A Bicriteria Clustering Approach for Collaborative Filtering

Emin T. Demirkiran¹ and Ahmet M. Turk, and Cihan Kaleli

Computer Engineering Department, Anadolu University, Eskisehir 26470, Turkey

Email: {ckaleli, etd, ahmetmuratturk}@anadolu.edu.tr

ABSTRACT

Clustering is one of the essential methods of data reduction. It is possible to find homogenous sub-sets of huge amount of data by employing clustering. In collaborative filtering schemes, clustering is used to form off-line user or item neighborhoods in order to enhance online performance. Classical clustering methods for collaborative filtering are only based on distances or correlations among entities. Thus, it is hard to form neighborhoods without sacrificing any useful entity by clustering. In this paper, we introduce a new bicriteria k-means clustering approach for collaborative filtering. We employ a degree of uncertainty of users along with similarities in order to obtain a single clustering criterion. We perform experiments on two benchmark data sets in order to measure the proposed approach’s accuracy. Experimental outcomes indicate that, it is possible to improve accuracy of a recommender system using bicriteria-based k-means clustering.

KEYWORDS

Collaborative filtering, nearest neighbors, bicriteria, k-means clustering, entropy.

1 INTRODUCTION

Collaborative filtering (CF) is one of the most popular recommendation techniques employed in various online applications, especially in e-commerce [1]. CF utilizes daily life notion of "word of mouth" [2] and it has a basic assumption that people who agreed in the past, tend to agree in the future. In a typical CF process, there is a user-item matrix which holds product preferences of customers and such matrix is employed in order to provide personalized recommendations for users. The preferences can be collected either as explicit indications given by users, or implicit clues extracted from log-based data. By employing CF services, users might request a recommendation about a book, movie, CD, hotel, or research paper.

The very first introduced CF algorithms depend on the correlations between users, which are users and items, and they are called memory-based CF algorithms [3]. These type of CF methods utilize entire user-item matrix during recommendation process in order to form
nearest neighbors (NN) of entities, and according to algorithm type, i.e., user- or item-based, they operate on users' or items' vectors in the database. In literature, the best accuracy level for recommendations are achieved by memory-based algorithms.

Although memory-based CF systems have the highest accuracy performance, they are more likely to encounter certain shortcomings such as scalability and data sparsity due to their way of data utilization. Scalability of a recommender system reduces in line with the number of users and ratable products it has, and typically, such numbers are huge. Therefore, forming NN of a user might require unreasonably high amount of time. Scalability challenge actually can be considered for only user-based CF algorithms. In a typical CF system, variety of items in a preference data does not change frequently, thus, it is possible to determine NN of items by an offline preprocess. In traditional CF systems, estimating users' correlations step is crucial in order to produce accurate predictions. However, since most of the users in a user-item matrix have preferences for a very small fraction of entire products, data sparsity problem arises and due to this problem, there might not have enough information about an entity in order to form its neighborhood.

During evolution of CF-based recommender systems, researchers lean in to introduce new methods in order to alleviate memory-based CF algorithms' challenges. To enhance scalability of recommendation process, model-based CF approaches are proposed. Such algorithms try to extract a useful prototype from a user-item matrix by employing different data mining methods, e.g., dimension reduction [4], matrix factorization [5], clustering [6], and decision trees [7]. Although model-based CF algorithms are useful to overcome scalability challenges, since tuning of their parameters is relatively hard, such methods might have trouble to provide desired accuracy level [6].

In order to overcome challenges caused by sparse data, recommender systems' capability at extracting useful knowledge from user-item matrix is needed to be improved. Otherwise, producing accurate and reliable predictions becomes very difficult. Therefore, CF methods try to consider all gatherable information from users' rating vectors. However, traditional memory-based CF algorithms consider only correlations among entities in order to form neighborhoods. Hence, researchers propose new methods that improve neighbor selection capability of memory-based systems by either proposing new similarity measures [8, 9, 10] or introducing new aspects for NN formation methodology [11, 12, 13].

