Research Article

A Kullback–Leibler divergence-based fuzzy *C*-means clustering for enhancing the potential of an movie recommendation system

S. V. Vimala[1] · **K. Vivekanandan**[1]

Abstract
Recommender systems (RS) are systems that filter information and help users to choose products from a large amount of information available online. RS recommend satisfactory and useful products (items) like movies, music, books and jokes to target users that they are interested in. In collaborative filtering (CF) movie recommendation system, timeliness and accuracy are considered as an indispensable entity since it needs to aggregate the emotions, reviews and preference of users in an optimal manner for aiding them to determine suitable movies of their interest. The potential factor of accuracy and timeliness purely depends on the significance of the utilized CF methods with effective nearest neighbors for facilitating recommendation to the users. In this paper, Kullback–Leibler divergence-based fuzzy *C*-means clustering is proposed for enhancing the movie recommendation system. In this proposed KLD–FCM–MRS scheme, KL divergence-based cluster ensemble factor is included in the fuzzy *C*-means clustering methods for enhancing the stability and robustness in the clustering process. The improved sqrt-cosine similarity is also used to find the effective nearest neighbors for an active user. This proposed movie recommendation scheme is compared to the baseline approaches for investigation. The experimental results of the proposed technique confirmed a superior accuracy, recall value, mean absolute error with reduced time on par with the benchmarked schemes used for analysis.

Keywords Recommender system · Kullback–Leibler divergence · Fuzzy *C*-means clustering · Improved sqrt-cosine similarity

1 Introduction

From the recent past, recommendation system is considered as the indispensable entity in a diversified number of e-commerce applications [1]. This recommendation system in e-commerce application either explicitly or implicitly gathers information pertaining to the users' taste of various products and service like movies, tourism, shopping, etc. [2]. In the explicit method of information acquisition, the user history and ratings of the user products and services are collected and in the implicit method of information acquisition, monitored user behavior like utilized services, watched movies and procured products are collected [3, 4]. In this context, the collaborative filtering process is considered as the method used for manipulating or filtering products and services based on the sentiments of a community of people [5]. In specific, the movie recommendation system gathers the movie ratings from a group of populations and recommends specific movies to the target users using similar-minded people with similar tastes and interests based on information collected from the past [6–8]. Further, clustering is considered as the potential unsupervised data mining aid that can be utilized in the activity of partitioning the dataset into similar groups by enforcing user preferences into significant groups. However, the stability and robustness are considered as the essential parameters that are necessary during the process of clustering [9].

✉ S. V. Vimala, vimsvickyvimal@gmail.com; K. Vivekanandan, kvivekanandan@pec.edu | [1]Department of Computer Science and Engineering, Pondicherry Engineering College, Puducherry, India.

SN Applied Sciences (2019) 1:698 | https://doi.org/10.1007/s42452-019-0708-9

Received: 11 January 2019 / Accepted: 4 June 2019 / Published online: 10 June 2019

In this paper, a Kullback–Leibler divergence-based fuzzy *C*-means clustering is proposed for effective clustering with the view to focus on greater accuracy during movie recommendation. In this approach, the user data ratings existing in the movie lens dataset are considered as input in the clustering process. The users are clustered into different groups using the degree of improvement based on the KL divergence-based cluster ensemble factor for maintaining robustness and stability in the clustering process. The improved sqrt-cosine similarity computes the similarity between two users which uses Hellinger distance measure. This Hellinger distance-based similarity is more effective for high-dimensional data. The KLD–FCM along with improved sqrt-cosine similarity gives better recommendation results than the existing systems.

The forthcoming sections of the paper are organized as follows. Section 2 explains the potential review of the most recent automated and movie recommendation system in the literature with the pros and cons. Section 3 describes in detail about the step-by-step process involved in the implementation process of the proposed KLD–FCM–MRS mechanism with the significance in the each phase of deployment. Section 4 highlights the predominance of the proposed KLD–FCM–MRS mechanism evaluated based on recommendation accuracy, recall value, standard deviation (SD), mean absolute error (MAE) and root-mean-square error (RMSE) under different intensities of clusters. Section 5 concludes the paper with major contributions and possible future enhancements that could be derived from the proposed KLD–FCM–MRS mechanism.

2 Related work

In this section, the recent contributions to the literature attributed toward the development of the potential movie recommendation system are presented with the merits and limitations.

Initially, an integrated trust and Bayesian model-based movie recommendation system is contributed for modeling the preference of the user, such that recommendation accuracy is improvised in a predominant manner [10]. The system is developed in such a way to resist the degree of noisy data in order to enhance the focus on the filtering process of past historical data collected from a community of users. This integrated trust and Bayesian model-based movie recommendation system was determined to be more potent than the compared collaborative filtering approaches since they possess the merits of improving efficiency with increased noise tolerant potential. Then, a precomputed clustering approach for effective movie recommendation system was proposed using the merits of machine learning [11]. This machine-learning-based approach was estimated to be potential in the construction of clusters by utilizing the benefits of distance matrix derived from the movie features for improving the effectiveness in movie recommendation. The effectiveness of this machine-learning-based approach was compared with the equally comparable random clustering, affinity propagation hierarchical clustering and density clustering techniques in order to understand its significance in the role of the remarkable movie recommendation system. The accuracy of this machine-learning-based approach was also determined to be superior than the integrated trust and Bayesian model.

