


Comparison of The Performances of Clustering and Dimensionality Reduction Approaches in Collaborative Filtering

Özge Tas^{1,*} 

¹ Cappadocia University, Cappadocia Vocational School, Department of Computer Programming

Abstract

Recommendation systems (RS) can be defined as systems that aim to offer personalized product and service recommendations to users based on users' past product preferences and similarities with other users in the system, especially in systems that provide e-commerce services. The main purpose of RS is to reveal meaningful information from large-scale data to users and to recommend systems that aim to simplify the analysis of user behaviors and product attributes. It is possible to divide the techniques used in RS into two main categories content-based and collaborative filtering (CF) according to the information they receive as input. Content-based recommendation systems focus on analyzing the attributes of items such as articles, movies or music to generate tailored recommendations. CF methods analyze user-generated scores for products and services to identify patterns and preferences. The success of CF techniques hinges on accurately identifying user similarities within large datasets. However, in CF techniques, large-scale data sets consisting of a large number of users and the scores given by users to these products are used. Consequently, identifying user similarities in such extensive datasets poses significant challenges. Two different methods are used to overcome this problem. The first method applies clustering analysis to divide the dataset into smaller subsets (user or product), followed by the application of CF techniques. In the other method, dimensionality reduction is performed on a product (object) basis using Singular Value Decomposition (SVD) and Principal Component Analysis (PCA) methods. Up to now, many studies have been carried out using clustering analysis and variable dimensionality reduction methods. Despite extensive research, a thorough comparison of clustering and dimensionality reduction methods on real-world datasets remains unexplored. This study aims to compare the performances of eleven clustering techniques of eleven clustering techniques, four of which are non-hierarchical seven of which are hierarchical clustering algorithms, and two variable dimensionality reduction techniques, consisting of SVD and PCA METHODS, in CF.

Keywords: *Recommender Systems, Collaborative Filtering, Cluster Analysis, Dimension Reduction, Big Data.*

1. Introduction

Collaborative Filtering (CF) techniques are categorized into two main types: model-based and memory-based. Model-based CF techniques are based on estimating a parametric model suitable for the training data set consisting of the scores that users give to products and predicting the scores that active users can give to products using this model. In these techniques, methods such as Bayesian Networks, Regression Analysis, Clustering Techniques, and Rule-Based and Latent Semantic Models are generally used for modeling [1-6]. Memory-based CF techniques are also divided into two main categories: user-based and object-based. User-based CF techniques assume that the best way to find products that may be of interest to the active user is to identify other users with similar interests [7]. Therefore, the first step in these methods is to identify users with whom the active user is similar. In the second step, the scores that the active user may give to the products are estimated based on the scores given to the products by the neighboring, i.e. the most similar users. Products with high predicted scores are recommended to the active user. Object-based CF techniques have a similar working principle. However, similarities between objects (products) are calculated instead of similarities between users [8]. Hybrid recommender systems integrate the strengths of content-based and CF techniques and eliminate the shortcomings arising from the individual use of these methods. Various methods, such as weighting, blending, and cascading, are employed to combine these techniques. However, hybrid systems are generally based on the use of CF techniques to score both products and their contents and to estimate the score that the active user can give to the products. Comparison of the performance of clustering and size reduction methods is necessary to improve the effectiveness of recommender systems. These comparisons help to understand under what conditions different algorithms and techniques give better results. For example, hierarchical clustering methods give better results in certain situations, while other methods such as K-Means can offer faster results [9]. Therefore, considering the advantages and disadvantages of both

*Corresponding author

E-mail address: ozge.tas@kapadokya.edu.tr

Received: 7/Dec/2024; Accepted: 28/Dec/2024.

approaches, choosing the most appropriate method will increase the success of recommendation systems. In this study, we provide an in-depth analysis of collaborative filtering techniques in recommender systems. In collaborative filtering systems, there are some disadvantages such as the high dimensionality of the data and the identification of products and users. Some techniques are used to minimize the computational risks of these disadvantages. Some of these techniques aim to reduce the number of users by clustering, while others aim to reduce the number of objects by using techniques such as Singular Value Decomposition (SVD) and Principal Component Analysis (PCA). Some of the studies on CF based on dimensionality reduction can be summarized. Dimensionality techniques play an important role, especially in large data sets. Big data is a factor affecting the performance of recommendation systems, and therefore appropriate pre-processing and size reduction methods need to be applied [10]. For example, the K-Means clustering algorithm makes datasets more manageable while also improving the accuracy of recommender systems [11]. In addition, size reduction methods help to obtain more meaningful results by reducing the noise in the dataset [12]. The researchers investigated the effect of clustering methods of K-means, SOM, and Fuzzy C-Means (BCO) on the predictive performance of CF. For this purpose, they used the MovieLens dataset. As a result of their study, they concluded that the prediction performance of the BCO clustering algorithm is better [13]. Chen et al. [14], proposed a new CF technique based on evolutionary heterogeneous clustering. To evaluate the performance of the proposed technique, raw CF (without clustering method), CF based on K Means and CF based on their proposed clustering technique were applied to MovieLens and CiaoDVD datasets and it was observed that the proposed clustering method improved the prediction performance of CF. Liao and Lee [15], proposed a new CF technique based on self-constructed clustering. Ba et al. [16], used the CF approach that combines CCA and clustering. They compared the performance of CF based on SVD and clustering, CF based on SVD and traditional CF techniques and concluded that the performance of the proposed approach is better. However, in most of the mentioned studies or studies conducted for similar purposes, both a small number of data sets and a limited number of clustering and dimensionality reduction techniques were used. This study aims to compare the performance of CF techniques based on clustering analysis and dimensionality reduction techniques on real data sets. For this purpose, 11 clustering techniques (7 hierarchical and 4 non-hierarchical), 2 dimensionality reduction techniques (SVD and PCA) and 9 real data sets were used.

2. Materials and Methods

2.1. Cluster Analysis

Cluster analysis identifies natural groupings within a distributed dataset. Cluster analysis identifies inherent groupings within a distributed dataset, aiming to maximize intra-cluster similarity while minimizing inter-cluster similarity. There are many cluster analysis techniques in the literature, and it is possible to group these techniques under different headings according to different criteria.

2.1.1. Hierarchical Clustering Analysis Methods

Hierarchical clustering methods measure the similarities between data points. Based on these measurements, similar data points are merged, while dissimilar ones are separated. As can be understood, clusters are formed in a stepwise manner in these methods. There are two types of hierarchical clustering methods in the literature: additive and partitional. In agglomerative clustering methods, all individuals in the data set are initially considered as a separate cluster. Then, the closeness or distance between individuals is calculated. At each step, clusters that are close to each other are merged and this process continues until there are no clusters to be merged according to the predetermined clustering criterion. Agglomerative clustering methods therefore have a top-down approach.