In this paper, we focus on clustering-based data reduction in user-based CF systems. The
proposed clustering methodologies for recommender systems depend on a single criterion, e.g., correlation and distance, in order to determine an entity's cluster membership. However, as shown in a previous study [13], it is possible to improve accuracy of predictions by considering users' interpretations about a rating domain along with correlations in order to form their NN. In [13], a new uncertainty-based measure, degree of uncertainty (DU), which is estimated by calculating entropy of their rating vectors, is introduced and it is also noted that two users having maximum similarity and minimum DU difference are the best correlated users. Therefore, we claim that it is possible to improve clustering performance by employing two criteria instead of using a single one in order to group users. In our new clustering methodology, we try to place users having the highest similarity and minimum DU differences into the same clusters. However, our new approach introduces the optimization problem of clustering satisfying all criteria considered. Such type of clustering is defined as Pareto efficient clustering [14]. According to authors [14], bicriteria clustering can be considered as a classical clustering by introducing a single criterion obtained as a combination of given criteria. In this paper, we employ these bicriteria clustering approach on k-means clustering algorithm. We evaluate success of our new clustering method on traditional user-based CF algorithm using two well-known data sets. Experimental outcomes indicate that it is possible to improve clustering performance for user-based CF algorithms by employing two criteria instead of a single one.

Major contributions of our study are listed below:

- A bicriteria clustering approach for CF is introduced.
- The way of combining two different criteria as a single one is proposed.
- Accuracy of traditional k-means clustering-based CF is improved.

The rest of the paper is organized, as follows. In the following section, we give brief information about related studies. In Section 3, we explain traditional user-based CF algorithm. After, describing our bicriteria clustering approach in Section 4, we display experiment results in Section 5. Finally, we conclude our paper and give new future directions in Section 6.

2 RELATED WORK

Clustering methods divide members of a given data set into homogeneous sub-groups in which dissimilarities among entities are minimized. In a similar manner, memory-based CF algorithms try to minimize dissimilarities between a user
and her NN in a recommendation process. Therefore, during CF algorithms evolution, various clustering techniques are utilized in order to enhance online performance of recommender systems. The first clustering method for CF is introduced by Ungar and Foster [15]. The authors propose a formal statistical model for and parameters in the model can be estimated by using \( k \)-means clustering. Besides, employing \( k \)-means clustering, Roh et al. [16] show that it is possible to change clustering approach for CF into a supervised learning problem by using machine learning methods, i.e., Self-Organizing Map (SOM) and Case Based Reasoning. Their method is successful at improving both accuracy and efficiency of a recommender system. Researchers also aim to utilize both memory- and model-based CF algorithms' advantages. Thus, Xue et al. [17] present a clustering-based smoothing method which outperforms other state-of-the-art CF algorithms. In another recent study, Shinde and Kulkami [18] offer a novel centering-bunching based clustering algorithm which is used for hybrid recommender systems. George and Merugu [19] introduce a co-clustering method which clusters users and items simultaneously. They combine their method with other dimension reduction approach, i.e., SVD. Experimental results indicate that the proposed method has a much lower computational cost than correlation and matrix factorization-based algorithms. Also, Khoshneshin and Street [20] offer another method which improve accuracy of co-clustering-based CF methods. In order to improve Fuzzy C-means (FCM) clustering-based collaborative filtering methods' accuracy, Wu and Li [21] propose to combine matrix factorization with FCM. Closely similar to our work, Kim et al. [22] aim to improve neighbor selection capability of \( k \)-means clustering for CF by introducing two new approaches. In their first method, the authors try to solve a classical clustering problem in which two similar users might be in different groups. Their second method performs clustering on numerical ratings along with categorical data. Besides, traditional rating-based clustering methods for CF, Birtoli and Ronca [23] take into account trust between users along with similarities and employ FCM in order to identify interesting products for customers. In addition, researchers introduce solutions in order to improve scalability of privacy-preserving CF algorithms [24, 25].