Further, an integrated movie recommendation system based on diversity measure and recall value was proposed for improving the recommendation accuracy [12]. In this approach, a user interaction process was used for accepting user request, recommending multiple movies to the user and recording user choice. Then, random and *k*-nearest neighbor algorithms were used for high influential recommendation and finally, the metrics of diversity and recall were utilized for evaluating and concentrating on enhancing the capability of the move recommendation system. The experimental results of this inferred that hybrid techniques are more significant than the non-hybrid movie recommendation system. A statistical evidence-based movie recommendation system was proposed for estimating the positive degree of association that could be determined from the rating of products or services with the propensity in selecting the recommended items [13]. This statistical evidence-based movie recommendation system was capable of enhancing the determination of ratings by exploiting the bias in selection through the incorporation of computational potent variation-facilitated mechanism. This statistical evidence-based movie recommendation system also enhances the trustworthiness in the process of recommendation by utilizing a neighborhood function that achieves collaborative filtering in order to facilitate better representation of the user population taste and preferences. A collaborative scheme based on regression equation was devised for filtering the multiple dimensions of user preferences in order to aid in optimal movie recommendation process [14]. The formulation of the regression equation in this collaborative scheme was confirmed to enhance the recall ratio and recommendation accuracy to a maximum level of 9% and 20% compared to the latest collaborative movie recommendation systems. This collaborative filtering mechanism was proved to handle the issues that emerge as a result of inadequate levels of preference data collected from the data sources.

Furthermore, a movie recommendation scheme based on the session-oriented temporal graph was proposed for resolving the degree of deviation that exists between

long-term and short-term preferences [15]. This recommendation system was also proved to handle the temporal impact of negative and positive ratings of the raters in the dataset used in the investigation. This temporal graph-based recommendation system was also determined to concentrate on the impacts of the time varying rating constraints and the influence of the positive and negative ratings over the future behavior of the user. The experimental results of this temporal graph-based recommendation system was determined to improve the recommendation accuracy, precision, coverage which is derived from the three predominants: netflix, movie lens and movie tweeting's dataset. An integrated collaborative framework for movie recommendation system (ICF–MRS) was propounded using the merits of item *k*-NN algorithm [16]. This ICF–MRS aided in effective classification of movies into rated and unrated movies through the incorporation of correlation parameters derived from the dataset is considered for implementation.

This ICF–MRS was determined to be more accurate than statistical evidence-based movie recommendation system due its degree of investigation enabled through optimal classification rule formulation. This ICF–MRS also included specific limits of classification such that only the potential rules are used during this process of collaborative filter-based categorization of the user ratings. Koohi [17] employed a user-based collaborative filtering using fuzzy *C*-means and its performance is evaluated against different clustering methods such as *k*-means, self-organizing map (SOM). Author also contributed an integrated fuzzy *C*-means and BAT-based movie recommendation scheme (FCM–BAT–MRS) for facilitating effective and collaborative recommendation to the target users [18]. This FCM–BAT–MRS was proposed for resolving the issues of scalability and improving the process of clustering that focuses on enhancing the quality of the recommendation process. In this FCM–BAT–MRS, fuzzy *C*-means played the vital role of clustering users into a diversified number of groups on which the BAT optimization is employed for estimating the initial cluster position such that accurate recommendations are provisioned to the target users. This FCM–BAT–MRS was also determined to improve MAE, precision and recall during its investigation with the movie lens dataset.

In addition, a potential movie recommendation system using *k*-means and cuckoo optimization algorithms (COA–MRS) is proposed for improving the rate of recommendation accuracy during the utilization of movie lens dataset [19]. This integrated *k*-means and cuckoo optimization mechanisms were determined to reduce mean absolute error, root-mean-square error, *t* value and standard deviation. This cuckoo optimization was determined to optimize the number of data ratings from the dataset such that data scalability issue is resolved to the predominant degree. However, the optimal measure in the cuckoo optimization-based movie recommendation system was not potent enough in ensuring maximum precision and recommendation accuracy.

From the review, it is clear that the stability and robustness of the clustering algorithm need to be enhanced for achieving essential accuracy in the process of movie recommendation to the target users. In this paper, fuzzy *C*-means clustering is enhanced using Kullback–Leibler divergence and gives effective clustering with the view to focus on greater accuracy during movie recommendation.

3 Proposed KLD–FCM–MRS-based movie recommendation scheme

In this proposed KLD–FCM–MRS method, three potential steps involved are as follows: (1) KL divergence-based ensemble fuzzy *C*-means user clustering (2) nearest cluster computation (3) determination of top recommended movies based on the neighborhood user's ratings in the environment.

This proposed approach utilizes the benefits of KL divergence-based cluster ensemble factor for robust and stable clustering process such that the accuracy in the process of recommendations is maintained. In addition, the architectural view of the proposed KLD–FCM–MRS is depicted through Fig. 1.

Figure 1 illustrates the complete workflow of the proposed KLD–FCM-based movie recommendation scheme. It is performed into two main phase including offline and online phases. In offline phase, the input is the user rating matrix derived from the movie lens dataset and utilizes ensemble fuzzy *C*-means clustering method to group the users to form different clusters. The basic idea of ensemble fuzzy *C*-means cluster approach is to apply the clustering method several times on the data (instead of just once) and then combine the results into a final single partition. The process of different homogeneous ensemble fuzzy *C*-means clustering method is applied over the extracted user ratings to cluster the users into different groups. In this work, single FCM clustering algorithm called as homogeneous FCM clustering is used to generate many partitions with different random initializations and by executing the FCM several times for each initialization with different fuzzy parameter values. (Here, 1.5, 2 and 2.5 are used.) Further, the incorporation of KL divergence-based cluster ensemble factor is utilized for generating effective user clusters. In online phase, the estimation of the nearest cluster for active user is computed using Euclidean distance. Then, the improved sqrt-cosine similarity method is used to find the nearest neighbors of the active user in

