A Hybrid Recommendation System based on Fuzzy C-Means Clustering and Supervised Learning

Li Duan[1], Weiping Wang[2,3,4,5*], and Baijing Han[2,3,4,5]
[1]Beijing Key Laboratory of Security and Privacy in Intelligent Transportation, Beijing Jiaotong University, Haidian, 100044, China
[e-mail: duanli@bjtu.edu.cn]
[2]School of Computer and Communication Engineering, University of Science and Technology Beijing
[3]Beijing Key Laboratory of Knowledge Engineering for Materials Science, Beijing 100083, China
[4]Institute of Artificial Intelligence, University of Science and Technology Beijing, Beijing 100083, China
[5]Shunde Graduate School, Beijing University of Science and Technology, Guangzhou 528399, China
[e-mail: weipingwangjt@ustb.edu.cn]
* Corresponding author: Weiping Wang

Received February 1, 2021; accepted June 24, 2021; published July 31, 2021

Abstract

A recommendation system is an information filter tool, which uses the ratings and reviews of users to generate a personalized recommendation service for users. However, the cold-start problem of users and items is still a major research hotspot on service recommendations. To address this challenge, this paper proposes a high-efficient hybrid recommendation system based on Fuzzy C-Means (FCM) clustering and supervised learning models. The proposed recommendation method includes two aspects: on the one hand, FCM clustering technique has been applied to the item-based collaborative filtering framework to solve the cold start problem; on the other hand, the content information is integrated into the collaborative filtering. The algorithm constructs the user and item membership degree feature vector, and adopts the data representation form of the scoring matrix to the supervised learning algorithm, as well as by combining the subjective membership degree feature vector and the objective membership degree feature vector in a linear combination, the prediction accuracy is significantly improved on the public datasets with different sparsity. The efficiency of the proposed system is illustrated by conducting several experiments on MovieLens dataset.

Keywords: Recommendation, Collaborative filtering, Clustering, Supervised learning.

1. Introduction

In the era of big data and artificial intelligence, network information resources are increasing quickly, which makes it costly for users to find information services that they are interested in. As an information filter, the recommendation system aims to filter out information services which are not related to user's interests and preferences, and thus mitigating information overload. Recommendation algorithms are generally divided into three categories: content-based recommendation, collaborative filtering, and hybrid recommendation [1].

In the early days of text information services, the content-based recommendation system aims to mine the user's interests and recommends items that meet the user's interest keywords to the user, the method is to use Term Frequency Inverse Document Frequency (TF/IDF) to learn keywords in the user profile. However, keywords have many different meanings at the semantic level, and users' interests cannot be accurately extracted without further analysis. Moreover, users' interests are not always clear, and effective keywords can't be extracted. Content-based recommendation method cannot fulfill the needs of users when keywords are incomplete.

The content-based recommendation algorithm [2] is commonly used in the recommendation system at the beginning. It relies on the historical behavior data of users and the content attribute information of the item. According to the similarity of the content attribute information of each item and the record of items browsed by the user, it recommends items that the user is most likely to like in the future. The recommendation result is simple, intuitive, and has good interpretability. However, the content-based recommendation algorithm cannot mine the implicit preferences. At the same time, when the target user has not yet commented on any items, the content-based recommendation will not be able to complete the recommendation task to the user due to the lack of historical behavior data. That is to say, the content-based recommendation algorithm has the cold start problem of users [3].

Compared with the content-based recommendation algorithm, the collaborative filtering (CF) recommendation algorithm [4] has a good ability to mine the implicit preferences. There are two kinds of CF recommendation algorithms: one is memory-based CF, the other is model-based CF. The former is divided into user-based CF and item-based CF. Memory-based CF [5] calculates the similarity between users or items, predicts the score of an item according to the score of the target user's nearest neighbor set, or find the neighbor set similar to the target item in the scored items, and predict the target score of the users on the target item based on the score value of each item in the neighbor set. When the number of users or items in the recommendation system increases rapidly, the CF recommendation algorithm based on the nearest neighbor needs to deal with the huge similarity calculation tasks, which results in the scalability problem of the recommendation system. The CF recommendation algorithm based on clustering model [6] divides users (or items) into disjoint clusters by the clustering analysis of user-item rating matrix. When recommending target users, it only searches all users in the cluster of target users, thus reducing the search space and reducing the calculation amount of similarity. However, due to the reduction of search space, the accuracy of finding similar users will also be reduced, and the sparsity of data may be increased, which will lead to the decline in the accuracy of scoring and prediction of the algorithm [7].

The existing algorithm [8] combines the clustering algorithm with the supervised learning algorithm. Through the clustering algorithm, the original user-item score matrix is clustered to obtain the cluster center of the user (or item), and then the distance between each sample point and each cluster center is obtained through a certain distance calculation method, as the membership degree of the sample belonging to each cluster center, the user membership

feature vector and the article membership feature vector are constructed, and the user membership degree feature vector and item membership feature vector finally are connected, so as to construct a supervised learning model to input sample space. When a user is a newly added user, he has not commented on any item, or an item is newly added and it has not been evaluated by any user. In this case, the method of constructing the membership degree by calculating the distance between the sample point and the cluster center often has the problem that the distance cannot be calculated.