Besides introducing clustering-based solutions in order to overcome online performance issues, researchers also focus on alleviating challenges caused by data sparsity. These type of works can be divide into two partitions. The first group includes the studies introducing new similarity measures and second one comprises the methods aiming to enhance
neighbor selection capability of CF algorithms. Ahn [8] offers a new similarity measure based on domain specific interpretations of user preferences to handle with cold start problem. In another work, Bobadilla et al. [26] propose a similarity measure utilizing non-numerical information of data. Choi and Suh [10] introduce a new perspective which employ target item similarities in classical PCC-based estimating of correlations. Herlocker et al. [3] state that neighbor selection part of a CF algorithm is crucial for success of the algorithm. In their paper, authors perform a comprehensive work which has been a milestone reference for researchers. In another work, Kim and Yang [27] propose an effective threshold-based neighborhood forming approach instead of determining k-NN of users. Koren [11] interprets selection of appropriate neighbors as a global cost function and his method introduce an improvement in accuracy. Anand and Bharadwaj [28] offer to utilize two different similarities, i.e, local and global, which are combined during a CF process. Kaleli [13] introduces a novel entropy-based neighbor selection approach depending on users’ own interpretations about a rating domain in order to express their preferences. The author claims that such user interpretations can be estimated by using entropy of their preference vector and differences between user entropies can be used along with similarities in 0-1 knapsack problem to form optimum neighborhoods.

Some applications might require clustering which cannot be solved by utilizing a classical method. Therefore, in order to optimize more than one criteria in a clustering problem, multicriteria clustering approaches are proposed. Delattre and Hansen [29] analyze bicriteria clustering in their work and propose some solutions. Similarly, Ferligoj and Bateglj [14] offer different ways of how to transform multicriteria into a single one. Smet and Eppe [30] focus on multicriteria relational clustering and propose a method building k-means clustering and relations between these clusters on the basis of a binary outranking matrix. Since multicriteria clustering is employed in various data mining applications, Thonnard et al. [31] present a multicriteria clustering approach that aims to solve a problem known as attack attribution in data mining.

Although there are various clustering-based CF methods, our work differs from them with the following aspects: Firstly, there is no previous study concerning clustering as a bicriteria approach in CF. However, it is shown that, all possible information gathered from user-item matrix is needed to be employed in order to boost prediction accuracy [13]. Therefore, our work is novel due to introducing a bicriteria clustering approach for CF. In addition, the
The proposed method can improve accuracy of $k$-means-based CF scheme’s accuracy. Finally, this work can be considered as a first step of bicriteria clustering methods for CF. We believe that our approach can be employed with other center-based clustering methods in recommender systems.

3 TRADITIONAL USER-BASED RECOMMENDATION ALGORITHM

In a traditional $k$-NN-based CF algorithm, predictions are produced by utilizing a matrix $D$ holding $n$ users' preferences on $m$ products. Such user-based algorithms are developed based on an assumption that each entity in the preferences data do not have the same utility for a particular recommendation process [3]. Thus, when an active user ($a$) requests a prediction for a target item ($q$), $k$-NN-based CF algorithm firstly determines $k$-NN of $a$ or $q$, and the users in these neighborhoods are involved into further steps in prediction process. Correlations among users are computed using Pearson’s correlation coefficient (PCC), as given in Eq. 1 [3].

$$w_{au} = \frac{\sum_{j \in J}(v_{aj} - \bar{v}_a)(v_{uj} - \bar{v}_u)}{\sigma_a \sigma_u} \quad (1)$$

where $v_{aj}$ and $v_{ui}$ are the ratings of user $a$ and $u$ for item $i$, $J$ indicates the set of commonly rated items by both users, and $\bar{v}_a$ and $\bar{v}_u$ are mean ratings of user $a$ and $u$, respectively. After correlations are computed, firstly, $k$-NN of $a$ is formed, prediction for $q$ ($p_{aq}$) is computed, as follows:

$$p_{aq} = \bar{v}_a + \frac{\sum_{u \in S}(v_{uq} - \bar{v}_u)w_{au}}{\sum_{u \in S}w_{au}} \quad (2)$$

where $S$ is the set of $a$’s neighbors who rated $q$. The overall $k$-NN-based CF process is described in Fig. 1.

![Figure 1. Overall Process of k-NN Prediction Algorithm](image)

4 COLLABORATIVE FILTERING VIA BICRITERIA-BASED $K$-MEANS CLUSTERING

Clustering-based CF solutions are useful for recommender systems due to their valuable assistance in order to improve scalability. If a clustering method is need to be applied, $k$-means clustering algorithm comes to mind since it is probably the most well-known clustering approach. In a classical $k$-means clustering, initially, $k$ random centers are chosen from entire data for each cluster [24]. Afterwards, rest of the objects are compared with each center by means of discrimination criterion, e.g., Euclidean distance, similarity, correlation, and assigned to
the closest cluster. This procedure is performed repeatedly and at each stage, cluster centers are then estimated as the average of objects assigned to corresponding cluster. Clustering might be terminated either by reaching a pre-defined iteration number or observing the modification in cluster centers between successive stages is close to zero or less than a default value. With completion of clustering process, all members of data are assigned to only one cluster.