Partitioning clustering has the opposite working principle of agglomerative clustering. In this method, the entire data set is initially treated as a single cluster. Similarly, the distances between individuals within the same cluster are calculated and the distant individuals are separated. This process continues until there are no clusters to be separated according to the predetermined clustering criterion. In practice, additive clustering is much more common than partitional clustering [17]. The biggest advantage of hierarchical clustering methods is that the number of clusters is automatically determined by the algorithm and therefore the number of clusters for the data set does not need to be known in advance. However, the necessity of calculating the distances between all individuals and repeating this calculation at each step makes it difficult to use hierarchical clustering methods, especially in large-scale data sets. Agglomerative hierarchical clustering processes are commonly visualized using dendrograms, which depict the hierarchical structure in a tree-like format. As an example, **Figure 1** shows the clusters obtained because of applying the agglomerative hierarchical clustering method to a data set consisting of eight data points.

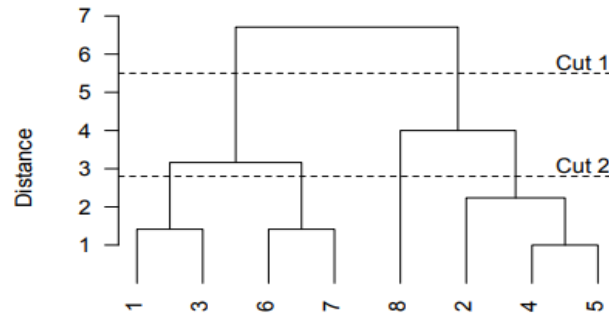


Figure 1. An example of a dendrogram for hierarchical clustering of eight observations is given for two cuts $K=2$ (Cut 1) and $K=4$ (Cut 2) [18].

In **Figure 1**, the values on the x-axis indicate the data points, while the values on the y-axis indicate the proximity (distance, remoteness) between the clusters. Accordingly, clusters are formed according to a predetermined distance value. For example, if the distance value is given as 2 units, the individuals that merge under 2 distance units in the dendrogram form clusters. According to **Figure 1**, when the distance is given as 2 units, the clusters obtained are Cluster 1 = {1,3}, Cluster 2 = {6,7}, Cluster 3 = {8}, Cluster 4 = {2}, Cluster 5 = {4,5}.

The standard algorithm followed by additive clustering methods is shown below. This algorithm starts with n clusters and iteratively merges clusters until only one cluster remains.

Algorithm 1: Standard algorithm for combinatorial clustering.

1. Start the algorithm with n clusters, where n is the number of individuals in the dataset.
2. Proximities between all individuals are calculated.
3. The two closest clusters are merged.
4. Steps 2 and 3 are repeated by decreasing the number of clusters by one.
5. There are seven different hierarchical clustering calculation methods based on this algorithm.

Table 1. Calculations used in clustering analysis

Clustering Methods	Formulas	
Single Linkage (TeB) [26]	$d_{k(i,j)} = \min (d_{kj}, d_{ki})$	(1)
Complete Linkage (TB) [31].	$d_{k(i,j)} = \max (d_{kj}, d_{ki})$	(2)
Centroid Linkage (MeB) [31].	$d_{ij} = \ \bar{x}_i - \bar{x}_j\ $	(3)
Average Linkage (OB) [31].	$d_{ij} = \frac{\sum_{k=1}^{n_i} \sum_{l=1}^{n_j} d(x_{ik}, x_{jl})}{n_i * n_j}$	(4)
Median Linkage (MB) [31].	$d_{ij} = \ \bar{x}_i - \tilde{x}_j\ _2$ $\tilde{x}_i = (1/2)(\bar{x}_k + \bar{x}_l)$	(5)
Weighted Average (AB) [31].	$d_{k(i,j)} = \frac{n_i}{n_i+n_j} d_{ki} + \frac{n_j}{n_i+n_j} d_{kj}$	(6)
Ward's Method (WB) [31].	$d_{ij} = \sqrt{\frac{2n_i n_j}{(n_i+n_j)}} \ \bar{x}_i - \bar{x}_j\ $	(7)

2.1.2. Non-Hierarchical Clustering Cluster Analysis Methods

Non-hierarchical clustering methods are based on starting from an initial clustering and iteratively repeating the process until the optimal cluster structure is found. In these methods, the number of clusters should be determined by the researcher in advance. In this study, non-hierarchical clustering methods such as K-Means, K-Medoid, Fuzzy C-Means and Self-Organizing Mapping (SOM) clustering methods are presented.

K-Means (KO): K-Means (KO) is a non-hierarchical clustering method widely used in many applications. The name K-Means comes from the fact that there are k clusters and the center of each cluster corresponds to the arithmetic mean of the clusters [19]. The KO clustering algorithm is based on minimizing the objective function given in Eq. 8.

$$J(X, V) = \sum_{j=1}^k \sum_{x \in S_j}^n \|x - v_j\|^2 \quad (8)$$

K-Medoidler (KM): Since the KO clustering algorithm is based on the arithmetic mean, it is highly sensitive to outliers and noisy values in the data set. To alleviate this disadvantage of the KO algorithm, Kauffman and Rousseeuw [20], proposed the KM clustering algorithm. In this algorithm, the cluster centers use the center point of the regions where clusters are dense, called medoids, instead of the arithmetic mean. It shows the difference between the cluster centers of KO and KM clustering algorithms. Medoid is calculated in Eq. 9:

$$z_j = \min \left(\sum_{t=1}^{n_j} \sum_{k=1}^{n_j} \|x_k - x_t\|^2 \right) \quad (9)$$

$$v_j = x_{z_j}$$

Fuzzy C-Averages (BCO): KM and KO clustering algorithms are based on classical logic. In other words, in these methods, an individual belongs to one and only one cluster. Therefore, these methods force an individual to belong to only one of the clusters, even if the individual is equidistant from more than one cluster. The BCO clustering algorithm is based on fuzzy logic. Therefore, BCO allows an individual to belong to multiple clusters simultaneously with different degrees of belonging. Here, membership degrees are used to determine how much individuals belong to the clusters and how much they have the characteristics of the clusters. In all clustering methods based on fuzzy logic, membership degrees have the following properties.

$$0 \leq u_{ij} \leq 1 \quad i=1,2,\dots,n \quad j=1,2,\dots,c$$

$$\sum_{j=1}^c u_{ij} = 1 \quad \forall i \quad (10)$$

$$\sum_{i=1}^n u_{ij} > 0 \quad \forall j$$

Similar to KO and KM, BCO is based on the minimization of an objective function and the objective function for this algorithm is in Eq. 11:

$$J(U, V, X) = \sum_{i=1}^n \sum_{j=1}^c u_{ij}^m d_{ij}^2(x_i; v_j) \quad (11)$$