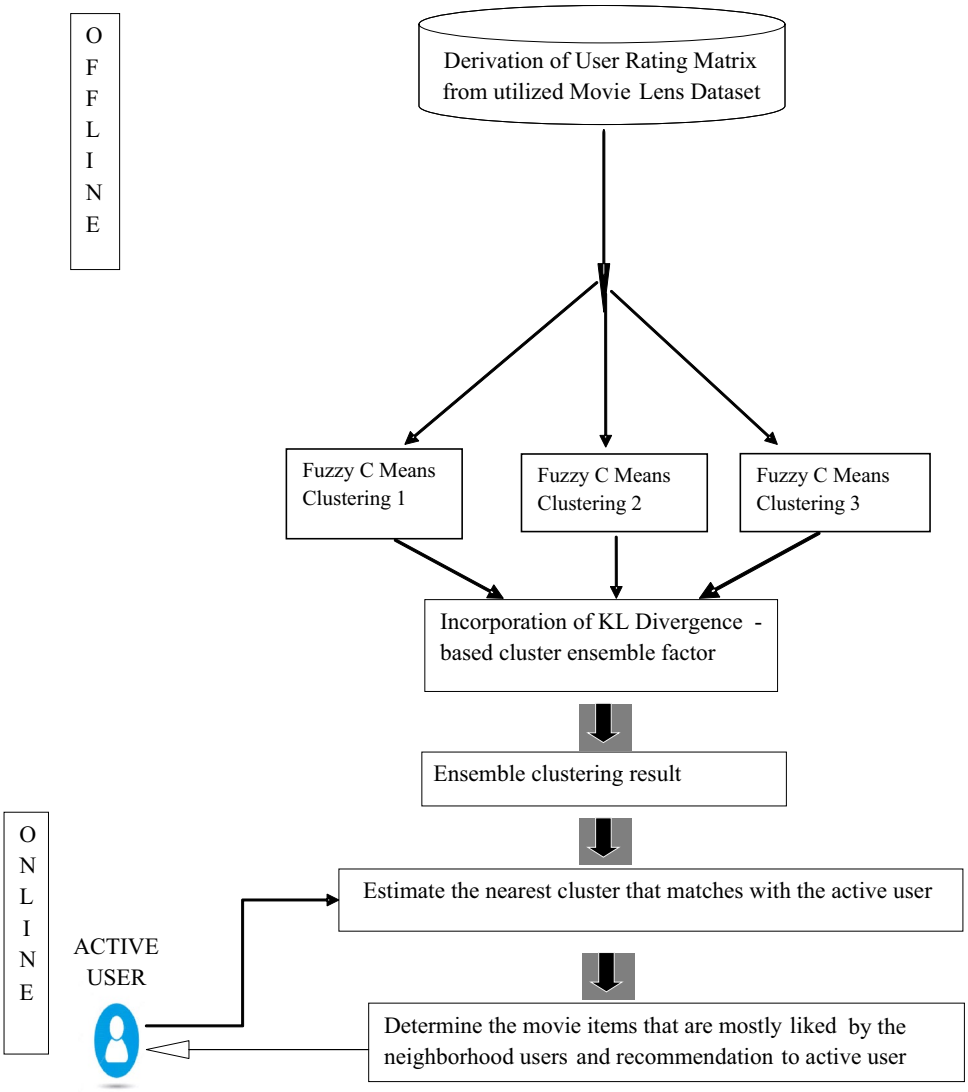

Fig. 1 Architectural view of the proposed KLD–FCM–MRS-based movie recommendation scheme

his/her nearest cluster. Finally, the movie items that are mostly recommended by the neighborhood users of the context are determined in an online mode for identifying the top list of recommended movie items.

3.1 KL divergence-based fuzzy *C*-means user clustering

3.1.1 Fuzzy *C*-means clustering algorithm

In fuzzy *C*-means clustering algorithm, the complete set of data derived from the dataset considered for investigation is partitioned into *k* clusters with the objective of minimizing the distances between the data value and its cluster centers using the concept of weighted aggregation. This objective of minimizing the inter-distance between each data point value $\left(D_{p(j)}\right)$ with its cluster centers $(C_{p(i)})$ is facilitated based on Eq. 1

$$O_{J(\mathrm{FCM})} = \sum_{j=1}^{n}\sum_{i=1}^{k}\left(v_{ij}\right)^{f_c} \times \left\|D_{p(j)} - C_{p(i)}\right\|^2 \tag{1}$$

where $\left\|D_{p(i)} - C_{p(i)}\right\|$ is the Euclidean distance between each data value and its cluster centers with the coefficient of fuzzy (f_c) assigned to 2. $f_c > 1$ is the fuzzy coefficient, which generally takes the value 2. The fuzzy parameter f_c is a weighting exponent in FCM to control the sharing level among fuzzy clusters. The parameter v_{ij} corresponds to the membership value associated with each data value to its *j*th cluster center ($C_{p(i)}$) based on the condition $\sum_{i=1}^{k} v_{ij} = 1$ with $v_{ij} \in [0, 1]$. Further, the value of v_{ij} and $C_{p(i)}$ are periodically updated based on Eqs. 2 and 3, respectively

$$v_{ij} = \frac{1}{\sum_{j=1}^{k}\left(\frac{\left\|D_{p(j)} - C_{p(i)}\right\|}{\left\|D_{p(j)} - C_{p(j)}\right\|}\right)^{\frac{2}{(f_c-1)}}} \tag{2}$$

$$C_{p(i)} = \left(\left(\sum_{j=1}^{n} (v_{ij})^{f_c} \times D_{p(j)} \right) \Big/ \left(\sum_{j=1}^{n} (v_{ij})^{f_c} \right) \right) \quad (3)$$

This process of updating is continued until the criteria of maximum iterations are satisfied for resulting in better clustering process. But, this process of updating the value of v_{ij} and $C_{p(i)}$ in the fuzzy *C*-means are determined to be enhanced predominantly only when the computation of neighborhood membership functions is included in the periodic enhancement process.

