Abstract: Albeit there exists many collaborative recommendation models, it is still a challenge to increase the performance of these models. In this manuscript, a peculiar approach for collaborative filtering based on fuzzy-c means clustering algorithm is proposed. We use cuckoo search optimization algorithm along with fuzzy-c means to overcome the local optima. Thenceforth Pearson correlation similarity is adopted to compute the similarity between users with in the cluster. Finally, we produce recommendations for the corresponding users. Precise analysis on a standard movie lens dataset reveals that our approach outer performs many other recommendation models and also solves the cold start problem in recommendation systems.

Key Words: Fuzzy-c means, cuckoo search, pearson correlation similarity.

1. Introduction

In the Internet generation, recommendation system has become an essential component. An enormous amount of websites exists today, but it becomes more difficult and sluggish for the users to find the information for what they are looking. In order to help users to find the information for what they are looking and filter out the unwanted information recommendation systems has been used. With the advent of this different types of recommendation techniques has been developed, comprising Collaborative Filtering(CF), Content based Filtering (CBF), Demographic based Filtering(DF), Community based, Knowledge based Filtering (KF) and Hybrid approach.

Collaborative Filtering is independent on the content generated from metadata like music, movies and products etc. It builds a utility matrix (user-item) which consists of preferences given by users for items and then calculates the similarities between the users and forms a neighborhood group.

Recommendations are generated to the user to those items that it was not rated but the users in the neighborhood were already rated positively to the items. Collaborative Filtering are broadly divided into Memory based and Model based techniques. In Memory based technique The past rating history of the users to the items shows a significant role in finding the neighbors that has similar taste preferences. Once a user neighbors are identified, various algorithms are applied to merge the neighbor preferences and generate recommendations to the user. Owing to the efficacy of these methods, they obtain an outspread success in real time applications. Memory based collaborative filtering can be attained through user based and item based techniques. In user based collaborative filtering similarity is calculated between users by considering the ratings on the same item, and then the predicted rating for an item is computed by considering the weighted average ratings of an item. In Item based collaborative filtering similarity is calculated between items and then the predicted rating for an item is computed by considering the weighted average ratings of items. In Model based approach, various datamining, machine learning models are applied to predict rating of unrated items. Singular Value Decomposition, Latent Semantic methods, Matrix completion Technique, Regression and clustering etc. comes under model based techniques. It analyzes the data and identifies the relations between data; from these relations they produce the top-N recommendations list.

In Content based approach items are recommended based on the past preferences of the users. In Demographic Filtering items are recommended based on the user profile. Community based approach recommends the items based on the preference of the users in the community (user friends). In Knowledge based approach items are recommended based on the inferences drawn from the user needs and tastes. A Hybrid approach combines any two recommendation techniques and provides better results than any individual model.

In this paper, a novel approach for user based collaborative filtering is proposed to solve the cold start problem and improves the model performance. Initially fuzzy-c means clustering is applied on the user’s data (age, gender and occupation). The reason to choose fuzzy-c means is that it can able to solve multiclass, ambiguous problems. However, fuzzy-c means algorithm highly depends on initial states which leads to the local optimum. Therefore, to overcome this problems we use cuckoo search optimization algorithm.
along with fuzzy-c means. Then we use Pearson correlation similarity to estimate the similarity between users. As a consequence, an utility matrix is generated which consists of ratings from a scale of 1-5. Finally, recommendations are generated to the users.

The main contribution of this paper is outlined as follows: we propose a novel approach in collaborative filtering to make good recommendations for the users. We import fuzzy-c means clustering into it in order to enhance the model performance. The remaining part of the paper is organized as follows. In section 2 a review of literature work is given. In section 3 the proposed work is illustrated. Result analysis and conclusion part is explained in the section 4.

2. Literature work

In this section, we will describe research work done on recommendation systems with relevant conceptions.

Chunhui et al [12] proposed a novel approach in collaborative filtering by adopting k-means and Artificial bee colony algorithm and the experimental results shows that the method has given best performance. Saveski et al [13] addressed the cold start problem in recommendation systems based on local collective embeddings. The experimental results reveal that this approach is performing well. Lu et al [17] surveyed different types of recommendation systems and observe the model performance in applications like e-government, e-business, e-commerce, e-library and e-tourism. Xu et al [18] proposed a new collaborative model with user generated content. This model learns the user’s preferences form the reviews; social tags and makes recommendations to the users. The results shows that the method is performing well compared to other methods. Lu et al [19] proposed a security model in collaborative filtering which concerns about the user’s security and the model has shown best performance.

