finally, Section 5 concludes the paper and discusses future research directions.

II. Related Work

Collaborative Filtering (CF) remains one of the most widely utilized recommendation algorithms, aiming to uncover latent preferences of users and items by analyzing historical interactions. Despite its effectiveness, CF often suffers from sparsity in the rating matrix, where a majority of user-item pairs are unrated. This excessive sparsity undermines prediction accuracy and overall recommendation performance. To address these challenges, researchers have proposed hybrid recommendation algorithms that combine multiple methods to leverage their respective strengths.

The hybrid approach aims to integrate different recommendation strategies to compensate for the limitations of individual methods while enhancing their complementary advantages. Ferdaous et al. [3] proposed a hybrid algorithm that linearly combines content-based item similarity with neighborhood-based similarity, improving interpretability and addressing sparsity. Similarly, Li et al. [4] introduced a hybrid system that integrates Fuzzy C-Means (FCM) clustering with supervised learning, combining subjective and objective membership degree vectors to improve prediction accuracy on sparse datasets. The algorithm demonstrated significant improvements in MAE and RMSE on the MovieLens dataset. Furthermore, matrix factorization techniques, including SVD, and dimensionality reduction methods, such as autoencoders [5], have been successfully employed to extract latent features, addressing sparsity and enhancing prediction quality.

Clustering techniques have emerged as a vital component in hybrid systems to overcome the limitations of traditional CF. Koohi and Kiani [6] demonstrated that Fuzzy C-Means (FCM) clustering can significantly enhance user-based CF by grouping users with similar preferences, thus improving recommendation accuracy and scalability. Vimala and Vivekanandan [7] further refined FCM by introducing Kullback-Leibler divergence-based clustering, which increased the stability and robustness of clusters, leading to superior recommendation performance. Noor et al. [8] combined FCM clustering with sparsity removal techniques, leveraging genre and rating similarities to fill missing data and create denser matrices, resulting in better computational efficiency and prediction accuracy.

There were also advancements in deep learning that have significantly expanded the potential of hybrid recommendation systems. Lund and Ng [9] introduced a deep learning-based recommendation approach that leverages autoencoders to address sparsity and improve prediction accuracy in collaborative filtering systems. Their method utilizes a neural network architecture with multiple hidden layers to learn a compressed representation of the user-item matrix, effectively capturing underlying patterns in the data.

This study builds upon these advancements by integrating FCM clustering, ALS-based SVD, and Pearson correlation within a user-based CF approach. By addressing data sparsity, improving computational efficiency, and enhancing recommendations, this work contributes a novel perspective to the evolving landscape of hybrid recommendation systems.

III. Proposed Approach and Implementation

This section details the proposed hybrid recommendation approach, which integrates Fuzzy C-Means (FCM) clustering, Alternating Least Squares (ALS)-based Singular Value Decomposition (SVD), and Pearson correlation to address the challenges of data sparsity and enhance recommendation accuracy. The proposed approach is designed to leverage the strengths of clustering, dimensionality reduction, and similarity measures, ensuring effective and scalable recommendations.

A. Fuzzy C-Means Clustering (FCM)

Fuzzy C-Means (FCM) clustering is utilized to partition users into overlapping groups based on their rating patterns. Unlike traditional hard clustering methods that assign users to a single cluster, FCM allows users to belong to multiple clusters with varying degrees of membership. This flexibility captures subtle relationships between users, ensuring a more accurate representation of their preferences.

The FCM algorithm clusters users based on their rating patterns and assigns membership degrees that indicate the strength of association between users and clusters. This iterative process ensures that users with similar preferences are grouped together and enables effective neighbor selection for collaborative filtering. By dynamically refining cluster assignments, FCM reduces computational complexity by limiting similarity calculations to users within the same clusters, facilitating better neighbor selection for collaborative filtering.

B. Singular Value Decomposition with ALS (ALS-SVD)

Singular Value Decomposition (SVD) with Alternating Least Squares (ALS) is employed to address the sparsity issue in the user-item rating matrix. ALS-based SVD decomposes the rating matrix into two lower-dimensional matrices: user latent factors and item latent factors.

Latent factors are abstract features that represent hidden patterns or relationships in the data. In the context of this implementation, latent factors capture the preferences of users for certain movies based on the ratings they have provided. For example, a latent factor might reflect a user's affinity for

action or drama movies, inferred from their rating patterns, while item latent factors represent how strongly a movie aligns with such preferences.