3.1.2 Neighborhood membership updating function

The neighborhood membership function is computed for periodic updating of data and cluster center value in fuzzy *C*-means for the following three reasons. The first reason emphasizes that the determination of the neighborhood membership function is capable of handling the issue of imprecision that arise during the process of data clustering. The second reason enforces the need for reducing the number of distance estimation since the computation of Euclidean distances for each and every data point without the context of importance increases the time incurred in the clustering process. Finally, the determination of neighborhood membership function increases the superiority in deriving the feasible data value with maximum fitness objective value during the process of optimization. The membership v_{ij} is updated by the average weighted value of its neighbors' membership values. Thus, the neighborhood membership function updation of $D_{p(j)}$ data point to the *i*th cluster is computed based on Eq. 4

$$\hat{v}_{ij} = \left(\sum_{d \in SN(D_{p(j)}} \frac{1}{|SN|} v_{id} + v_{ij} \right) / 2 \quad (4)$$

where $SN(D_{p(j)})$ represents the neighborhood region related to the data point $D_{p(j)}$. SN denotes the size of the neighborhood (user-specified) and '*d*' is the particular neighborhood. v_{ij} is the membership of original data point $D_{p_{(j)}}$. v_{id} is the neighborhood membership value. In accordance with (4), when most of the neighborhood of a pixel is from the same cluster, then this cluster has a large membership function.

3.1.3 KL divergence-based fuzzy *C*-means user clustering

The process of cluster ensemble determination is divided into two phases, in which the first phase focuses on the derivation of ensemble clustering generator and second phase concentrates on the computation of the consensus function. In the first phase, the data are clustered with different fuzzy clustering methods. In the second phase, the fuzzy clustering results are aggregated by a KL divergence-based objective function. In this proposed KLD–FCM–MRS-based movie recommendation scheme, the method of KL divergence-based cluster ensemble factor is utilized for improving the rate of stability and robustness degree in the clustering process. The KL divergence defines a distance measure between two instances. The KL divergence-based cluster ensemble is used for measuring the distance between the data and cluster center point by considering them as two independent discrete probability distributions. KL divergence is employed to build cluster ensemble through integrating multiple FCM clustering's results.

Initially, the process of homogeneous center-based fuzzy *C*-means clustering algorithms is applied for generating a number of membership matrices for primitive partitioning process in order to build cluster ensemble factor derived based on primitive partitions $\{S_k^T\}_{k=1}^r$. In this context, S_k^T relates to the transpose membership matrix determined as the result of *r*th clustering process. The entries of this transpose membership matrix S_k^T highlight the degree of data pertaining to each cluster considered for investigation. Further, the count of primitive partitions is similar to that of the utilized homogeneous center-oriented soft clustering mechanism. Thus, any kind of clustering mechanism can be optimally selected if that clustering scheme is capable of generating membership matrices with utmost significance. Then, the primitive transpose membership matrices are concatenated into a single matrix $\{S_k^T\}^{CON} = \{S_1^T, S_2^T, \ldots, S_k^T\} \in R^{pxm}$, where *p* and *m* represents the number of data and the number of membership functions derived after the application of different center-oriented soft clustering process. For clarity, $m = kr$ is considered if each primitive partition contains *r* clusters. Moreover, it is not necessary to have a diversified number of clusters in each of the derived primitive partitions. Further, the matrix $\{S_k^T\}^{CON}$ is normalized based on $\{S_k^T\}^{NOR} = \frac{\{S_k^T\}^{CON}}{r}$. Thus, the entries of the matrix $\{S_k^T\}^{NOR}$ can be represented through $\{S_k^T\}^{NOR} = \{\{S_1^T\}^{NOR}, \{S_2^T\}^{NOR}, \ldots, \{S_k^T\}^{NOR}\}$ such that $\left\| \{S_k^T\}^{NOR} \right\| = 1$, where $k = 1, 2, \ldots, p$; Furthermore, each and every row of the normalized matrix $\{S_k^T\}^{NOR}$ is considered as the discrete probability vector over which KL divergence process of estimation is to be initiated. $\{S_k^T\}^{NOR}$ is the input data to the FCE_KL approach.

In the ensemble method proposed, the membership values of '*d*' partitions of various different homogeneous center-oriented soft (fuzzy) clustering schemes are concatenated and normalized to form a new data representation whose weighted aggregation is considered to be a single discrete probability distribution. The KL divergence process of estimation is used in this proposed approach for

better measurement of discrete probability distributions. On this KL divergence process of estimation, desired numbers of clusters are created based on the available number of '*d*' discrete probability distributions. Thus, the KL divergence process-based cluster ensemble factor (distance measure between cluster center of each cluster to their membership vectors) is determined based on Eq. 5

$$\mathrm{FCE}_{\mathrm{KL}} = \sum_{j} V_{\mathrm{p}(j)} \log \frac{V_{\mathrm{p}(j)}}{V_{Q(j)}} \tag{5}$$

where $V_{\mathrm{p}(j)}$ and $V_{Q(j)}$ refers to the two discrete probability distributions considered in investigating KL divergence estimation process-based cluster ensemble factor. Let $f_{kj} = \{S_k^{\mathrm{T}}\}^{\mathrm{NOR}}$ ($k = 1, 2, \ldots, p$; $j = 1, 2, \ldots, m$). It is the normalized matrix which is derived by ensemble homogeneous fuzzy *C*-means clustering. It is the input data to FCE_KL approach (fuzzy *C*-means ensemble KL divergence). The FCE_KL splits them into a fixed number of clusters again. This FCE_KL divergence process-based cluster ensemble factor is responsible for partitioning the '*d*' discrete probability vectors f_{kj} with *m* dimensions. This method of KL divergence-based partitions increases the possibility of converting f_{kj} with *m* dimensions into *r* clusters by modifying Eq. 1 into Eq. 6 with the objective of minimizing the fitness function

$$O_{J(\mathrm{FCM})}(KL) = \sum_{j=1}^{n} \sum_{i=1}^{k} (v_{ij})^{f_c} \times \mathrm{FCE}_{\mathrm{KL}}(f_{kj} || C_{ij}) \tag{6}$$