In the equation, m is the fuzziness index, $d_{ij}^2(x_i; v_j)$, is the Euclidean distance between individual i and cluster center j . The update equations for cluster centers and membership degrees that minimize the objective function are obtained in Eq. 12.

$$u_{ij} = \frac{1}{\sum_{k=1}^c (d_{ij}^2 / d_{ik}^2)^{1/(m-1)}} \quad i = 1, 2, \dots, n \quad j = 1, 2, \dots, c \quad (12)$$

$$v_j = \frac{\sum_{i=1}^n u_{ij}^m x_i}{\sum_{i=1}^n u_{ij}^m} \quad j = 1, 2, \dots, c \quad (13)$$

Self-Organizing Mapping (SOM): SOM is a clustering algorithm with an architecture similar to artificial neural networks. It was proposed in 1995 by Kohonen [21]. It is therefore also known as Kohonen networks. SOM transforms high-dimensional data sets into a two-dimensional map. SOM consists of two layers, input and output. The input layer contains the number of features of the individuals to be clustered, and the output layer contains as many neurons as the number of clusters.

$$\varphi_{zk} = \exp\left(\frac{\|r_k - r_z\|^2}{\sigma^2(t)}\right) \quad (14)$$

2.2. Size Reduction Methods

2.2.1. Principal Components Analysis

PCA is one of the linear size reduction methods based on the covariance matrix of variables. The main goal in PCA is to transform the p variable in the original data set into a smaller number of orthogonal linear

components with the highest variance explanation rate, so that the relationship between the variables is eliminated [22]. As can be understood from this, the correlation between the linear components obtained at the end of PCA, also called the principal component, is zero. The size reduction with PCA can be summarized as p-dimensional X_1, X_2, \dots, X_p let be the original data matrix. As mentioned earlier, the main purpose of the PCA method is to find the basic components in its shape. Mathematically, PCA is based on the spectral decomposition of the covariance matrix, which is defined in eq. 15. $TB_1, TB_2, \dots, TB_{d \ll p}$

$$\Sigma = AAA' \tag{15}$$

In other words, the principal components are generally calculated in eq. 16:

$$TB = A'(X - \mu) \tag{16}$$

After calculating the principal components as given in Eq. 16, the dimensionally reduced data is obtained by selecting the number of principal components that explain most of the variance [23].

2.2.2. Singular Value Decomposition

SVD is one of the most widely used dimension reduction techniques in CF. SVD is basically based on decomposing an NxN matrix X (where n is the number of users and p is the number of products and objects) into 3 matrices in eq. 17:

$$X = U_{n \times p} \lambda_{p \times p} V^T_{p \times p} \tag{17}$$

2.3. Recommendation Systems

GLCM Recommender systems are filtering systems that are used to generate information based on the behavior of users, to examine their interests and behaviors, and to predict the products they may be interested in by using the information entered online [24, 25]. There are different ways to design recommendation systems. The first and simplest of these is to provide a streaming style service by directing recommendation around the content stream. Examples of this are music services such as Spotify and Youtube music. After each item, the user is allowed to evaluate the item and the user is presented with content based on these evaluations. These reviews influence the algorithm for the next song or item. This process is repeated until the user leaves the platform. Another way is the catalog-based website method. An example of this is the Netflix website, which is one of the movie websites. It helps users to make ratings on the movie and then categorize the movie content and have information about the movies. There are usually pages dedicated to each movie with detailed movie content specifications. This collaborative filtering method is further enhanced by using algorithms to encourage users to rate any movie they see, and to generate prediction ratings of personalized features next to the movie cover image in the detailed information. These predictions help users quickly decide whether a movie is worth learning more about. As a whole, recommendation systems can be divided into 3 types, content-based recommender systems are used to recommend new objects and information that may be of interest to users by taking into account the content information of objects that have previously attracted users' attention on the Internet and the basic characteristics of users. Collaborative filtering systems usually take into account the scores that users give to products or objects in the system. The main purpose here is to examine the objects that users give high scores and suggest new objects to them. Hybrid Methods are a combination of contextual and collaborative filtering systems.

2.3.1. Collaborative Filtering Methods

The CF method is the most successful recommendation system [26]. The main purpose of the CF methods is to make suggestions and predictions to the active user based on the opinions of like-minded users. In CF, user opinions can be obtained directly or implicitly from users differently.

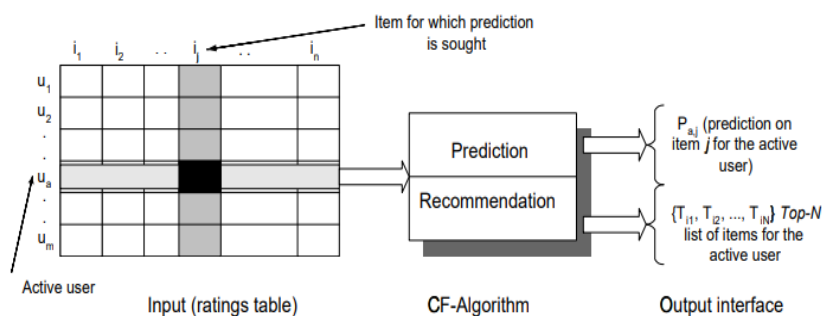


Figure 2. The collaborative filtering process [35].

Collaborative filtering algorithms represent all $m \times n$ user-item data as a rating matrix. Each entered image in user set A represents the preference score of user i on item j . Researchers have developed collaborative filtering algorithms as memory-based (user-based) and model-based (item-based) algorithms [3].

There are many different methods for calculating the similarity or weight between users and items. Generally, in the similarity calculation, the number of users is considered as the size of the active user's neighborhood relations, and the similarity-based collaborative filtering method is considered as neighbor relation-based collaborative filtering.

Correlation-based similarity: In this case, the similarity $w_{u,v}$ between user u and v and the similarity $w_{i,j}$ between two items i and j are measured using Pearson's correlation or other correlation measures. Pearson correlation measures the relationship between two variables. User-based algorithm for similarity between users u and v ;

$$w_{u,v} = \frac{\sum_{i \in I} (r_{u,i} - \bar{r}_u)(r_{v,i} - \bar{r}_v)}{\sqrt{\sum_{i \in I} (r_{u,i} - \bar{r}_u)^2} \sqrt{\sum_{i \in I} (r_{v,i} - \bar{r}_v)^2}} \quad (18)$$

The item-based algorithm calculates the Pearson correlation of the degrees of the items i and j for the user set $u \in U$,

$$w_{i,j} = \frac{\sum_{u \in U} (r_{u,i} - \bar{r}_i)(r_{u,j} - \bar{r}_j)}{\sqrt{\sum_{u \in U} (r_{u,i} - \bar{r}_i)^2} \sqrt{\sum_{u \in U} (r_{u,j} - \bar{r}_j)^2}} \quad (19)$$

Some variations of item-based and user-based Pearson correlations can be found [15]. Pearson correlation calculation is widely used in collaborative filtering.