In order to solve the cold start problem of traditional supervised learning hybrid recommendation algorithm based on clustering and distance calculation, this paper proposes a content-based recommendation algorithm by means of fuzzy C-means (FCM) clustering.

The FCM clustering algorithm [9] is used to directly calculate the membership matrix of users and items, and construct the feature vector of user membership and item membership feature vector. At the same time, the content-based recommendation algorithm is mixed, and the subjective membership vector and the objective membership vector of the linear combination of items are further improved to further improve the prediction accuracy of the algorithm. Specifically, our contributions are summarized up as follows:

(1) The service recommendation scheme FCM-ML is proposed, it not only considers content information, but also resolves the cold start problem. FCM clustering technique has been applied to the item-based collaborative filtering framework to solve the cold start problem.

(2) The content information is integrated into the collaborative filtering. The algorithm constructs the user and item membership degree feature vector, and adopts the data representation form of the scoring matrix to the supervised learning algorithm, as well as by combining the subjective membership degree feature vector and the objective membership degree feature vector in a linear combination.

(3) The detailed experiments on MovieLens data set are conducted to verify the effectiveness of our recommendation scheme. Firstly, we carry out various parameters that have influence on the recommendation effect, and then compare the proposed scheme with other baseline algorithms in recommendation performance and prediction performance, the experimental results show that the proposed scheme is superior to the baseline algorithm in recommendation accuracy and prediction accuracy.

The rest of this paper is organized as follows: In Section 2, the related work on service recommendation is shown. In Section 3, we introduce the overall architecture design of the proposed service recommendation scheme HCFCM-SL. In Section 4, we compare the accuracy and efficiency of the proposed recommendation scheme with other recommendation solutions through the detailed experiments. Conclusions of this paper are provided in Section 5.

2. Related Work

The CF recommendation algorithm is the most mainstream recommendation algorithm. The goal of the algorithm is to mine the hidden preferences of users and items from the rating matrix, extract the potential features of users and items, and predict the missing values in the rating matrix. However, the scoring matrix is usually very sparse, that is, the proportion of the user's evaluation of the item recorded in the entire scoring matrix is very small. The CF recommendation algorithm that relies on the scoring matrix as the original input data tends to have a negative impact on the prediction accuracy of the algorithm due to the excessive sparsity of the scoring matrix. In order to overcome the data sparsity problem of a single CF recommendation algorithm, researchers have proposed many hybrid recommendation

algorithms [10][11][12].

Hybrid recommendation algorithm is a method of combining several recommendation algorithms. By combining different recommendation algorithms, it can make up for the shortcomings of a single recommendation algorithm and complement the advantages for each other. Ferdaous et al., "[13] proposed a more interpretable hybrid recommendation algorithm by linearly combining the item similarity matrix based on content recommendation and the similarity matrix based on item neighbor. In order to solve the problem of data sparsity. A hybrid recommendation algorithm was proposed in [14], which based on user clustering and rating matrix filling collaborative filtering technology. The method of hybrid supervised learning and CF recommendation algorithm is also an effective way to solve data sparsity. This method is based on a scoring matrix, extracts the user feature vector and the item feature vector from the two perspectives of the user and the item, and then concatenates the user and item feature vectors. The concatenated vector represents the user-item pair feature representation (Feature Representation), all users-the film became the input feature space of the supervised learning algorithm for Zhang [15], and finally used the existing score value of the score matrix as the sample label, and input it into the regression prediction model for training and prediction. The combination of hybrid supervised learning and CF recommendation algorithm changes the form of the traditional CF recommendation algorithm using the scoring matrix through data conversion, and uses Zhang Cheng's input feature space for regression prediction, which can solve the data sparsity problem of the scoring matrix to a certain extent. The SVD matrix factorization method and autoencoder dimensionality reduction method were used in [16] and [17] to extract user and item feature vectors, and mix supervised learning and CF recommendation algorithms with the idea of dimensionality reduction, then obtain satisfactory results. By constructing the membership degree matrix, according to the clustering analysis method, the user and item membership vectors are used as the user and item feature vectors, and the supervised learning and CF recommendation algorithms are mixed with the idea of clustering.

The core step of the hybrid supervised learning and CF recommendation algorithm based on the idea of clustering is to construct the user and item membership vector. The traditional method is based on a certain distance calculation formula by sequentially calculating the distance between each sample point and each cluster center. But due to the sparseness of the scoring matrix, when a user has not yet evaluated any item; or an item has not been evaluated by any user, there will be a problem that the membership based on distance is not available. That is, the cold start of the user or the item will result in the user or the item's membership vector construction method being inconsistent with other users or items, then affecting the accuracy of the algorithm.

In this research, we use the FCM clustering algorithm to directly construct the user and item membership vector, thus solving the problem of cold start based on distance calculation of membership is not available, and by further mixing content-based recommendation, linear combination of subjective items attribute characteristics and objective attribute characteristics further improve the prediction accuracy of the algorithm. This paper compares the average absolute error (MAE) and root mean square error (RMSE) of the HFCCM-SL algorithm and existing algorithms in related fields on the two movie public data sets ml-100k and ml-1m. Through experimental results and analysis, the superiority of the HFCCM-SL algorithm is verified.