The ALS algorithm alternates between optimizing the user and item latent factor matrices iteratively. This process ensures robust reconstruction of missing ratings, effectively addressing data sparsity. By capturing latent factors, ALS-based SVD provides a compact representation of user-item interactions, making it well-suited for handling large-scale datasets.

C. Pearson Correlation

Pearson correlation is used within each cluster to measure the similarity between users, enabling the identification of the most relevant neighbors for generating recommendations. Pearson correlation quantifies the linear relationship between users rating vectors, comparing their deviations from their respective average ratings. This similarity measure helps identify the most relevant neighbors for generating recommendations. By focusing on intra-cluster similarities, the approach ensures that predictions are generated using the most pertinent user subsets.

D. Overall Recommendation Process

The hybrid recommendation approach follows a multi-step process to generate personalized recommendations. Initially, the sparse user-item rating matrix is preprocessed, with missing values initialized.

Next, Singular Value Decomposition (SVD) with Alternating Least Squares (ALS) is applied to the entire matrix. This global application of ALS-SVD reduces data sparsity by reconstructing missing ratings through the learning of latent factors. The resulting dense matrix provides a robust input for the subsequent clustering step, capturing broad patterns in user-item interactions.

The processed matrix is then utilized to group users into clusters using Fuzzy C-Means (FCM) clustering, allowing each user to belong to multiple clusters with varying degrees of membership. By grouping similar users, FCM ensures that subsequent calculations are more focused and computationally efficient.

Within each cluster, ALS-SVD is re-applied to refine predictions further. This localized application of SVD tailors the reconstruction of missing ratings to the preferences of users within the cluster, capturing more specific patterns. This step enhances the quality of predictions by focusing on the preferences of similar users.

Pearson correlation is then used to compute similarity scores between users within each cluster. This step identifies the top-k neighbors for each user, those whose preferences most closely align with the target user's preferences.

Finally, the reconstructed ratings from ALS-SVD are combined with weighted ratings derived from the top-k neighbors. For the top-k neighbors, their ratings are averaged to produce a representative score. This score is then combined with the ratings reconstructed by ALS-SVD using a simple average. By blending these two components, the system ensures that the final predictions effectively capture more global patterns identified by ALS-SVD in clusters and the localized preferences reflected in the neighbors ratings.

Once the predictions are generated, the system can produce TOP-N recommendations for any user. This is achieved by filtering out movies the user has not yet rated and selecting the ones with the highest predicted ratings. This approach ensures that users receive recommendations tailored to their inferred preferences, prioritizing movies they are likely to enjoy.

The entire workflow, including sparsity removal, clustering, and recommendation generation, is illustrated in Figure 1. This structured approach leverages the strengths of FCM, ALS-SVD, and Pearson correlation to overcome the challenges of sparsity and scalability, ultimately improving recommendation accuracy.

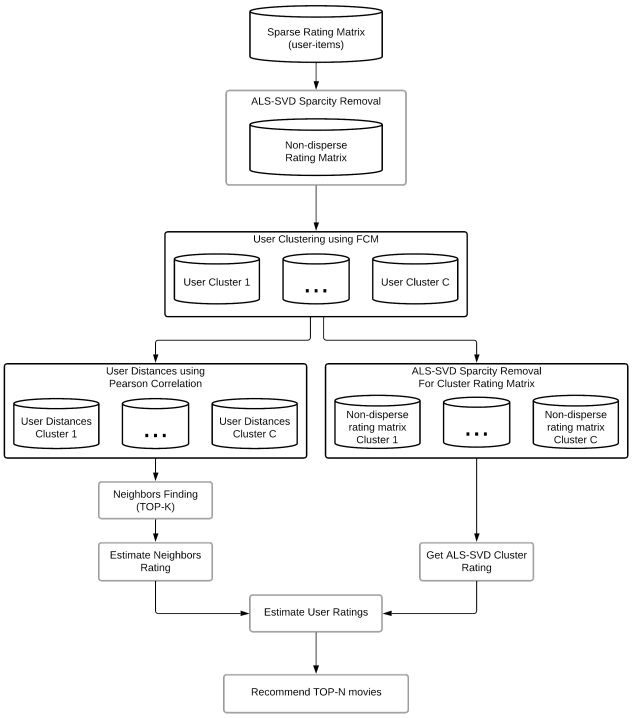

Fig. 1. Workflow of the proposed movie recommendation system.