Subject to the constraints specified in Eqs. 7 and 8

$$\sum_{i=1}^{m} v_{ij} = 1 \quad \text{with} \quad k = 1, 2, \ldots, m \tag{7}$$

and

$$\sum_{j=1}^{k} C_{ij} = 1 \quad \text{with} \quad i = 1, 2, \ldots, k \tag{8}$$

where v_{ij} represents the membership value which is related to the *r*th discrete probability vector in the *k*th cluster $C_{ij} = (C_{i1}, C_{i2}, \ldots, C_{ir})$ and $\mathrm{FCE}_{\mathrm{KL}}$ highlighting $\sum_{j=1}^{r} f_{kj} \times \log \frac{f_{kj}}{C_{ij}}$.

The neighborhood membership function is determined for updating the centroid of the clusters in an iterative manner. The input to this phase is the discrete probability vectors that possess the KL divergence ensemble factor very close to the centroid of the clusters with high membership value. In contrary, the discrete probability vectors that possess the KL divergence ensemble factor very far to the centroid of the clusters are considered to have low membership value in order to focus on the minimization of objective function presented in Eq. 6. In this proposed KLD–FCM-based movie recommendation scheme, the Lagrange formula is used for solving the cluster center and updated membership by modifying Eq. 6 into Eq. 9

$$\begin{aligned}(O_{J(\mathrm{FCM})}(KL)_{\mathrm{UP}}) = &\sum_{j=1}^{n} \sum_{i=1}^{k} (v_{ij})^{f_c} \times \mathrm{FCE}_{\mathrm{KL}}(f_{kj} || C_{ij}) \\ &+ \sum_{j=1}^{n} \beta_j \left(\sum_{i=1}^{k} v_{ij} - 1 \right) + \sum_{i=1}^{k} \gamma_i \left(\sum_{j=1}^{n} C_{ij} - 1 \right).\end{aligned} \tag{9}$$

Under the enforcement of satisfying stop criterion constraints highlighted in Eqs. 10 and 11, respectively,

$$C_{ij} = \frac{\sum_{i=1}^{k} (v_{ij})^{f_c} \times f_{ij}}{\sum_{g=1}^{r} \sum_{i=1}^{k} (v_{ij})^{f_c} \times f_{ig}} \tag{10}$$

and

$$v_{ij} = \frac{1}{\sum_{h=1}^{r} \left(\frac{\mathrm{FCE}_{\mathrm{KL}}(f_{kj} || C_{ij})}{\mathrm{FCE}_{\mathrm{KL}}(f_{kj} || C_{kj})} \right)^{\left(\frac{1}{f_c - 1} \right)}}. \tag{11}$$

This process of updating the cluster center and membership is facilitated by the process of converging objective function derived within the maximum number of iteration number used in the process of implementation. Finally, the neighborhood membership updating function derived in Eq. 4 is included in Eq. 12

$$\hat{v}_{ij} = \left(\sum_{d \in \mathrm{SN}(D_{\mathrm{p}(j)})} \frac{1}{|\mathrm{SN}|} v_{id} + v_{ij} \right) / 2 \tag{12}$$

Finally, the KL divergence ensemble method clusters the users into different effective clusters based on the KL divergence and nearest neighborhood membership function.

3.2 Nearest cluster computation

After clustering the users into different clusters, the nearest cluster for active user is computed using Euclidean distance

$$\mathrm{sim}_i(\mathrm{Cent}_i, U) = \sum_{j=1}^{d} \left(\mathrm{Cent}_{i,j} - U_j \right)^2 \tag{13}$$

Cent_i is the centroid of *i*th cluster, *U* is the active user profile, *d* is the dimension of data (number of attribute), U_j represents the *j*th attribute of the active user profile, $\mathrm{Cent}_{i,j}$ is the *j*th attribute of centroid profile in cluster *i*.

3.3 Determination of top recommended movies based on using neighborhood user rating in the environment

The nearest neighbors (neighborhood) for active user are computed using improved sqrt-cosine similarity (Eq. 14). The similarity between users $u1$ and $u2$ is calculated as follows:

$$\text{sim}(u1, u2) = \frac{\sum_{i=1}^{m} \sqrt{R_{u1,i} R_{u2,i}}}{\sqrt{\left(\sum_{i=1}^{m} R_{u1,i}\right)} \sqrt{\left(\sum_{i=1}^{m} R_{u2,i}\right)}} \tag{14}$$

m set of common items rated by user $u1$ and user $u2$, $R_{u1,i}$ is the rating given to item 'i' by user $u1$, $R_{u2,i}$ is the rating given to item 'i' by user $u2$.

The improved sqrt-cosine similarity uses Hellinger distance to compute the similarity between two vectors. This Hellinger distance-based similarity is more effective for high-dimensional data than cosine similarity. This step is essential for comparing each individual user ratings with the other user ratings existing in the clusters. Finally, the movie items that are mostly recommended by the neighborhood users of the context is estimated for concluding the top list of recommended movie items using Eq. 15 that could be possibly suggested to an active user at any instant of time.

In the process of recommendation, the movies are recommended to target users that most likely used by other neighbor users which are not seen by him/her. The prediction rating of unrated items for active user is calculated based on weighted average of the rating of items in the same cluster neighbor's by using (15) and then make top-N recommendations list to active user. The rating of unrated movie (item) 'i' for an active user 'a' is predicted by $P_a(i)$

$$P_a(i) = \bar{R}_a + \frac{\sum_{N \varepsilon Cx} \text{Sim}(a, N) \times (R_N(i) - \bar{R}_N}{\sum_{N \varepsilon Cx} (|\text{Sim}(a, N)|} \tag{15}$$

where a active user, $\bar{R}_a$ average of active user a, Cx set of nearest neighbors of active user a belonging to one common cluster, N nearest neighbor in Cx set, $\bar{R}_N$ average rating score given by active user's neighbor N, $\text{Sim}(a, N)$ similarity between active user a and neighbor.