In theory, PLSA model can be regarded as the probability version of LSI, and it is a generation model of soft clustering. Wang et al., "[18] proposed a graph-based PLSA model, the topic of the text was discovered by PLSA model, and then the topic was mapped to an undirected graph,

and the relevance of the topic was managed by the difference between discrete probabilities, so as to better understand the semantic composition of the text.
Compared with PLSA, LDA model is a topic model with completely generated semantics, which is sometimes used as a tool to reduce the dimension of data.
In addition, Unger et al., "[19] put forward context-Aware Recommendation systems based on deep learning frameworks, Liu et al., "[20] proposed a hybrid neural recommendation with joint deep representation learning of ratings and reviews. Khan et al., "[21] proposed a joint deep recommendation model exploiting reviews and metadata information. Matrix decomposition is often used to predict scores in recommendation systems. Jiao et al., "[21] \cite{jiao2019novel} proposed an SVD++ algorithm based on adaptive learning rate, By using adaptive learning rate, the convergence of matrix decomposition algorithm can be accelerated, and the time cost can be reduced without affecting the scoring performance.
Early researchers often regard recommendation as the fitting problem of user rating. Although matrix decomposition performs well in rating prediction, the effectiveness of these models in generating recommendation list should be further explored. Scoring is actually only one part of recommendation, when new users have no scoring behavior but have calling service behavior, semantic feature analysis of service is more important than service scoring. Because NMF has no negative value, some scholars have applied it to the field of feature extraction. Mutinda et al., "[22] put forward the method of combining NMF and Holt-Winters to recommend the link service of sequences, the recommendation idea is to extract features by using NMF, then capture the changes of features with time by using Holt-Winters method, and finally recommend the link service of sequences to users by using unchanged features. Unlike this work, in this paper, we propose a hybrid recommendation based on combining content and the FCM clustering algorithm.

3. The Proposed Recommendation Algorithm Scheme

The existing algorithms based on clustering and supervised learning mainly construct user and item membership vectors by calculating the distance between each sample point and the cluster center. However, due to the sparsity of the scoring matrix, direct distance calculation methods often encounter cold start sxual issues. This paper improves on this point. The proposed HCFCM-SL algorithm uses the FCM clustering algorithm to solve the membership matrix iteratively to construct the user and item membership vectors in an attempt to reduce the impact of data sparsity on the performance of the algorithm.
Through the equation (4) and the equation (5), we can find that the FCM clustering algorithm iteratively solves the fuzzy membership matrix and each clustering center, the HCFCM-SL algorithm can directly return to the equation (1). The membership matrix can directly perform data conversion and construct the input feature space, which solves the cold start problem that the membership distance cannot be calculated in the mixed recommendation process of clustering algorithm and supervised learning.

3.1 FCM algorithm

FCM algorithm is a fuzzy clustering algorithm based on partition. The FCM clustering algorithm uses the degree of membership to indicate the degree to each sample point belongs to each cluster. The so-called degree of membership refers to the degree to which an object x belongs to the set A. It is usually recorded as $\mu_A(x)$, and the value range of $\mu_A(x)$ is $[0,1]$,

which is $0 \le \mu_A(x) \le 1$. When $\mu_A(x) = 1$ is equal to object $x \in A$. Assuming that there are a total of n sample points, after clustering by the FCM algorithm, c clustering center vectors and an $n \times c$ dimensional fuzzy membership matrix U will be output. The fuzzy membership matrix U satisfies:

$$\sum_{j=1}^{c} \mu_{ij} = 1, \forall i = 1, \ldots n \tag{1}$$

where, u_{ij} means that the sample point i belongs to the membership degree of the cluster where the cluster center j is located, and $u_{ij} \in [0,1]$. The FCM algorithm has two important parameters, one is the number of cluster centers c, and the other is the membership factor m, which controls the flexibility of the algorithm. If parameter m is too large, the clustering effect will be poor; if m is too small, the clustering effect will be closer to K-means clustering. The generalized form of the objective function of the FCM algorithm is:

$$J(U, c_1, \ldots c_c) = \sum_{j=1}^{c} J_j = \sum_{j=1}^{c} \sum_{i}^{n} u_{ij}^m d_{ij}^2 \tag{2}$$

then, d_{ij} represents the Euclidean distance between the j cluster center and the i sample point. Using Lagrangian multiplier method, according to the constraint condition equation (1) and the objective function equation (2), a new objective function is obtained:

$$\begin{aligned} \bar{J}(U, c_1, \ldots c_c, \lambda_1, \ldots \lambda_n) &= J(U, c_1, \ldots c_c) + \sum_{i=1}^{n} \lambda_i (\sum_{j=1}^{c} u_{ij} - 1) \\ &= \sum_{j=1}^{c} \sum_{i}^{n} u_{ij}^m d_{ij}^2 + \sum_{i=1}^{n} \lambda_i (\sum_{j=1}^{c} u_{ij} - 1) \end{aligned} \tag{3}$$

Finding the derivatives of variables u_{ij} and c_j in the equation (3) to obtain the extreme value of the objective function. Take the derivative of u_{ij}, we have

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{d_{ij}}{d_{ik}} \right)^{2/(m-1)}} \tag{4}$$

Taking the derivative of c_j to get：

$$c_j = \frac{\sum_{i=1}^{n} u_{ij}^m x_i^T}{\sum_{i=1}^{n} u_{ij}^m} \tag{5}$$

Where x_i is the i sample data. By calculating u_{ij} and c_j, it is found that the two are related to each other. Therefore, when the FCM algorithm is executed, it will first initialize u_{ij} or c_j, and then iterate repeatedly to make the objective function J gradually stabilize. The algorithm description of FCM clustering is shown in **Table 1**:

Table 1. Steps of FCM clustering algorithm.