In user-based CF applications, as previously mentioned, the crucial part of algorithms is forming of a’s neighborhood. k-means clustering algorithm is particularly employed as an off-line preprocess step to gather similar users in the same cluster in order to enhance online performance of recommendation process. After a requests a prediction for a particular item, clustering-based CF algorithm firstly determines a’s cluster, than, as well as all users in that clusters can be considered as k-NN of a, also the most similar pre-defined number (s) of them might be selected to form neighborhood. Note that, in whole k-means-based CF process, only one discrimination criterion, e.g., PCC, is employed.

As previously mentioned, Kaleli [13] claim that it is possible to improve neighbor selection capability of user-based CF algorithms by employing an additional measure DU along with user correlations. DU measure is utilized to reveal users’ interpretations about a rating domain. According to the study [13], some users might express their admirations about products in a binary manner by only employing 1 or 5 rating values even though CF service provider has a 5 star preference domain. On the other hand, there might be other users employing all members of rating domain in order to express their feelings. Consequently, DU measure provides additional information about users and we can conclude that users having minimum DU difference and maximum similarity might be considered as neighbors. DU measure depends on entropy in information theory and a user’s DU value is calculated as follows [13]:

\[
p_{uv} = \frac{\text{Number of vote } v \text{ in user } u \text{'s vector}}{\text{Total number of votes in } u \text{'s vector}}
\]

where \( p_{uv} \) shows probability of occurrence of vote \( v \) belonging to the rating domain (RD) in user \( u \)’s preference vector. After each computing probabilities of each rating in RD, a user’s DU is estimated as follows:

\[
DU_u = \frac{\sum_{v \in RD} p_{uv} \log_2(p_{uv})}{\text{Total number of votes in } u \text{'s vector}}
\]

To date, all k-means clustering algorithm approaches consider only users’ correlations, similarities, or distances in order to construct clusters. In this paper, we introduce a new k-means clustering methodology in which users’ correlations and DU values determines centers and users’ cluster memberships, together. Since, we aim to group users having maximum
correlation and minimum DU differences in a cluster, a Pareto efficient clustering problem occurs [14]. In Pareto efficient clustering, it is not possible to find an optimum solution without sacrificing any of the given criteria and one of the ways to cluster given data according to criteria is converting them as a single criterion. In the case of user-based CF algorithm, we propose to combine correlations with DU differences for two users $a$ and $u$ as follows:

$$\text{ClsCrt}_{au} = \alpha \times w_{au} + (1 - \alpha) \times |DU_a - DU_u|$$

(4)

where $\text{ClsCrt}_{au}$ indicates the new clustering criterion and it can be employed in $k$-means clustering step of recommendation process instead of correlations.

5. Experimental Evaluation

We performed trials on two benchmark data sets, i.e., MovieLens (MLM) and Netflix (NF), in order to how bicriteria $k$-means clustering approach effects accuracy and scalability in user-based CF. MLM data set is introduced by GroupLens (http://www.grouplens.org) and it contains 3,952 users’ preferences on 6,040 movies. NF is also a movie database holding 480,189 users’ ratings for 17,770 movies. Since NF data set has huge number of users to be handled, we sampled 10,000 users from different density ranges. Detailed information about the data sets is given in Table 1.

Experiments were realized using 10-fold cross-validation experimentation methodology. Data sets were uniformly randomly partitioned into ten subgroups and at each iteration $i$ ($i = 1, 2, \ldots, 10$), associated subset was employed as the test set and the remaining ones were as the training set. After training and test sets were constructed, five rated items’ actual votes were withheld for each test (active) user. Such entries were replaced with null; their values were tried to be predicted, and estimations were compared with actual values.