4 Experimental results and investigations

4.1 Dataset

The performance of the proposed KLD–FCM–MRS is compared with the baseline recommendation systems by conducting experiments with the help of publicly available movie lens dataset. The movie lens dataset utilized for investigating the superiority of the proposed KLD–FCM–MRS scheme with the compared COA–MRS, FCM–BAT–MRS, FCM–MRS and ICF–MRS consists of 100,000 rating that are potentially rated by 943 users [20]. This movie lens dataset (ML 100 K) is comprised of reviews about nearly 1500–1682 movies that are rated on the scale of 1–5, respectively. The performance of the proposed KLD–FCM–MRS approach is investigated by partitioning the entire movie lens dataset using k-cross-validation method. Fivefold cross-validation was performed for evaluating the results. The original dataset is partitioned equally into five subsets. One is the test set (20%), while the other are used as training set (80%). The process is repeated five times, each time a different set is chosen as the test set and the average results were reported. Further, the proposed method was compared with some non-clustering methods which includes basic CF (BCF) [21], user-based CF (UBCF) [22], SVDM [23] [a variant of single value decomposition (SVD) that uses batch learning with a learning momentum], RSVD [24] (regularized SVD model) algorithms in terms of MAE and RMSE.

All the algorithms are coded in MATLAB R2017a, the experiments have been carried out on a PC with 2.3 GHz Intel Core i7 CPU, and 8 GB RAM, 500 GB Hard Disk and 64-bit Windows 10 Enterprise Operating System.

4.2 Evaluation criteria

The comparative investigation of the proposed KLD–FCM–MRS scheme over the existing methods in order to determine its excellence is achieved based on the evaluations conducted using recommendation accuracy, recall, speed, mean absolute error (MAE), standard deviation and root-mean-square error (RMSE) under varying number of clusters of the movie lens dataset.

4.2.1 Mean absolute error

MAE computes the deviation between the predicted ratings and actual ratings which is defined in Eq. 16.

$$\text{MAE} = \frac{\sum \left| P_{ij - a_{ij}} \right|}{N} \tag{16}$$

where P_{ij} the predicted rating value for user i on item j, N the total number of predicted items, a_{ij} the real rating of user i on item j.

4.2.2 Root-mean-square error

RMSE is the square root of the MAE which gives higher penalty when the deviation is higher which is defined below

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{u,i}(p_{u,i} - r_{u,i})^2} \tag{17}$$

4.2.3 Standard deviation

SD is calculated by squaring of differences of single values and the mean and taking the roots (how far away the individual values are from the mean). The SD of errors is computed by the equation stated below

$$\sqrt{\frac{\left(\sum(E - \bar{E})^2\right)}{N}} \tag{18}$$

$\bar{E}$ mean absolute error $E=\{\in_1,\ldots,\in_N\}=\{p_1 - r_1,\ldots,p_N - r_N\}$ $\{p_1,\ldots,p_N\}$ → predicted ratings $\{r_1,\ldots,r_N\}$ → real ratings.

4.2.4 Recall

Recall is the fraction of related recommended items collected by target user to the total number of items that is actually considered as relevant. Better performance is achieved for larger values of recall. The recall for top-N recommendation is calculated using the formula given in Eq. 19

$$\text{Recall} = \frac{t_\text{p}}{t_\text{p} + f_\text{n}} \tag{19}$$

f_n false negative (an interesting item is not recommended to the user), t_p true positive (an interesting item is recommended to the user).

4.2.5 Accuracy

This is used to estimate how far the algorithm proposed recommends an item aptly. The accuracy rate is based on selecting high-quality items from the set of all items. The accuracy is calculated by the equation stated below

$$\text{Accuracy} = \frac{t_\text{p} + t_\text{n}}{\text{total number of population}} \tag{20}$$

t_n an uninteresting item is not recommended to the user.