Yang et al [20] proposed a fused method for news recommendation system it is a hybrid approach and it avoids the cold start problem. The experimental results proven that this method is performing well. Panigarhi et al [25] implemented a hybrid collaborative recommendation system. This approach is very useful in making recommendations on large amount of data and avoids the cold start problem.

Yuan et al [26] developed a new model in collaborative filtering by using deep learning approach. This model has efficiently overcome the item cold start problem. Mishra et al [28] has given an overview on types of recommendation systems and different type so evaluation metrics which are used to measure the performance of recommendation systems. Katarya et al [32] developed a new collaborative recommendation system by combining k-means clustering and cuckoo search optimization algorithm. The results proven that this method is performing well. Haruna et al [33] proposed a new approach of collaborative filtering in research paper recommendations. The uniqueness of this approach is it provides personalized recommendations irrespective of research field.

3. Methodology of Collaborative Filtering Using Fuzzy-cmeans and Cuckoo Search

Collaborative filtering is affected by cold start problem i.e. when a new user enters into the system, it is difficult to generate recommendations because his past ratings are not recorded. To overcome that limitation and also to improve model performance among the existed systems, we propose a hybrid approach of collaborative filtering using fuzzy-c means clustering and cuckoo search optimization algorithm. Our intention is to design a consolidated model that uses users data (users profile and users rating) to make recommendations and also able to withstand with cold start problem. In our approach fuzzy-c means clustering is applied on user’s data (age, occupation and gender) by consider the analogy “Users who are under same age, who are in same occupation and who belongs to same gender” will have similar preferences.

In fuzzy-c means clustering, initial centroids are selected at random; then the users are observed in a sequence by estimating the degree of membership between the users and initial centroids, and the users are allocated to clusters whose membership degree is high. Then the new centroid points are generated from the clusters by measuring the cluster means. This repetitive relocation process would continue until there is no change in centroid points. The steps of fuzzy-c means algorithm is explained here:

1. Choose „C” points among the users that will be clustered. These initial „C” points can be treated as centroids.

2. Calculate the Euclidian distance between the users to centroids, $D_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$ where $(x_i,y_i)$ represents users data and $(x_j,y_j)$ represents centroids.

3. Measure the membership function for user to

$$\mu_q = \left[\frac{1}{\sum_{j=1}^{n-1} \left|\frac{1}{d_{ij}}\right|}\right]$$

those centroids. Where „$i$” represents number of users and „$j$” represents cluster number.

4. Allocate each user to the cluster that has degree of membership.

5. Once all the users are assigned, recomputed the centroids of each cluster.
6. Rerun the steps two to five till centroid points doesn’t change.

The problem with fuzzy-c means algorithm is selection of initial centroids, it leads to local optima problem. To overcome that, an optimization algorithm should be used. There exists many optimization algorithms such as genetic algorithm; particle swam optimization, Bat and cuckoo search algorithm etc. Among these optimization algorithms we have chosen cuckoo search algorithm, because it outperforms many other optimization algorithms. Cuckoo search algorithm was influenced by obligate brood parasitism of a few cuckoo species. The following rules can be used to depict the cuckoo search algorithm.

1. Every single cuckoo lays one egg at a time and places its egg in an arbitrary nest.

2. The finest nest with immense quality of eggs will move to later generations.

3. The number of available host nests is fixed, and the egg laid by cuckoo is identified by the host bird with a probability Pa € [0,1].

The procedure of cuckoo search algorithm commences with the initialization of arbitrary n host nests. Then levy flight equation is used to obtain a cuckoo, here levy flight is based on random walk in which step lengths follows a levy distribution. Once cuckoo egg is found, then an arbitrary nest let us assume “j” is selected and compares the fitness quality of cuckoo egg in the host nest with the random nest (j) using objective function. If the fitness quality of cuckoo egg in the random nest (j) is greater than the fitness quality of cuckoo egg in the host nest „then the cuckoo egg is replaced in the random nest otherwise it won’t”, and this process will repeat till it meet the stop criterion or maximum number of iterations. In this framework, we considered the user as an egg and nest as cluster.