IV. Experimental Evaluation

To validate the effectiveness of the proposed hybrid recommendation approach, experiments were conducted using the MovieLens ml-100k dataset provided by the GroupLens Research Group [10]. This dataset is widely recognized as a benchmark for evaluating recommendation systems, containing 100,000 ratings from 943 users on 1,682 movies. The ratings in the dataset range from 1 to 5, representing the users level of preference, with higher values indicating stronger preferences. The dataset also includes pre-defined splits for training and testing, facilitating direct comparison with other methods.

A. Experimental Setup

The evaluation employed five-fold cross-validation, as defined by the dataset. Each fold contained 80% of the data for training and 20% for testing, ensuring consistency with prior research and comparability across methods. The proposed approach utilized the following parameters:

- **Number of clusters (FCM):** 2
- **Number of neighbors (TOP-K):** 200
- **Latent factors (ALS-SVD General):** 9
- **Latent factors (ALS-SVD Cluster-Specific):** 2

B. Results

The performance of the proposed method was measured using two standard metrics:

- **Mean Absolute Error (MAE):** Captures the average magnitude of prediction errors. It is calculated as:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

  where y_i is the actual rating, $\hat{y}_i$ is the predicted rating, and N is the total number of predictions.
- **Root Mean Square Error (RMSE):** Provides an error measure that penalizes larger deviations. It is calculated as:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$$

  Like MAE, y_i represents the actual rating, $\hat{y}_i$ the predicted rating, and N the total number of predictions.

The table I compares the results of the proposed approach with benchmark algorithms reported in the literature (extracted from Li et al [4] and Lund and Ng [9]). The comparison includes seven baseline algorithms: UCF, which uses collaborative filtering based on user neighbors; ICF, which applies collaborative filtering based on item neighbors; Clust, a recommendation algorithm that employs user clustering; CRF, which integrates collaborative filtering with clustering and random forest techniques; FCM-SL, a supervised learning recommendation algorithm based on Fuzzy C-Means clustering; HCFCM-SL, a hybrid supervised learning recommendation algorithm that combines content and Fuzzy C-Means clustering; and Deep Learning, which employs autoencoders to capture latent patterns in user-item interactions, leveraging a deep neural network architecture to address data sparsity and improve prediction accuracy. Each of these approaches represents a distinct strategy for addressing challenges in collaborative filtering, providing a benchmark for evaluating the proposed method.

TABLE I
COMPARISON OF MAE AND RMSE ON ML-100K DATASET

Algorithms	MAE	RMSE
UCF	0.7539	0.9600
ICF	0.7571	0.9637
Clust	0.8081	1.0215
CRF	0.7582	0.9534
FCM-SL	0.7540	0.9582
HCFCM-SL	0.7490	0.9504
Deep Learning	0.1500	0.3544
Proposed Approach	0.7405	0.9396

C. Analysis of Results

The proposed approach achieved a MAE of 0.7405 and RMSE of 0.9396, outperforming all benchmark algorithms that do not utilize deep learning. These results highlight the effectiveness of the hybrid integration of FCM clustering, ALS-SVD, and Pearson correlation in addressing challenges such as data sparsity and computational efficiency. However, the Deep Learning method achieved significantly lower MAE (0.15) and RMSE (0.3544), demonstrating the potential of neural network-based approaches for recommendation tasks.

The impact of FCM clustering is evident in its ability to group users with similar preferences and focus similarity calculations within clusters. This approach reduces noise and enhances neighbor selection, enabling more precise recommendations. By capturing subtle relationships among users, FCM clustering contributes significantly to improving prediction accuracy.

ALS-SVD further enhances the model by addressing sparsity in the rating matrix. Its application both globally and within clusters provides a robust mechanism for reconstructing missing ratings. The dual use of ALS-SVD allows the approach to leverage both broad and localized patterns in user-item interactions, ensuring more accurate predictions across diverse datasets.

The role of Pearson correlation is critical in refining the recommendation process. By calculating similarities within clusters, Pearson correlation ensures that only the most relevant neighbors contribute to the rating predictions. This targeted approach enhances the reliability of the recommendations, reducing the impact of irrelevant neighbors.

While the proposed hybrid method balances computational efficiency and prediction accuracy effectively, the significantly better results of the Deep Learning approach suggest that deep neural networks may be better suited for handling large-scale datasets and capturing complex non-linear relationships in user-item interactions.