Algorithm1：FCM Clustering
Input: cluster center number c, weighting index m, iteration termination threshold δ. Output: C cluster center vectors, membership matrix $U_{n\times c}$. BEGIN （1）Randomly initializing the membership matrix U with a value between 0 and 1, which satisfies the formula (1); （2）Calculating c cluster center by means of the formula (5); （3）Using the formula (2) to calculate the objective function J. If the two changes of J are less than the termination threshold δ, the algorithm stops; otherwise, continue to perform the step (4); （4）using the formula (4) to calculate the membership matrix U; END

3.2 HCFCM-SL algorithm

The method of constructing the membership degree vector of users and items based on FCM clustering analysis can solve the problem of user cold start or item cold start cannot obtain the solution membership degree. From the item perspective, we find that each item belongs to each category. Because the item membership matrix is iteratively calculated from the user-item rating matrix, the fuzzy membership matrix of the cluster is actually obtained based on the user's subjective evaluation. To classify an item, the calculated results will inevitably be biased, if the user's subjective evaluation value is only considered, and the objective attributes of the item itself are ignored. In response to this, the HCFCM-SL algorithm mixes content-based and cluster-based recommendation algorithms, then uses content-based recommendation to assist cluster-based collaborative filtering recommendation algorithms.

The HCFCM-SL algorithm needs to obtain the content attribute information of all items. The content attribute information reflects the objective characteristic attributes of all items. Taking the ml-100k movie data set of MovieLens as an example, the content attribute information of each movie includes movie name, movie brief summary, director, actor, movie type, and so on. Then the membership degree vector of the item is extracted from the content attribute information. The processing flow of content attribute information is as follows:

(1) Obtaining: Obtaining the content attribute information of all items. If the content attribute information is not included in the public data set, it can be obtained through crawler technology;

(2) Word segmentation: For Chinese information, word segmentation is required to obtain the result of text segmentation, while English information does not require word segmentation;
(3) Vectorization: Using the term frequency-inverse document frequency (TF-IDF) method to vectorize the content attribute information of all items to obtain a high-dimensional space sparse vector representation;
(4) Dimensionality reduction: Using PCA dimensionality reduction technology, the high-dimensional sparse representation of the step (3) is reduced to a low-dimensional spatial dense vector representation, and feature representations that take into account both statistical and semantic information are selected;
(5) Clustering: Clustering the low-dimensional dense vector representation of the step by using K-means clustering method;
(6) Calculating the degree of membership: Calculating the low-dimensional vector of each item and the cosine value of each cluster center to obtain the item membership matrix.

Supposing that the user membership matrix obtained by the scoring matrix is $U_{m\times k1}$, the movie membership matrix is $V_{n\times k2}$, where the membership feature vector of each user is represented by u_i and the membership feature vector of each movie is represented by v_j, which is obtained from the content attribute information The movie membership matrix of is $W_{n\times k2}$, and the membership feature vector of each movie is w_j. In order to make the movie membership degree vector more fully express the subjective evaluation information and objective attribute information of the movie, the HCFCM-SL algorithm first normalizes the movie membership moment vector w_j, and then uses the normalized w_j and v_j Perform linear combination with a combination factor of c, as shown in the formula, to obtain a new movie membership vector.

$$v_j = (1-c)v_j + cw_j \tag{6}$$

The HCFCM-SL algorithm uses the user-item rating matrix to perform FCM clustering analysis from the user and item perspectives. Through FCM clustering, the HCFCM-SL algorithm can directly obtain the user and item membership matrix. At the same time, the HCFCM-SL algorithm mixes the content-based recommendation algorithm, by vectorizing the objective content attribute information, then using the clustering algorithm to calculate the membership feature vector of each item, and finally by linearly combining the subjective membership vector of the item and the objective membership vector, the prediction accuracy of the algorithm can be further improved. Finally, the algorithm connects the membership vectors of users and items, and inputs the feature space according to the existing score values in the score matrix. The hybrid supervised learning algorithm predicts the missing score values through the model to make recommendations. The detailed algorithm steps are shown in **Table 2**:

Table 2. Steps of HCFCM-SL algorithm.