The steps of cuckoo search algorithm are illustrated here.

**Objective function** \( f(y_i,y_j,y_k,-----y_d) \)

Generate initial population of n host nests \( y_i \) (i=1,2,---n)

While (k<Max Iteration) or (stop criterion)

Acquire a cuckoo randomly by levy flights

Measure its quality fitness Fi

Select a nest among n, consider (j) arbitrarily

If (Fi>Fj),

Replace j by the new solution; End.

A fraction (Pa) of the bad nests are abandoned and built the new ones. Keep the supremacy solutions.

Assign rank to the solutions and select the best. End while

Process the results

End.

### 3.1. Methodology

Overall, A novel framework for user based collaborative filtering is proposed using fuzzy-c means with cuckoo search optimization algorithm. The pseudo code of this novel framework is shown here.

**Pseudo code**

1. Implement fuzzy-c means clustering on the user’s profile.

2. To get the optimized clusters, cuckoo search algorithm is executed on the outcomes generated by fuzzy-c means.

3. For each cluster j

   For each user i in cluster j

   Substitute the ratings data by performing join operation on user profile data using user id.

4. For each cluster j

   For each user (u) in cluster j

   Identify the similar users (su) who has rated items in common with user (u). Measure the similarity between user (u) and similar users (su)

   For i in items rated by su and i not in items rated by u

   Estimate the predicted rating (\( r^u_{i} \)) of an item (i) by the user (u).

5. Estimate the performance of model by comparing the predicted ratings with the actual ratings.

The proposed approach is applied on the Movie lens data set, it constitutes Users data, Movies data and ratings data. User’s data consists of users information such as user id, age, occupation and country. Movies
data consists of movies information such as Movie id, title and genres. Ratings data dwell up with rating information in which users has given rating to movies and the fields in the rating data comprises of user id, movie id and rating. The architecture of the proposed methodology is shown in the Figure 1.

Clustering algorithm fuzzy-c means is applied on the user’s data to categorize the user’s into different clusters based on the context of user’s who are in same context, same profession and belongs to same nation will have analogous preferences. Then cuckoo search algorithm is applied on the clusters, in cuckoo search clusters are treated as nests and each user in the cluster is considered as egg. As a consequence optimized clusters are obtained.

To perform user based collaborative filtering, rating information of users given to movies must be needed. To get the ratings data from the optimized clusters we performed join operation on the users data and rating data. As a result, rating data is retrieved for each user in the cluster. Now users are grouped into different clusters based on the user preferences, then user based collaborative filtering technique is applied on each cluster. To compute the similarity between users in the cluster we use Pearson correlation coefficient, let us assume there are two persons x, y to compute the similarity between them the Pearson correlation coefficient can be defined as follows:

\[ \text{sim}(x, y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - x)^2} \sqrt{\sum (y_i - y)^2}} \]

Here ,,i” is the items rated by both users x and y, x and y are the mean value of ratings. Once the similarity between users in the cluster is computed, then the estimated rating for the unrated item by the user will be predicted.

3.2. Approach for handling the cold start problem

When a new user enters into the system, user will provide information of demographic details like age, gender and occupation. Based on his demographic details, fuzzy-cmeans cuckoo search algorithm makes the user fall into the cluster of similar users group. Then the movies which are frequently seen and highly rated by the similar users in the clusters will be recommended.

In addition to this problem, the proposed approach can also able to reduce scalability and sparsity problems. Clustering is performed on user’s demographic information, based on users demographic details user fall into which cluster can be identified. Then collaborative filtering is applied to that cluster only instead of applying on whole data.

4. Experiment Results and Analysis

We use movie lens m1-lm data set to perform experiments, it can be easily found in online, constitutes 1,000,209 ratings for 3,900 movies made by 6,040 users. The dataset consists of files like user’s data, ratings data and movies data. User’s data constitute of user information like user id, age, gender, occupation and zip code. Ratings data constitute of fields like user id, movie id, rating and timestamp. Movies data consists of information like Movie id, title and genres. In movie lens data set, every user has rated at least 20 movies. The data set is partitioned into train set and test set in such a way that the test set holds 10 ratings given by each user. To observe the performance of the proposed framework we have used metrics like MAE, RMSE.