Cosine-based similarity measures the similarity between two documents by treating each document as a vector of word frequencies and calculating the cosine of the angle formed by the frequency vectors. Generally, this type of similarity is preferred for collaborative selection based on users and items rather than on the frequencies of documents and ratings.

If R is an $n \times m$ -dimensional user item matrix, the similarity between two items i and j is calculated as the cosine of the n -dimensional vectors corresponding to items i and j in the R matrix.

Cosine similarity between element vectors i and j ,

$$w_{i,j} = \cos(\vec{i}, \vec{j}) = \frac{\vec{i} \cdot \vec{j}}{\|\vec{i}\| * \|\vec{j}\|} \quad (20)$$

Jaccard similarity is the easiest way to calculate the similarity between two users. It looks at the common items rated by both users, regardless of their ratings from the user (Charikar, 2002). Jaccard similarity is useful when items do not receive a reliable rating.

$$sim_{u,i} = \left| \frac{r_{u,i} \cap r_{v,i}}{r_{u,i} \cup r_{v,i}} \right| \quad (21)$$

The most important part of collaborative filtering is to determine the recommendation for the active user. Once the active user items are identified, the following two different equations are used to determine the user and item score.

$$r_{u,i} = \bar{r}_u + \frac{\sum_{v \in U} (r_{v,i} - \bar{r}_v) \cdot sim_{u,v}}{\sum_{v \in U} sim_{u,v}}, \quad (22)$$

$$r(a,p) = 1/n \sum_{i=1}^n r_{ip}, \quad (23)$$

3. Application and Result

The This study aims to compare the performance of collaborative filtering methods based on clustering and dimensionality reduction techniques. For this purpose, 9 real data sets were used. The data sets and their characteristics are given in **Table 2**.

Table 2. Data sets used and their features.

Dataset	Number of users	Number of Objects	Scoring
Anime	200	16384	1 with 10 between
BookDataset	671	150	1 with 10 between
Jester	24938	100	-10 with +10 between
Laptop	671	300	0 with 5 between
Mobile	671	183	0 with 5 between
Movie	943	1683	0 with 5 between
Restaurant 1	139	2551	0 with 2 between
Restaurant 2 (Lunch)	139	2551	0 with 2 between
Restaurant 3 (Service)	139	2551	0 with 2 between

Anime, Jester, Restaurant 1, Restaurant 2, Restaurant 3 data sets were downloaded from Link 1 [27] and BookDataset, Laptop, Mobile, Movie data sets were downloaded from Link 2 [28], Link 3 [29]. For performance comparisons, 6 different scenarios were run. For each scenario, 90% of the dataset was selected as the users in the system and 10% as the active user whose score would be calculated. From the 10% portion, the score of the products that each user rated was estimated and compared with the actual values. Three goodness-of-fit measures were used as comparison criteria: Root Mean Square Error (HKOK) in eq. 24, Mean Absolute Percentage Error (OMYH) in Eq. 25 and Mean Absolute Error (OMH) in Eq. 26:

$$HKOK = \sqrt{\frac{\sum_{i=1}^n (gp_i - tp_i)^2}{n}} \quad (24)$$

$$OMYH = \frac{\sum_{i=1}^n |gp_i - tp_i| / gp_i}{n} * 100 \quad (25)$$

$$OMH = \sum_{i=1}^n \frac{|gp_i - tp_i|}{n} \quad (26)$$

Comparison results are presented based on the average performance across nine datasets, followed by specific evaluations of the Jester dataset, secondly according to the Jester dataset with the highest number of users, and thirdly according to the Anime dataset with the highest number of objects.

Comparison Results for Scenario 1

In this section, we compare the performance of CF techniques based on BCO, KO, KM, SOM, WB, OB, MeB, MB, AB, TeB, TB clustering algorithms. Here, each clustering technique was run with 20, 30, 40, 50 and 60 clusters respectively. According to the average results obtained for all data sets for Scenario 1,

- The results indicate that the number of clusters does not significantly influence performance across all goodness-of-fit metrics,
- Across all scenarios, the BCO algorithm demonstrates the highest average performance,
- The performance rankings of the clustering methods are BCO, WB, TB, KO, TeB, MeB, AB, OB, SOM, KM, MB according to the HKOK criterion, BCO, WB, TeB, MB, OB, MeB, TB, KM, KO, AB, SOM according to the OMYH criterion, and BCO, WB, TeB, KO, MeB, KM, SOM, MB, TB, OB and AB according to the OMH criterion. Thus, for Scenario 1, it can be said that the WB clustering algorithm also provides better prediction results than the other methods, while SOM, AB and MB generally perform poorly.

For the Jester data set,

- When comparisons are made according to the HKOK and OMH criteria, the top three best-fit clustering algorithms are SOM, KO and BCO, respectively, and when comparisons are made according to the OMYH criterion, the top three algorithms are AB, OB and TB,
- The OLSR values are quite high for Scenario 1 and the OLSR values of the Jester dataset can significantly affect the overall performance,

- It is observed that the number of clusters does not have a significant effect on performance. For the anime dataset,
- It is seen that the CF technique based on the BCO clustering algorithm performs better than the other methods according to all three goodness of fit criteria, while the worst fit is obtained from the SOM clustering method.
- Apart from this, it can be said that the WB clustering method also performs well for the Anime dataset.

Comparison Results for Scenario 2

For the comparisons in this section, SVD dimension reduction technique was first applied to all datasets. In the next stage, users were clustered separately using 11 clustering algorithms and scores were calculated according to the clusters. According to the average goodness of fit values obtained from all datasets,

- Compared to non-hierarchical clustering methods in all three goodness-of-fit criteria, CF methods based on hierarchical clustering methods have a higher prediction success,
- The highest success was obtained from the WB clustering method and the worst success was obtained from the SOM clustering method,
- Among the non-hierarchical clustering methods, BCO and KM are the most successful algorithms,
- It is observed that the number of clusters does not significantly change the prediction success.

SVD feature extraction method for the Jester dataset,

- The best fit is when WB is run with the HKOK and OMH criteria and TB with the OMYH criterion in combination with the clustering method,
- The worst prediction performance was achieved when run with the KO clustering method according to all three criteria.
- In general, hierarchical clustering methods perform better for the Jester dataset.
- The number of clusters did not have a positive effect on performance.