Algorithm 2: HCFCM-SL algorithm
Input: user-item rating matrix $R_{m\times n}$, item content attribute data set. Output: the predicted value of the missing score value. BEGIN

(1) From the perspective of users, the FCM clustering algorithm is adopted to calculate the membership matrix $U_{m\times k1}$ of each user belonging to $k1$. clusters, and the user's membership feature vector $u_i \in U_{m\times k1}$;
(2) From the perspective of items, the FCM clustering algorithm is adopted to calculate the membership matrix $V_{n\times k2}$ of each item belonging to $k2$ cluster, and the item's membership feature vector $v_j \in V_{n\times k2}$;
(3) Basing on the content recommendation, we get the membership matrix $W_{n\times k2}$ in which each item belongs to $k2$ clusters, and also get the membership feature vector $w_j \in W_{n\times k2}$ of the item;
(4) Calculating the subjective membership degree vector and the objective membership degree vector of linear combination items：$v_j = (1-c)v_j + cw_j$;
(5) Concatenating the membership eigenvectors u_i and v_j of the users and items corresponding to the existing scoring items, creating the input sample space, and using the corresponding scoring value in the scoring matrix as the sample label;
(6) Dividing the data set in (5) into training set and test set, and training the random forest regression prediction model;
(7) Making predictions on the model trained in (6) on the test set, and evaluating the prediction accuracy of the algorithm;
(8) Using the training model in (6) to predict the sample set of missing score values, and recommending items that he may like to a user based on the predicted score values, or recommending an item to users who are likely to like it.
END

4. Experiments and Analysis

4.1 Data Set and Evaluation Metrics

In this paper, MovieLens's ml-100k and ml-1m datasets are used for the experiment. Among them, the ml-100k dataset contains 100000 ratings, 943 users, and 1,682 movies with each user ratings at least 20 movies, and each movie has been evaluated by at least one user. The ml-1m dataset contains 1,000,209 ratings, 6,040 users, and 3,952 movies, with each user ratings at least 20 movies, but 246 movies with no ratings. The sparsity of two datasets is 93.7% and 95.81% respectively. The dataset used by the content-based recommendation algorithm is the text summary data of the movie, which is composed of the movie. The ml-100k dataset contains summary data for 1,682 movies, and the ml-1m dataset contains summary data for 3,952 movies. We divide the original data set into 80% training set and 20% test set. The regression model is trained by the training set sample, and then the test set is used to predict the score, and finally the algorithm is evaluated.
MAE (Mean Absolute Error) and RMSE (Root-Mean-Square Error) [23] have been widely used in evaluating the accuracy of a recommender system. The method is to compare the numerical recommendation scores against the actual user ratings in the test data. The MAE is calculated by summing these absolute errors of the corresponding rating-prediction pairs and then computing the average.

$$\mathrm{MAE} = \frac{\sum u,i \in T \left| r_{ui} - \hat{r}_{ui} \right|}{\left| \mathrm{T} \right|} \tag{6}$$

$$\mathrm{RMSE} = \sqrt{\frac{\sum u,i \in T \left(r_{ui} - \hat{r}_{ui}\right)^2}{|\mathrm{T}|}} \tag{7}$$

Where r_{ui} represents the user u to the item i true score value, $\hat{r}_{ui}$ represents the predicted score value, T represents the test set, and |T| represents the sample size of the test set. The smaller MAE and RMSE values are, the higher the prediction accuracy of the algorithm is [20].

4.2 Evaluation Methods and Analysis

In this experiment, we compared the CF recommendation algorithm based on User-neighbor (UCF), the CF recommendation algorithm based on Item-neighbor (ICF), the recommendation algorithm based on user clustering (Clust), the collaborative filtering recommendation algorithm based on clustering and random forest (CRF), the supervised learning recommendation algorithm based on fuzzy C-means clustering (FCM-SL) and the hybrid supervised learning recommendation algorithm based on content and fuzzy C-means clustering (HCFCM-SL). Among them, compared with HCFCM-SL algorithm, FCM-SL algorithm in this paper lacks the part of objective membership degree feature vector in a linear combination.

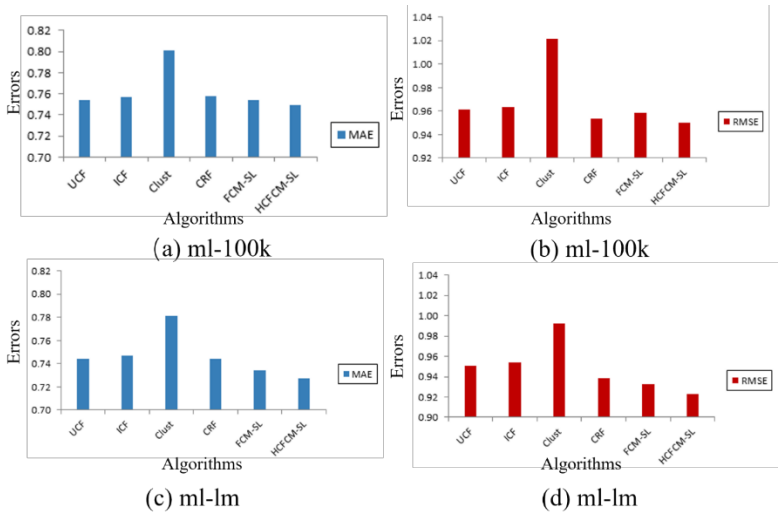

Fig. 1. Comparison of prediction accuracy of algorithms, (a) and (b) represents the results on the ml-100k dataset; (c) and (d) represents the results on the ml-1m dataset.