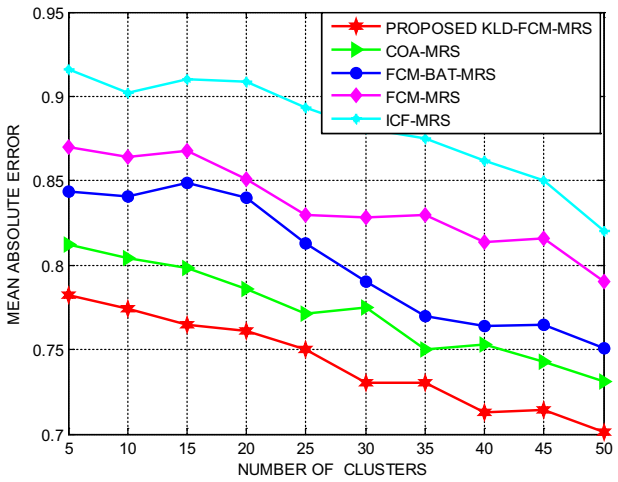

Fig. 2 MAE comparison under different cluster size

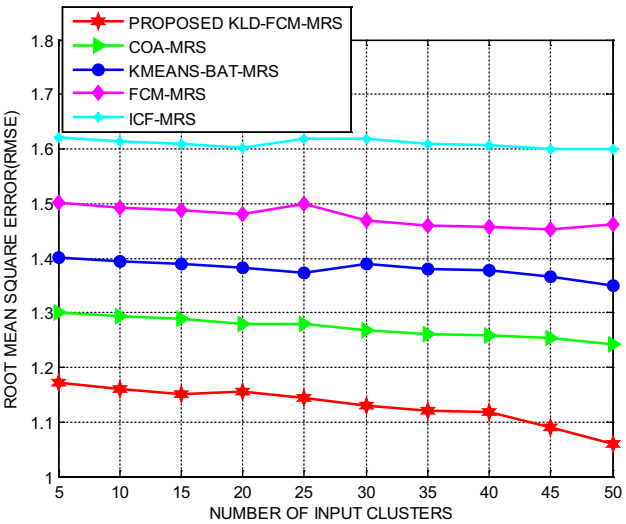

Fig. 3 RMSE comparison under different cluster size

4.3 Result analysis and discussion

Figures 2 and 3 quantify the performance of the proposed KLD–FCM–MRS scheme evaluated using MAE and RMSE under a different number of clusters. The MAE and RMSE of the proposed scheme are proving to get predominantly reduced on par with the existing schemes. Thus, MAE of the proposed scheme is minimized than the existing COA–MRS, FCM–BAT–MRS, FCM–MRS and ICF–MRS approaches. From Fig. 2, it is understood clearly that the MAE is inversely proportional to the number of clusters, i.e., when the numbers of clusters are increased from 5 to 50, it is detected that there is a gradual decrease in the value of MAE. Further, it is added that when there is an

Table 1 Comparison among various methods for MAE and RMSE

Algorithm	MAE	RMSE
BCF	0.762	0.966
SVDM	0.773	0.979
UBCF	0.846	1.083
KLD–FCM–MRS (proposed)	0.7	0.95

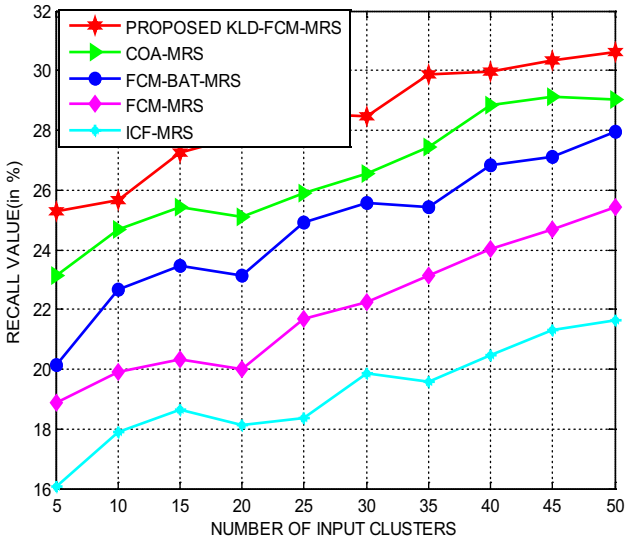

Fig. 4 Recall value comparison under different cluster size

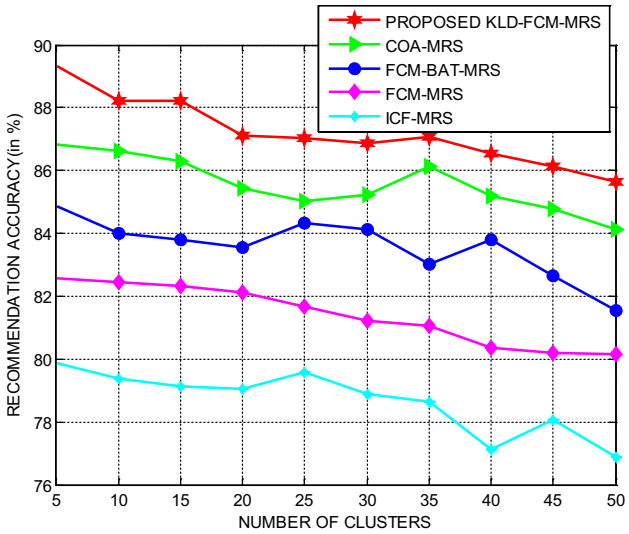

Fig. 5 Comparison of accuracy under different cluster size

Table 2 Comparison of standard deviation of different methods

Method	SD
COA–MRS and	0.1292
FCM–BAT–MRS	0.1558
FCM–MRS	0.1881
ICF–MRS	0.2307
KLD–FCM–MRS	0.0986

Table 3 Comparison of running time of different methods

Methods	ICF–MRS	FCM–MRS	FCM–BAT–MRS	COA–MRS	Proposed KLD–FCM–MRS
Time (s)	35	45	67	64	53

increase in number of cluster, the closeness between the objects going into the cluster also increases. Closer elements remain in the same cluster and prediction becomes extremely accurate. The RMSE of the proposed scheme is also verified to be significantly minimized than the baseline schemes considered for analysis. Figure 3 clearly shows that RMSE is inversely proportional to the number of clusters, i.e., RMSE value decreases gradually as the number of clusters increases. When MAE and RMSE are less, the system proposed predicts accurate user ratings which in turn provide a good recommendation.

Table 1 compares the method proposed with non-clustering methods on movie lens dataset in terms of MAE and RMSE. For ML 100 K, the method proposed outperforms other existing techniques.

Figures 4 and 5 highlight the performance of the proposed KLD–FCM–MRS scheme evaluated using recall and recommendation accuracy under a different number of clusters. The recommendation accuracy and recall value of the proposed KLD–FCM–MRS scheme is determined to be excellent over the baseline approaches since the guidance of KL divergence factor in the process of clustering is responsible for predominant success. The recall is directly proportional to the number of clusters, i.e., the recall increases when the number of clusters increases. This proves that the system proposed provide improved predictions. The performance of the proposed approach is considerably increased than existing methods. Because it utilizes the benefits of KL divergence fuzzy *C*-means clustering with improved sqrt-cosine similarity.