Table 1. Descriptions of data sets

<table>
<thead>
<tr>
<th>Name</th>
<th>User x Item</th>
<th>Rating Scale</th>
<th>Total Ratings</th>
<th>Density (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLM</td>
<td>6,040 x 3952</td>
<td>5-star</td>
<td>1M</td>
<td>4.25</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>10,000</td>
<td>2,337,295</td>
</tr>
<tr>
<td>NF</td>
<td>x 17,770</td>
<td>5-star</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

CF-based systems’ accuracy can be evaluated through several metrics. In this study, we utilized mean absolute error (MAE), which measures how close the predictions are to the actual ratings as an average of absolute errors. Thus, the lower the MAE, the better the results are. In order to show how accuracy results are statistical significant, we also performed $t$-tests.
In the first group of experiments, we tried to compare accuracy level of classical \(k\)-means-based CF (CKMCF) and selected \(k\)-means-based CF (SKMCF) on both data sets. We aimed to determine our base clustering methodology for CF. During experiments we set number of clusters (\(c\)) to 1, 2, 3, 5, 7, and 10 and \(s\) to 60. Note that, 1 assigned to \(c\) indicates that there is no clustering and \(s\) shows the number of users, selected from \(a\)’s cluster, assigned as NN of \(a\). We repeat the experiments 100 times and displayed the results in Fig. 2 and Fig. 3 for MLM and NF, respectively.

According to results presented in Fig. 2 and Fig. 3 for both data sets SKMCF approach has better accuracy level than CKMCF method. Traditional user-based CF algorithm’s MAE values are 0.7478 and 0.7587 for MLM and NF, respectively. If we employ classical \(k\)-means clustering in prediction producing process, we obtain the best accuracy level with \(c\) equals 2 for both data sets and MAE values are 0.7563 and 0.7745 for MLM and NF, respectively. On the other hand, by utilizing modified version of \(k\)-means (SKMCF), it is possible to decrease error of predictions to 0.7489 and 0.7623 for MLM and NF, respectively. Since probability of losing required information among users increases with greater \(c\) values, MAE values for \(c\) bigger than 2 become worse. Eventually, we can conclude that SKMCF method can boosts accuracy of CKMCF approach for recommendation process.

After determining clustering methodology, we performed experiments in order to measure accuracy performance of our bicriteria \(k\)-means clustering approach for CF. In the experiments, we set \(c\) to 2 for both data sets and varied \(a\) from 0.1 to 0.9. Thereby, we tried to find a balance between user similarities and DU difference values to determine a single criterion. We run experiments 100 times for both data sets using the same 10-fold methodology and presented the outcomes in Fig. 4.
The displayed results shows that it is possible to improve SKMCF methods’ accuracy by employing two criteria instead of a single one. Also, according to outcomes, it is possible to improve traditional user-based CF algorithm’s accuracy by setting $\alpha$ to 0.7 and 0.6 for MLM and NF, respectively. As a remainder, traditional user-based CF has MAE values 0.7563 and 0.7745 for MLM and NF, respectively. On the other hand, by employing our bicriteria clustering approach, it is possible to improve accuracy and obtain MAE values as 0.7322 and 0.7435 for MLM and NF, respectively. Moreover, applied $t$-test results indicates that, improvements in the prediction accuracy are statistically significant at 95% confidence level. The experiment results also show that, similarities among users are more valuable than DU differences. However, if we yield a balance between such measures, the resulted criterion is more beneficial than similarities and DU differences. Consequently, it is shown that utilizing any possible information obtainable from user preference matrix is so important for success of a recommender system.

6 CONCLUSION
In this paper, we introduce a new clustering methodology for collaborative filtering. Up to present, all clustering approaches proposed for recommender systems employ a single criterion in order to group users. Since, clustering of any data might cause to lose useful information, traditional clustering algorithms suffer from decrease in accuracy level. On the other hand, it is possible to improve clustering performance by considering all possible information belonging to users. Entropy-based degree of uncertainty measures users interpretations about a rating domain. In a traditional recommender system, since we have only users’ correlations information, we aim to group similar users in the same cluster. Whereas, we also need to consider having minimum degree of uncertainty in a cluster. Therefore, we propose a novel bicriteria clustering approach in which we convert two clustering criteria into a single one. We employ our approach on $k$-means-based collaborative filtering algorithm and we obtain that it is possible to improve accuracy of predictions by employing two clustering criteria instead of a single one.
In future, we are planning to employ our approach with other clustering algorithms proposed for collaborative filtering. Also, we are considering to extract our possible information from user-item vector and introduce a multi-criteria clustering for recommender systems.

7 REFERENCES