For the anime dataset, the best predictions according to all goodness-of-fit criteria were obtained by running SVD with the WB clustering method, while the worst predictions were obtained by running it with the KM clustering method.

Comparison Results for Scenario 3

The results in this section include the goodness of fit values obtained by first applying the PCA dimensionality reduction technique to all data sets, and then applying the 11 clustering methods to the reduced dimensionality data sets. When run with the PCA dimensionality reduction technique,

- According to all three goodness of fit criteria, the WB clustering technique provides the best estimation performance,
- The worst results are obtained from the BCO clustering method according to the HKOK and OMH criteria, and from the KM clustering method according to the OMYH criteria,
- The estimation results obtained from the hierarchical clustering methods are closer to the real values,
- It can be seen that there is no relationship between the number of clusters and performance.

According to the CF method, which was performed by first applying PCA and then cluster analysis methods to the Jester data set,

- When HKOK and OMH criteria are taken into consideration, the best estimation result is obtained from the WB clustering method, and according to the OMYH criterion, from the AB clustering method.
- The worst estimation result is obtained from the SOM clustering method according to all criteria.

For the anime dataset,

- It can be seen that the prediction performance of hierarchical clustering methods is quite good compared to non-hierarchical clustering methods according to all goodness of fit criteria,
- The WB clustering method provides the best prediction results,
- The worst prediction results are obtained from the KO clustering algorithm according to all criteria.

Comparison Results for Scenario 4

This section gives the estimation results obtained when 11 clustering algorithms are applied to the raw data and the score estimation is made according to Jaccard similarity.

- According to HKOK and OMH criteria, the best point estimate on average was obtained when the KO clustering algorithm was applied to the data sets, and according to OMYH criteria, AB clustering algorithm was applied.
- It is seen that BCO algorithm behaves differently compared to other clustering methods and has the worst estimation results.
- The performance of all algorithms except BCO and SOM improved as the number of clusters increased.
- WB clustering method showed good performance in Scenario 4 of CF estimation.

For the Jester dataset,

- According to the HKOK and OMH criteria, the best estimation results were obtained from SOM and KO, and according to the OMYH criteria, AB and MB clustering algorithms.
- According to the HKOK and OMH criteria, BCO and according to the OMYH criterion, SOM provided the worst performance.

For the anime dataset,

- The performance of hierarchical clustering methods is better,
- Of the non-hierarchical clustering methods, the BCO algorithm appears to provide the worst prediction performance for all criteria.

Comparison Results for Scenario 5

Scenario 5 is based on the combined use of the SVD size reduction technique, clustering algorithms, and the score calculation given by the Jaccard similarity. According to the results of Scenario 5,

- Hierarchical clustering methods are more successful,
- According to all three criteria, the SOM clustering algorithm gives the worst prediction results,
- The WB algorithm, which is one of the hierarchical clustering methods, provides the best performance according to all criteria,
- In hierarchical clustering methods, it can be seen that performance improves as the number of clusters increases.

As a result of applying Scenario 5 to the Jester dataset,

- On average, the best estimation results were obtained from the WB algorithm according to the HKOK and OMH criteria, and from the AB algorithm according to the OMYH criteria.
- The worst performance was obtained from the SOM algorithm according to all criteria.

As a result of the application of Scenario 5 to the anime dataset,

- The best estimation results were obtained from the MeB clustering method according to all criteria, and the worst performance was obtained from the SOM clustering method.
- Apart from this, the success of hierarchical clustering methods is quite high compared to non-hierarchical methods.

Comparison Results for Scenario 6

In this section, the results obtained as a result of using PCA as a size reduction method and Jaccard similarity as a score estimation are included. Scenario 6 for average results

- Hierarchical clustering methods are more successful in prediction,
- That the worst predictions are derived from the SOM algorithm
- For non-hierarchical clustering methods, the most successful method is KM,
- It can be seen that there is no significant relationship between the number of clusters and performance.

Results on the Jester Dataset The results obtained as a result of the application of Scenario 6 to the Jester dataset are as follows. Hierarchical clustering methods have produced prediction values that are closer to reality. BCO and SOM showed worse predictive performance than other methods. Among the hierarchical clustering methods, the best estimation success was obtained from the WB method according to the HKOC and OMH criteria, and from the MB according to the OMYH criteria.

For the anime dataset,

- Similar to the general mean and Jester data, the best estimates are obtained from the CF technique based on hierarchical clustering,
- Among the non-hierarchical clustering methods, BCO has the worst performance and SOM has the

best performance,
Overall, BCO has the worst predictive success in almost all cluster numbers.

3.1. Comparison of Overall Performance of Clustering Methods and Size Reduction Techniques

In this section, firstly, the CF performances of clustering methods for 9 data sets, 6 different scenarios, and a total of 54 different situations were compared. **Figure 3** shows the average goodness of fit values by averaging 54 different conditions. When **Table 3** and **Figure 3** are examined, it can be concluded to say that the overall prediction performance of hierarchical clustering methods is better compared to non-hierarchical clustering methods. Apart from this, in **Table 3** and **Figure 3**, the best prediction results are obtained from the WB clustering algorithm on average, and the worst performance is obtained from the SOM algorithm according to the HKOK and OMYH criteria, and the KM algorithm according to the OMH criteria. Looking at the box-plot graphs, it is seen that the number of clusters does not have a significant effect on performance.