MAE: It can be defined as Mean Absolute Error; it helps to calculate the absolute difference between predicted ratings and actual ratings which can be depicted form the equation.

\[ \text{MAE} = \frac{\sum |p_{i,j} - r_{i,j}|}{n} \]

Where pi,j is the predicted rating of user i for movie j,
is the actual rating of user i given to movie j and “n” is the number of movies.

**RMSE:** It can be defined as Root Mean Square Error. It calculates root mean square difference between the ratings predicted by the model and actual ratings, which can be viewed from the below equation.

In the proposed framework, fuzzy-cmeans cuckoo search algorithm is applied on the user’s demographic data and the data is partitioned into different clusters. Then for each cluster, rating data is generated and the collaborative filtering is applied on rating data of each cluster to generate recommendations. The performance of the model has estimated by varying different values of k (number of clusters). And it can be viewed from the Table 1.

Table 1: Comparison of MAE and RMSE Values with Varying Clusters

<table>
<thead>
<tr>
<th>Number of clusters (k)</th>
<th>MAE</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>0.816630</td>
<td>1.290652</td>
</tr>
<tr>
<td>10</td>
<td>0.795269</td>
<td>1.28070</td>
</tr>
<tr>
<td>15</td>
<td>0.781423</td>
<td>1.26101</td>
</tr>
<tr>
<td>20</td>
<td>0.744569</td>
<td>1.25802</td>
</tr>
<tr>
<td>25</td>
<td>0.7124378</td>
<td>1.25079</td>
</tr>
<tr>
<td>30</td>
<td>0.684214</td>
<td>1.24342</td>
</tr>
<tr>
<td>35</td>
<td>0.671330</td>
<td>1.23921</td>
</tr>
<tr>
<td>40</td>
<td>0.657293</td>
<td>1.21962</td>
</tr>
<tr>
<td>45</td>
<td>0.662341</td>
<td>1.21085</td>
</tr>
</tbody>
</table>

Performance of the proposed approach is measured on movie lens data sets using metrics like MAE and RMSE with varied number of clusters. From the Table-1 it is observed that MAE and RMSE values have been reduced as the cluster number k-value increases. The reason for such attitude is, as the number of clusters increases and there will be more possibility for user to allocate to the respective cluster in which similar users existed. As a result the difference between predicted rating and the actual rating will be decreases. The proposed model has given least MAE value as 0.657293 at k=40 when compared to different k-values and RMSE value is 1.21162.

Table 2: Comparison of the Proposed Model with Different Approaches

<table>
<thead>
<tr>
<th>Approaches on collaborative filtering</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>SOM cluster</td>
<td>0.75</td>
</tr>
<tr>
<td>k-means</td>
<td>0.69</td>
</tr>
<tr>
<td>k-means cuckoo</td>
<td>0.68</td>
</tr>
<tr>
<td>Fuzzy-cmeans cuckoo</td>
<td>0.65</td>
</tr>
</tbody>
</table>

In [33] the authors performed experiments on movie lens dataset using SOM, k-means and k-means cuckoo. From these results the performance of the proposed model (fuzzy-c means cuckoo search) is compared by using mean absolute error and it is shown in the Table 2. From the Table 2 it is identified that the fuzzy-cmeans cuckoo outer performs many other approaches.

5. Conclusion

In this paper, we propose fuzzy-cmeans cuckoo search collaborative filtering, a novel approach. It is used for generating movie recommendations to the users and also able to handle the cold start problem. Experimental analysis has done on movie lens dataset with the proposed approach and it has given 0.65 MAE. It shows that, fuzzy-cmeans cuckoo collaborative filtering approach outer performs many other approaches. In future, we would to improve our approach to scale large amount of data by using Hadoop technology.

References


[9] Yang X.S., Deb S., Cuckoo search via Lévy


[27] Shou Qiang L., Ming Q., Qing-Zhen X., Research and Design of Hybrid Collaborative Filtering Algorithm Scalability Reform Based on Genetic Algorithm Optimization, 6th International Conference on Digital Home (2016), 175-179.


◆◆◆