Fig. 1 shows the MAE and RMSE values of each algorithm on the ml-100k dataset and the ml-1m dataset. Lower values of the MAE and the RMSE indicate greater accuracy in predicatio. It can be found that the prediction accuracy of FCM-SL algorithm in this paper is not obviously improved, but by further mixing the content-based recommendation algorithm and using the content attribute information of items to supplement the membership degree matrix of original user-item scoring matrix, the prediction accuracy of HCFCM-SL algorithm is further improved. The results of each algorithm are compared in detail as shown in **Table 3** and **Table 4**. The positive and negative numbers in the table represent the changes of MAE and RMSE values of the proposed algorithm and the classical algorithms. On the ml-100k dataset, the MAE and RMSE values of HCFCM-SL are 0.7490 and 0.9504 respectively, which are 0.5% and 0.78% lower than those of FCM-SL and 0.92% and 0.3% lower than those of

CRF. On the ml-1m dataset, the MAE value of HCFCM-SL is 0.7270 and the RMSE value is 0.9226, which are respectively 0.71% and 1.02% lower than that of FCM-SL algorithm and 1.72% and 1.56% lower than that of CRF. In the ml-1m dataset, there are 246 movies have no score value, that is, there is the problem of cold start of articles. In this case, the prediction accuracy of the HCFCM-SL algorithm is improved even more obviously, proving that the algorithm has the ability to deal with the sparsity of data.

This paper also analyzes the influence of cluster center of FCM clustering algorithm on the HCFCM-SL algorithm, as shown in **Fig. 2(a)**. From **Fig. 2(a)**, we can see that the prediction accuracy of the algorithm will be gradually improved with the increase of the number of clustering centers on the ml-100k dataset. This is because as the number of clustering centers increases, the dimension of potential feature vectors of users and items will increase, sample features used in model training will increase, and the prediction accuracy of the algorithm will be improved accordingly. But the **Fig. 2(b)** found that the running time of the algorithm will increase greatly with the increase of the number of clustering

Table 3. Comparison of error values of algorithms on the ml-100k dataset.

Algorithms	MAE	RMSE
UCF	0.7539(-0.49%)	0.9600(-1.07%)
ICF	0.7571(-0.81%)	0.9637(-1.33%)
Clust	0.8081(-5.2%)	1.0215(-7.11%)
CRF	0.7582(-0.92%)	0.9534(-0.3%)
FCM-SL	0.7540(-0.5%)	0.9582(-0.78%)
HCFCM-SL	0.7490(0%)	0.9504(0%)

Table 4. Comparison of error values of algorithms on the ml-lm dataset.

Algorithms	MAE	RMSE
UCF	0.7439(-1.69%)	0.9511(-1.07%)
ICF	0.7471(-2.01%)	0.9537(-3.11%)
Clust	0.8081(-5.4%)	0.9925(-6.99%)
CRF	0.7442(-1.72%)	0.9382(-1.56%)
FCM-SL	0.7341(-0.7%)	0.9328(-1.02%)
HCFCM-SL	0.7270(0%)	0.9226(0%)

centers, when the number of clustering centers of users and items are 21 and 19 dimensions respectively, the performance of the algorithm is the best, namely the abscissa is 40 points of algorithm performance is best.

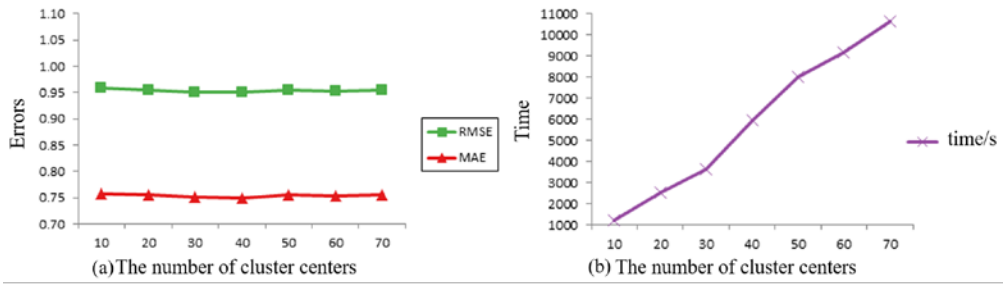

Fig. 2. Influence of the number of cluster centers on the HCFCM-SL algorithm.(a) Change of algorithm prediction accuracy with the number of cluster centers; (b) Change of algorithm running time with the number of cluster centers.

This experiment also analyzed the effect of the combination factor on the experimental results when the item membership degree feature vector is linearly combined. **Fig. 3** shows the influence of the combination factor c on the MAE and RMSE in the ml-100k data set and ml-1m data set. **Fig. 3 (a)** and **Fig. 3 (b)** show the results of the ml-100k data set while **Fig. 3 (c)** and **Fig. 3 (d)** show the results of the ml-1m data set. From **Fig. 3 (a)** and **Fig. 3 (b)**, we can see that when the combination factor is 0, which means without hybrid content-based recommendation algorithms, the MAE and the RMSE is 0.7540 and 0.9582 respectively. When the combination factor reaches 0.1, the MAE is 0.7490, RMSE is 0.9504, respectively reaching the minimum value. At this time, the prediction accuracy of the algorithm is the highest. When the combination factor is 0.5, the MAE and the RMSE is 0.7518 and 0.9521 respectively. When the combination factor is in the range of 0.1 to 0.5, the MAE and RMSE of the algorithm fluctuates which means the prediction accuracy of the algorithm does not change significantly. However, when the combination factor exceeds 0.5 which means that the weight of Content-based Recommendation increases. Both MAE and RMSE show an upward trend while the prediction accuracy of the algorithm is decreasing. Finally, the experiment found that when the combination factor is 0.1, the algorithm performs best; when the combination factor is in the range from 0.1 to 0.5, the algorithm performs relatively well.