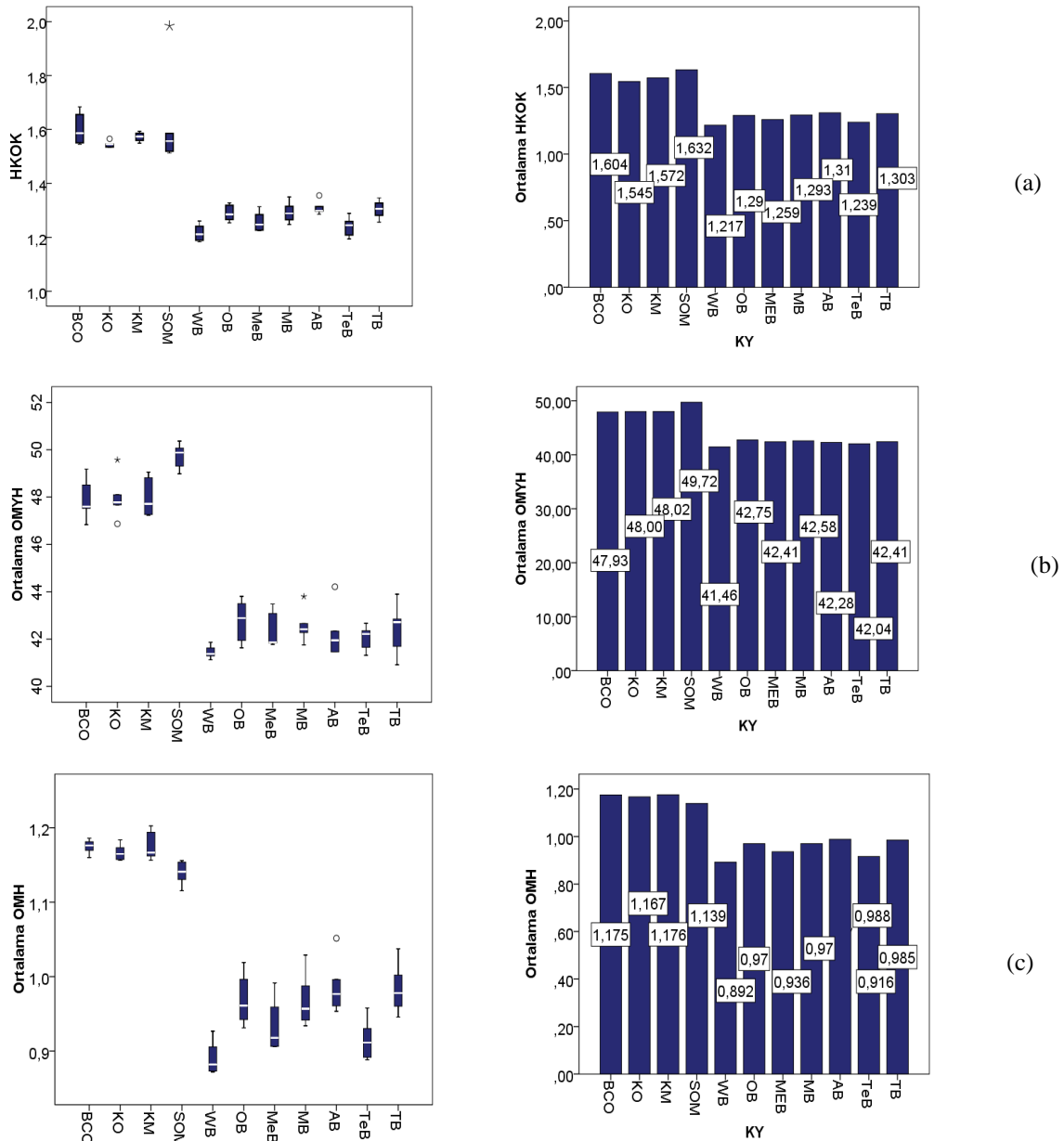


Figure 3. Overall mean values of goodness of fit measures (a) Average HKOK (b) Average OMYH (c) Average OMH

Table 3. Overall average of goodness of fit measures for size reduction methods

CN	HKOK		OMYH		OMH	
	SVD	PCA	SVD	PCA	SVD	PCA
20	1.430	1.475	45.720	47.930	1.061	1.121
30	1.408	1.471	45.080	46.939	1.037	1.107
40	1.375	1.499	44.521	47.061	1.015	1.093
50	1.375	1.441	44.792	47.321	1.011	1.091
60	1.366	1.564	44.221	46.784	1.005	1.095
Mean	1.391	1.490	44.867	47.210	1.026	1.101

Looking at the chart, it is seen that the values of the goodness of fit criteria of the SVD for all cluster numbers are lower than the PCA. From this, it is possible to say that the prediction performance of the SVD size reduction technique is better on average.

3.2. Average Comparison Tests

In this section, the non-parametric Wilcoxon test was performed to test whether the difference between the performances of the scenarios was statistically significant. **Table 4** shows the test results according to the HKOK criterion.

Table 4. Wilcoxon Test significance values according to the HKOK criterion

	Scenario1	Scenario2	Scenario3	Scenario4	Scenario5	Scenario6
Scenario1	-	0.657	0.594	0.003	0.010	0.328
Scenario2	-	-	0.003	0.003	0.006	0.006
Scenario3	-	-	-	0.003	0.003	0.004
Scenario4	-	-	-	-	0.010	0.003
Scenario5	-	-	-	-	-	0.197
Scenario6	-	-	-	-	-	-

Important conclusions that can be drawn from **Table 4** are as follows. Since the significance values were greater than 0.05, there was no statistically significant difference between Scenario 1-2, Scenario 1-3, Scenario 1-6, Scenario 5-6. There was no significant difference between the performances of the scenarios using Eq. 22 for score estimation. From this, it is possible to say that using raw data or reduced data does not have a significant effect on performance in scenarios where Eq. 22 is used for score estimation. If Eq. 23 is used for score estimation, there is a significant difference between the performances of the raw data and the use of data with reduced size with TKA and the use of raw data and PCA with reduced size.

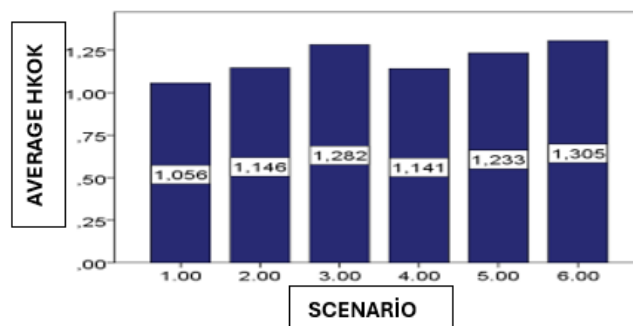


Figure 4. Average HKOK values according to scenarios

As can be seen from **Figure 4**, the best predictive performance on average for 9 data sets was obtained from Scenario 1, where clustering algorithms were applied to the raw data and the score

estimation was made according to Eq. 22. The worst performance was obtained from Scenario 6, which corresponds to the PCA size reduction technique and its use with Eq. 23 for score estimation.

Table 5. Wilcoxon Test significance values according to OMYH criteria

	Scenario1	Scenario2	Scenario3	Scenario4	Scenario5	Scenario6
Scenario1	-	0.657	0.004	0.003	0.003	0.026
Scenario2	-	-	0.003	0.003	0.003	0.003
Scenario3	-	-	-	0.003	0.003	0.003
Scenario4	-	-	-	-	0.003	0.003
Scenario5	-	-	-	-	-	0.534
Scenario6	-	-	-	-	-	-

According to **Table 5**, the scenario binaries whose significance values are greater than 0.05 and therefore there is no significant difference between their performances are Scenario 1-Scenario 2, Scenario 5-Scenario 6. From this point of view, it is possible to say that there is no significant difference between the OMYH values obtained as a result of the use of raw data and the use of data reduced in size with SVD in cases where Eq. 22 is used for score estimation, and between the OMYH criteria obtained as a result of the use of data reduced in size with SVD and reduced in size with PCA in cases where Eq. 23 is used for score estimation. According to the OMYH criterion, the difference between all other scenario pairs is statistically significant. **Figure 5** shows the average OMYH criteria for all scenarios.