These findings underscore the strengths and limitations of both approaches. While the hybrid method excels in balancing computational cost and accuracy, the superior performance of the Deep Learning method highlights its potential for future research, particularly in scenarios where computational resources are less constrained.

V. CONCLUSION AND FUTURE WORK

In this study, a hybrid recommendation approach was proposed to address the challenges of data sparsity and scalability in collaborative filtering. By integrating Fuzzy C-Means (FCM) clustering, Alternating Least Squares (ALS) based Singular Value Decomposition (SVD), and Pearson correlation, the approach effectively combines clustering, dimensionality

reduction, and similarity measures to enhance recommendation accuracy and efficiency.

The experimental results demonstrated that the proposed approach outperforms traditional state of the art methods that do not utilize deep learning on the MovieLens ml-100k dataset, achieving competitive Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) values. These improvements can be attributed to the combined strengths of FCM in grouping users with similar preferences, SVD in reconstructing missing ratings, and Pearson correlation in refining neighbor selection within clusters. The use of both global and cluster-specific SVD ensured a comprehensive understanding of user-item interactions, while the integration of FCM reduced noise and computational complexity.

However, the results also highlight the significant potential of deep learning approaches, as the Deep Learning algorithm evaluated in this study achieved substantially lower MAE and RMSE. This underscores the value of neural network-based models in capturing complex non-linear patterns and suggests that hybrid methods could benefit from incorporating deep learning techniques.

There are areas for future improvement. First, exploring hybrid models that integrate the strengths of traditional methods, such as clustering and SVD, with deep learning architectures like autoencoders or neural collaborative filtering could further enhance performance. Second, investigating alternative clustering methods, such as deep clustering or graph-based clustering, could provide a more nuanced representation of user preferences. Third, incorporating additional contextual information, such as temporal dynamics, item attributes, or user demographics, could improve personalization and adaptability of the recommendation systems. Finally, extending the evaluation to larger and more diverse datasets, such as MovieLens 1M or Netflix Prize, would validate the scalability and generalizability of the approach.

Overall, this work contributes a novel perspective to hybrid recommendation systems, offering a robust and scalable solution for enhancing the accuracy and efficiency of collaborative filtering. While the proposed approach demonstrates strong performance, the insights gained suggest that integrating deep learning techniques represents a promising direction for future research, potentially leading to even more powerful and adaptable recommendation systems.

REFERENCES

[1] Y. Koren, S. Rendle, and R. Bell, *Advances in Collaborative Filtering*. New York, NY: Springer US, 2022, pp. 91–142.

[2] T. Zhao, "Performance comparison and analysis of svd and als in recommendation system," *Applied and Computational Engineering*, vol. 49, pp. 142–148, 03 2024.

[3] F. H., B. F., B. O. *et al.*, "Recommendation using a clustering algorithm based on a hybrid features selection method," *Journal of Intelligent Information Systems*, vol. 51, pp. 183–205, 2018.

[4] L. Duan, W. Wang, and B. Han, "A hybrid recommendation system based on fuzzy c-means clustering and supervised learning," *KSII Transactions on Internet and Information Systems*, vol. 15, no. 7, pp. 2399–2413, 2021.

[5] B. J., A. L. G. M., B. F. *et al.*, "Autoencoders and recommender systems: Cofils approach," *Expert Systems with Applications*, vol. 89, pp. 81–90, 2017.

[6] K. H. and K. K., "User based collaborative filtering using fuzzy c-means," *Measurement*, vol. 91, pp. 134–139, 2016.

[7] S. V. Vimala and K. Vivekanandan, "A kullback–leibler divergence-based fuzzy c-means clustering for enhancing the potential of an movie recommendation system," *SN Applied Sciences*, vol. 1, no. 698, 2019.

[8] N. Ifada, E. H. Prasetyo, and Mula'ab, "Employing sparsity removal approach and fuzzy c-means clustering technique on a movie recommendation system," in *2018 International Conference on Computer Engineering, Network and Intelligent Multimedia (CENIM)*, 2018, pp. 329–334.

[9] J. Lund and Y.-K. Ng, "Movie recommendations using the deep learning approach," in *2018 IEEE International Conference on Information Reuse and Integration (IRI)*, 2018, pp. 47–54.

[10] GroupLens, "MovieLens 100K Dataset," 3 2021. [Online]. Available: https://grouplens.org/datasets/movielens/100k/