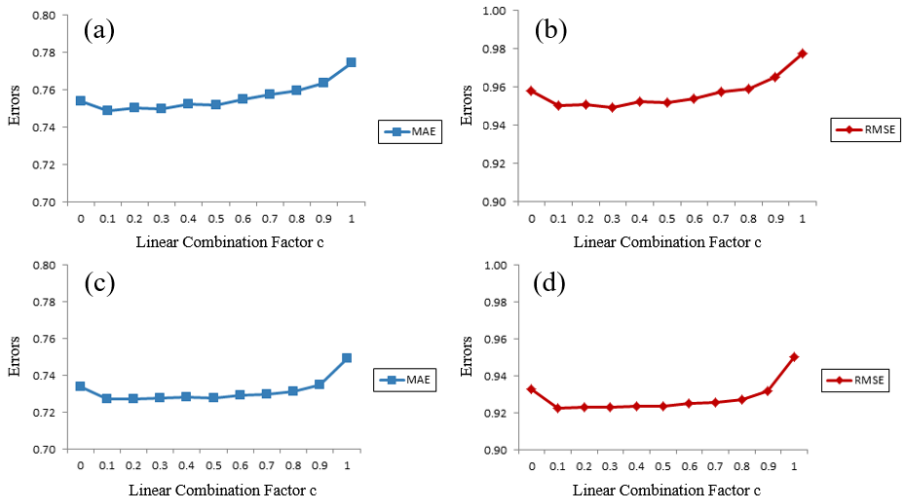

Fig. 3. Influence of linear combination factor c on the HCFCM-SL's prediction accuracy: (a) and (b) represents the results on the ml-100k dataset; (c) and (d) represents the results on the ml-1m dataset.

Experimental results prove that hybrid membership vector obtained based on FCM clustering recommendation and the membership vector generated based on content recommendation can improve the prediction accuracy of the recommendation algorithm. When the weight of the membership vector generated based on content recommendation is relatively small, the algorithm performance better.

Fig. 3 (c) and **Fig. 3 (d)** continue to analyze the effect of the combination factor c of the ml-1m data set on the value of MAE and RMSE, and the experimental conclusion are consistent with that of the ml-100k data set. By experimenting on data sets with different sparseness, results show that the HCFCM-SL algorithm proposed in this paper performs better than the single FCM-SL algorithm. This indicates that hybrid content-based recommendation can further improve the prediction accuracy of the algorithm. In addition, the experiment pointed

out that between the above two recommendation algorithms, the weight of the membership vector obtained based on content recommendation should not exceed the weight of the membership measure based on FCM clustering recommendation and meanwhile, the algorithm effect is relatively good.

5. Conclusion and Future Work

As for the cold start problem of Traditional Hybrid Clustering and distance clustering to Supervised Learning, this paper proposes Hybrid Content and Fuzzy C-Means Clustering to Supervised Learning. The algorithm uses fuzzy C-means clustering to solve the user and item membership degree matrix, constructs the user and item membership feature vector, and solves the problem that the traditional distance-based membership degree is not available due to the data sparseness. In addition, the algorithm improves the accuracy of the algorithm's scoring prediction by mixing the content-based recommendation algorithm and the subjective and objective membership degree feature vector of linearly combination. The HCFCM-SL optimization algorithm proposed in this paper uses the random forest regression algorithm as a supervised learning algorithm. In the future, more predictive models can be explored for collaborative filtering and supervised learning recommendation algorithms.

Acknowledgement

This research was supported by the National Natural Science Foundation of China under Grant No.61902021, Beijing Natural Science Foundation under Grant No.4212008, the Basic Scientific Research Project of Beijing Jiaotong University (No. 2019RC050).

References

[1] Isinkaye F O, Folajimi Y O, Ojokoh B A, “Recommendation systems: Principles, methods and evaluation,” *Egyptian Informatics Journal*, 16(3), 261-273, 2015. Article (CrossRef Link)
[2] Lops P, Jannach D, Musto C, et al., “Trends in content-based recommendation,” *User Modeling and User-Adapted Interaction*, 29(2), 239-249, 2019. Article (CrossRef Link).
[3] Pereira A L V, Hruschka E R, “Simultaneous co-clustering and learning to address the cold start problem in recommender systems,” *Knowledge-Based Systems*, 82, 11-19, 2015. Article (CrossRef Link).
[4] Su X, Khoshgoftaar T M, “A survey of collaborative filtering techniques,” *Advances in artificial intelligence*, 2009. Article (CrossRef Link)
[5] Stephen S C, Xie H, Rai S, “Measures of similarity in memory-based collaborative filtering recommender system: A comparison,” in *Proc. of the 4th Multidisciplinary International Social Networks Conference*, 1-8, 2017. Article (CrossRef Link)
[6] Zhang L, Liu X, Zhou X, “A Novel Recommendation Algorithm Based on Clustering Dissimilarity Measures,” *Journal of Engineering Science & Technology Review*, 13(3), 2020.
[7] Kumar B, Sharma N, “Approaches, issues and challenges in recommender systems: a systematic review,” *Ind J Sci Technol*, 9(47), pp. 1-12, 2016. Article (CrossRef Link)
[8] Kumar N, Ozakin A, Gray A, et al., “Supervised Learning Based Recommendation System,” U.S. Patent Application 15/249, 386, 2017-3-2. https://patents.google.com/patent/US20170061286A1/en
[9] Kumar P, Thakur R S, “Recommendation system techniques and related issues: a survey,” *International Journal of Information Technology*, 10(4), 495-501, 2018. Article (CrossRef Link).
[10] Çano E, Morisio M., “Hybrid recommender systems: A systematic literature review,” *Intelligent Data Analysis*, 21(6), 1487-1524, 2017. Article (CrossRef Link)