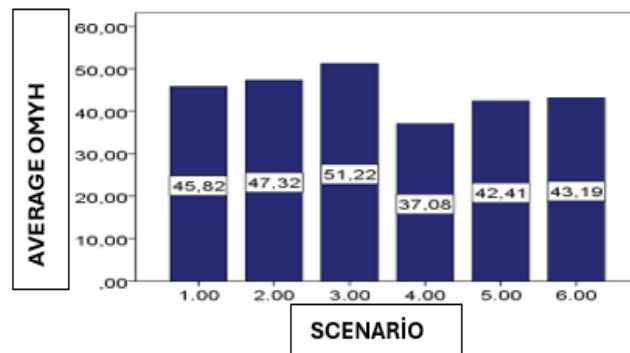


Figure 5. Average OMYH values according to scenarios

Looking at **Figure 5**, the best estimation results according to the OMYH criterion are obtained from Scenario 4. The worst performance was obtained from Scenario 3. Finally, in this section, Wilcoxon test results according to the OMH criterion are given in **Table 6**.

Table 6. Wilcoxon Test intelligibility values according to OMH criteria

	Scenario1	Scenario2	Scenario3	Scenario4	Scenario5	Scenario6
Scenario1	-	0.657	0.656	0.004	0.006	0.013
Scenario2	-	-	0.003	0.003	0.003	0.003
Scenario3	-	-	-	0.003	0.003	0.003
Scenario4	-	-	-	-	0.009	0.010
Scenario5	-	-	-	-	-	0.154
Scenario6	-	-	-	-	-	-

According to **Table 6**, the scenario pairs that did not have a significant difference between their performances according to the OMH criterion were determined as Scenario 1-Scenario 2, Scenario 1-Scenario 3, Scenario 5-Scenario 6. The difference between all other scenario pairs was found to be statistically significant.

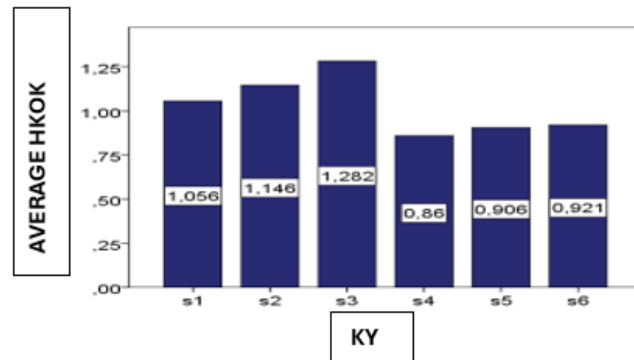


Figure 6. Average OMH values according to scenarios

In **Table 6**, it is seen that the scenario with the highest performance according to the OMH criterion is Scenario 4, the performance of the scenarios using Eq. 10 and Eq. 16 for score estimation is higher, and the scenario with the worst performance is scenario 3.

4. Conclusion

CF is a popular recommendation algorithm that uses the ratings or ratings of users in the system for the prediction of products that the active user might like. The main purpose of this technique is to identify the users who behave similarly to the active user among the users in the system in the most accurate way. Various approaches are used for this purpose. The basic steps of the most popular of these approaches are as follows. In the first step, similarities between the active user and all users in the system are calculated using similarity measures such as Pearson correlation coefficient, cosine similarity, and adjusted cosine similarity. In the second step, the similarities are sorted from largest to smallest, and k predetermined number of users are selected that are most similar to the active user. In the last step, it is tried to estimate the score that the active user can give to that product by using the scores or degrees given by the most similar users to the product to be predicted. Here, if it is predicted that the user will give a high rating to the product, the product is recommended to the user, otherwise it is not recommended. However, such an approach requires a high computational cost if the number of users or the number of products is high. For this reason, clustering analysis techniques are used to reduce the number of users in user-based CF techniques and the number of objects in object-oriented CF techniques. Similarly, size reduction techniques are used to reduce the number of objects in user-based CF techniques and the number of users in object-oriented CF techniques.

In approaches based on clustering, first of all, the data set is divided into c number of clusters using various methods or algorithms. In the next step, it is first determined which cluster the active user is closest to. The next step is to determine the k users that are the most similar from the cluster they are closest to and the score is calculated based on these users. The main goal of this approach is to reduce the number of users for whom similarity will be calculated.

The size reduction process, on the other hand, aims to reduce the number of objects and thus reduce the number of terms in the similarity calculation if we are talking about the user-based CF technique.

So far, various cluster analysis and size reduction techniques have been used for size reduction. However, there has not been a comprehensive study to compare the performance of these methods. The main purpose of this study is to compare the performances of the most popular 11 clustering algorithms and 2 dimensional reduction techniques using 9 datasets with different user and product numbers. For this purpose, 6 different scenarios were carried out.

As a result of the Wilcoxon tests carried out in order to determine whether there is a significant difference between the performances of the scenarios;

- A statistically significant difference was found in all scenario pairs except Scenario 1-Scenario2, Scenario1-Scenario3, Scenario5-Scenario6, Scenario1-Scenario6 according to the HKOC criterion. When the HKOC averages were examined, it was seen that the best performance was obtained from Scenario 1 and the worst performance was obtained from Scenario 6.
- When the scenarios were compared according to the OMYH criterion, the difference between all scenario pairs except Scenario1-Scenario2 and Scenario5-Scenario6 was found to be significant. When the averages were examined, it was seen that the best performance was obtained from Scenario 4 and the worst performance from Scenario 3.
- Finally, when the comparisons were made according to the OMH criterion in this section, the

difference between the performances of all scenario pairs except Scenario1-Scenario2, Scenario1-Scenario3, Scenario5-Scenario6 was found to be significant, similar to the HKOK criterion. When the averages were examined, it was determined that the scenario that provided the best performance was Scenario 4 and the scenario that provided the worst performance was Scenario 3.

The key findings of the study can be summarized as follows.