From Table 2, the SD of the proposed scheme is proved to be considerably minimized than the benchmarked schemes considered for investigation. The number of clusters is considered as 50 for every method. Table 3 is the exemplar of the significance of the proposed KLD–FCM–MRS scheme evaluated using execution time of proposed and existing methods which is measured in seconds.

Speed is the execution time that refers to clustering and similarity calculation cost in clustering-based methods. Clustering cost, optimization cost and similarity calculation cost in optimization with clustering-based CF methods. For example, COA–MRS and FCM–BAT–MRS (optimization with clustering-based CF methods) take more time than other techniques. It can be observed that computational cost of proposed KLD–FCM–MRS is mainly focused on the clustering part. Even though clustering techniques cost some time, prediction error is extremely reduced. ICF–MRS and FCM–MRS methods require less time than the other methods, while MAE and RMSE have higher value for these methods. Further, the proposed method KLD–FCM–MRS requires less time and provides lower MAE and RMSE than other existing optimization-based methods. Thus, it is concluded that the proposed method KLD–FCM–MRS provides better value in terms of MAE, RMSE, accuracy, recall and speed.

5 Conclusion

The proposed KLD–FCM–MRS scheme is presented as a reliable contribution that enhances the potential of movie recommendation to a predominant degree by the merits of KL divergence-based fuzzy *C*-means clustering process and improved sqrt-cosine similarity. The proposed scheme emphasized and presented the key role of the KL divergence-based cluster ensemble factor that aids in enhancing the stability and robustness in the clustering process. The improved sqrt-cosine similarity was used for calculating effective similar neighbor users for prediction. The combination of KLD–FCM with improved sqrt-cosine similarity improves the recommendation performance. The experimental investigations of the proposed KLD–FCM–MRS scheme was determined to be superior in recommendation metrics compared to the COA–MRS, FCM–BAT–MRS, FCM–MRS and ICF–MRS approaches and some non-clustering-based methods considered for analysis.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Shrestha J, Jo GS (2009, Apr) Enhanced content-based filtering using diverse collaborative prediction for movie recommendation. In: 2009 1st Asian conference on intelligent information and database systems. IEEE, New York, pp 132–137
2. Halder S, Sarkar AJ, Lee YK (2012, Nov) Movie recommendation system based on movie swarm. In: 2012 2nd international conference on cloud and green computing. IEEE, New York, pp 804–809
3. Zhao D, Xiu J, Yang Z, Liu C (2016, Oct) An improved user-based movie recommendation algorithm. In: 2016 2nd IEEE international conference on computer and communications (ICCC). IEEE, New York, pp 874–877
4. Tan E, Seaman I, Leung H, Ng YK (2016, Nov) Making personalized movie recommendations for children. In: Proceedings of the 18th international conference on information integration and web-based applications and services. ACM, New York, pp 96–105
5. Portugal I, Alencar P, Cowan D (2018) The use of machine learning algorithms in recommender systems: a systematic review. Expert Syst Appl 97:205–227
6. Sun M, Li F, Zhang J (2018) A multi-modality deep network for cold-start recommendation. Big Data Cogn Comput 2(1):7
7. Pirasteh P, Jung JJ, Hwang D (2014, Apr) Item-based collaborative filtering with attribute correlation: a case study on movie recommendation. In: Asian conference on intelligent information and database systems. Springer, Cham, pp 245–252
8. Yuan J, Li L (2014) Recommendation based on trust diffusion model. Sci World J 2014:159594
9. Treerattanapitak K, Jaruskulchai C (2012) Exponential fuzzy *C*-means for collaborative filtering. J Comput Sci Technol 27(3):567–576
10. Wei D, Junliang C (2013) The Bayesian network and trust model based movie recommendation system. In: Intelligence computation and evolutionary computation. Springer, Berlin, pp 797–803
11. Li B, Liao Y, Qin Z (2014) Precomputed clustering for movie recommendation system in real time. J Appl Math 2014:1–9
12. Zhang HR, Min F, He X, Xu YY (2015) A hybrid recommender system based on user-recommender interaction. Math Probl Eng 2015:331–370
13. Vernade C, Cappé O (2015, Oct). Learning from missing data using selection bias in movie recommendation. In: 2015 IEEE international conference on data science and advanced analytics (DSAA). IEEE, New York, pp 1–9
14. You SH, Park J, Choi J (2016) Personal preference based movie recommendation system. Int J Multimed Ubiquitous Eng 11(9):11–18
15. Li WJ, Dong Q, Fu Y (2017) Investigating the temporal effect of user preferences with application in movie recommendation. Mobile Inf Syst 2017:10
16. Sang A, Vishwakarma SK (2017) Design and implementation of collaborative filtering approach for movie recommendation system. Int J Comput Appl 167(12):18–24
17. Koohi H, Kiani K (2016) User based collaborative filtering using fuzzy *C*-means. Measurement 91:134–139
18. Vellaichamy V, Kalimuthu V (2017) Hybrid collaborative movie recommender system using clustering and bat optimization. Int J Intell Eng Syst 10(5):38–47
19. Katarya R, Verma OP (2017) An effective collaborative movie recommender system with cuckoo search. Egypt Inf J 18(2):105–112
20. http://grouplens.org/datasets/movielens/100k/
21. Koren Y (2010) Factor in the neighbors: scalable and accurate collaborative filtering. ACM Trans Knowl Discov Data (TKDD) 4(1):1

22. Wang J, De Vries AP, Reinders MJ (2006, Aug) Unifying user-based and item-based collaborative filtering approaches by similarity fusion. In: Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval. ACM, New York, pp 501–508
23. Ma CC (2008) A guide to singular value decomposition for collaborative filtering. Computer (Long Beach, CA) 2008:1–14
24. Funk S (2006) Netflix update: try this at home. http://sifter.org/~simon/journal/20061211.html