[11] Walek B, Fojtik V., "A hybrid recommender system for recommending relevant movies using an expert system," *Expert Systems with Applications*, 158, 113452, 2020. Article (CrossRef Link)
[12] Dooms S, De Pessemier T, Martens L., "Online optimization for user-specific hybrid recommender systems," *Multimedia Tools and Applications*, 74(24), 11297-11329, 2015. Article (CrossRef Link)
[13] Ferdaous H, Bouchra F, Brahim O, et al., "Recommendation using a clustering algorithm based on a hybrid features selection method," *Journal of Intelligent Information Systems*, 51, 183-205, 2018. Article (CrossRef Link)
[14] Chen Z, Li Z, "A collaborative recommendation algorithm based on user cluster classification," in *Proc. of 2016 4th International Conference on Cloud Computing and Intelligence Systems (CCIS)*, IEEE, 260-263, 2016. Article (CrossRef Link)
[15] Braida F, Mello C E, Pasinato M B, et al., "Transforming collaborative filtering into supervised learning," *Expert Systems with Applications*, 42(10), 4733-4742, 2015. Article (CrossRef Link)
[16] Barbieri J, Alvim L G M, Braida F, et al., "Autoencoders and recommender systems: COFILS approach," *Expert Systems with Applications*, 89, 81-90, 2017. Article (CrossRef Link)
[17] Koohi H, Kiani K, "User Based Collaborative Filtering using Fuzzy C-Means," *Measurement*, 91, 134-139, 2016. Article (CrossRef Link)
[18] X. Wang, M.-C. Chang, L. Wang, S. Lyu, "Efficient algorithms for graph regularized plsa for probabilistic topic modeling," *Pattern Recognition*, 86, 236–247, 2019. Article (CrossRef Link)
[19] Unger M, Tuzhilin A, Livne A., "Context-Aware Recommendations Based on Deep Learning Frameworks," *ACM Transactions on Management Information Systems (TMIS)*, 11(2), 1-15, 2020. Article (CrossRef Link)
[20] H. Liu, Y. Wang, Q. Peng, F. Wu, L. Gan, L. Pan, P. Jiao, "Hybrid neural recommendation with joint deep representation learning of ratings and reviews," *Neurocomputing*, 374, 77–85, 2020. Article (CrossRef Link)
[21] Khan Z Y, Niu Z, Yousif A, "Joint deep recommendation model exploiting reviews and metadata information," *Neurocomputing*, 402, 256-265, 2020. Article (CrossRef Link)
[22] F. Mutinda, A. Nakashima, K. Takeuchi, Y. Sasaki, M. Onizuka, "Time series link prediction using nmf," *Journal of Information Processing*, 27, 752–761, 2019. Article (CrossRef Link)
[23] Brassington G, "Mean absolute error and root mean square error: which is the better metric for assessing model performance?," in *Proc. of EGU General Assembly Conference Abstracts*, p.3574, 2017. Article (CrossRef Link)

LI DUAN received her Ph.D degree in Computer Science and Technology from Beijing University of Posts and Telecommunications, Beijing, China, in 2016. She is currently an Assistant Professor with the School of Computer and Information Technology, Beijing Jiaotong University. She was a research fellow of Nanyang Technological University and University of Science and Technology Beijing, her research interests are services computing and internet of thing, Data security and privacy protection, and Blockchain security and applications. She received the National Natural Science Foundation of China, the Postdoctoral Fund, and the Basic Scientific Research Project.

WEIPING WANG received the Ph.D. degree in telecommunications physics electronics from the Beijing University of Posts and Telecommunications, Beijing, China, in 2015. She is currently an Associate Professor with the Department of Computer and Communication Engineering, University of Science and Technology Beijing. Her current research interests include auto-driving vehicle formation control, brain-like computing, memrisitive neural networks, associative memory awareness simulation, complex networks, network security, and image encryption. She received the National Key Research and Development Program of China, the State Scholarship Fund of China Scholarship Council, the National Natural Science Foundation of China, the Postdoctoral Fund, and the Basic Scientific Research Project.

BAIJING HAN received the B.Sc. degree from Southwest University, in 2018. She is currently pursuing the master's degree with the University of Science and Technology Beijing. Her current research interests include auto-driving vehicle for mation control, brain-like computing, and intelligent control.