When the averages of the CF performances of the clustering methods for 9 data sets, 6 different scenarios, and a total of 54 different situations were compared, it was found that the top three most successful clustering methods were WB, TeB, MeB according to the HKOK and OMH criteria, and WB, TeB and AB according to the OMYH criteria. Hierarchical clustering methods consistently outperform other techniques in CF. Among dimensionality reduction techniques, SVD outperformed PCA, particularly in scenarios involving high-dimensional datasets. In the light of all this information, it was concluded that the performance of CF techniques was better than the use of Eq. 21, Scenario 4, SVD size reduction technique, WB and TeB clustering methods for score estimation

Declaration of interest

This study is derived from the thesis and there is no conflict of interest

Acknowledgements

The author would like to thank Nevin Guler Dincer for her contributions

Nomenclature

Abbreviations

RS	Recommendation Systems
CA	Cluster Analysis
CF	Collaborative Filtering
SOM	Self-Organizing Mapping
SVD	Singular Value Decomposition
PCA	Principal Component Analysis
TeB	Single Connection
TB	Complete Connectivity
MeB	Central Connectivity
OB	Average Connection
MB	Median Connection
AB	Weighted Connection
WB	Ward Link
KO	K-Averages
KM	K-Medoids
BCO	Fuzzy C-Averages
HKOK	Square root of the mean of squared error
OMYH	Average Absolute Percentage Error
OMH	Mean Absolute Error

References

- [1] Cai, D., Wang, X., & He, X. (2009, June). Probabilistic dyadic data analysis with local and global consistency. In *Proceedings of the 26th annual international conference on machine learning* (pp. 105–112).
- [2] George T, Merugu S., (2005), A scalable collaborative filtering framework based on co-clustering. In Proc. the 5th IEEE Int. Conf. Data Mining, Nov. pp.625-628.
- [3] Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements Of Statistical Learning: data mining, inference and prediction* (2 ed.). Springer, pp 745.
- [4] Heckerman D., Chickering D., Meek C., Rounthwaite R. and Kadie C., (2001) Dependency networks for inference, collaborative filtering, and data visualization. *The Journal of Machine Learning Research*, 1:49–75.
- [5] MacQueen, J. B., (1967), Some Methods for Classification and Analysis of Multivariate Observations, Proc. Symp. Math. Statist. and Probability (5th), 281–297.
- [6] Şenol, A., Kaya, M. ve Canbay, Y. (2024). Akan veri kümeleme probleminde ağaç veri yapılarının performans karşılaştırması. *Gazi Üniversitesi Mühendislik Mimarlık Fakültesi Dergisi*, 39 (1), 217-232.
- [7] Groth, D., Hartmann, S., Klie, S. ve Selbig, J. (2013). Başlıca Bileşenler analizi. *Hesaplamalı Toksikoloji: Cilt II*, 527-547.

- [8] Bakır, Ç., & Albayrak, S. (2014, April). User based and item based collaborative filtering with temporal dynamics. In *2014 22nd Signal Processing and Communications Applications Conference (Siu)* (pp. 252-255). IEEE.
- [9] Sarwar B., Karypis G., Konstan J. and Riedl J., (2001) Item-based Collaborative Filtering Recommendation Algorithms. In *Proceedings of the 10th International Conference on World Wide Web (WWW '01)*. ACM, 285– 295. DOI:<http://dx.doi.org/10.1145/371920.372071>.
- [10] Şenol, A., Kaya, M. ve Canbay, Y. (2024). Akan veri kümeleme probleminde ağaç veri yapılarının performans karşılaştırması. *Gazi Üniversitesi Mühendislik Mimarlık Fakültesi Dergisi* , 39 (1), 217-232.
- [11] Xu R,Wunsch D., (2005) . Survey Of Clustering Algorithms, *IEEE Transactionson Neural Networks*, 16(3):645–678.
- [12] Altinisik, A., Yildirim, U., & Topcu, Y. I. (2022). Evaluation of failure risks for manual tightening operations in automotive assembly lines. *Assembly Automation*, 42(5), 653-676.
- [13] Koochi, H., Kiani, K. (2016), User based collaborative filtering using fuzzy c-means, *Measurement*, 91:134-139.
- [14] Chen, J., Wang, H., & Yan, Z. (2018). Evolutionary heterogeneous clustering for rating prediction based on user collaborative filtering. *Swarm and Evolutionary Computation*, 38, 35-41.
- [15] Liao, C.L., Lee, S.J. (2016) A clustering based approach to improving the efficiency of collaborative filtering recommendation, *Electronic Commerce Research and Applications*,18:1-9.
- [16] Ba, J. ve Frey, B. (2013). Derin sinir ağlarını eğitmek için uyarlanabilir bırakma. *Sinirsel bilgi işleme sistemlerindeki gelişmeler* , 26 .Chicago
- [17] Hastie,T ,R.Tibshirani and J. Friedman (2009). *The Elements Of Statistical Learning: datamining, inference and prediction* (2 ed.). Springer, pp 745.
- [18] Roelofsen, P. (2018), *Time Series Clustering*, Master Thesis, Vrije Universiteit, Amsterdam, 83s.
- [19] MacQueen, J. B., (1967), *Some Methods for Classification and Analysis of Multivariate Observations*, Proc. Symp. Math. Statist. and Probability (5th), 281– 297.
- [20] Kaufman, L. ve Rousseeuw, PJ (2009). *Verilerde grupları bulma: kümeleme analizine giriş* . John Wiley & Sons.
- [21] Kohonen T. (1995) *Learning Vector Quantization*. In: *Self-Organizing Maps*. Springer Series in Information Sciences, vol 30. Springer, Berlin, Heidelberg pp 175-189.
- [22] Groth, D., Hartmann, S., Klie, S. ve Selbig, J. (2013). Başlıca Bileşenler analizi. *Hesaplamalı Toksikoloji: Cilt II*, 527-547.
- [23] X. Zhang, D. Rajan, and B. Story, “Concrete crack detection using context-aware deep semantic segmentation network,” *Computer-Aided Civil and Infrastructure Engineering*, 34(11) (2019) 951–971; <https://doi.org/10.1111/mice.12477>.
- [24] Konstan, J.A., Riedl, J. (2012) *Recommender systems: from algorithms to user experience* , *Adapt Interact* 22: 101–23 .
- [25] Pan, C., Li. W. (2010) *Research paper recommendation with topic analysis*. In *Computer Design and Applications IEEE* 4, pp V4-264 .
- [26] Konstan J.A., Miller B.N., Maltz D., Herlocker J.L., Gordon L.R., Riedl J., (1997), *Applying collaborative filtering to Usenet news*.*Commun ACM*; 40(3):77-87.
- [27] Link 1 , (<https://www.kaggle.com/datasets>) , (Jester Collaborative Filtering Dataset) , (Restoran_tavsiye_sistemi) , (Recommendation System (CF) | Anime),01.08.2023
- [28] Link 2, <https://github.com/Ramakrishna05/Recommendation-Algorithm>, 01.08.2023.
- [29] Link 3, Web: <https://bookdown.org/egarpor/PM-UC3M/lm-ii-dimred.html>