Soft Computing (2019) 23:9253–9263
https://doi.org/10.1007/s00500-019-03796-9

FOCUS

Performance analysis of efficient data distribution in P2P environment using hybrid clustering techniques

S. Raju[1] · M. Chandrasekaran[2]

Published online: 5 February 2019

Abstract
In this paper, *K*-means algorithm has been applied for distributed large data using hybrid clustering techniques. *K*-means is a simple and scalable algorithm which can be applied on large datasets. It is one of the well-known unsupervised clustering algorithms that fail in providing structured to unstructured data to enable extraction of valuable information. Peer-to-peer (P2P) technologies divide the data or resources between the peers for managing the network bandwidth, network participants and processing powers. During the data distribution process in the P2P environments, accuracy, computation complexity and distributed clustering accuracy are the important issues as they reduce the entire system performance. So, the author in this paper considered the system for the distribution of data in P2P environment using mining techniques. The data have been distributed using the hybrid map reducing method which analyzes the large volume of data by performing filtering and sorting. The cluster approach analyzes and manages the neighboring relationship about the peer nodes that helps in the management of the cluster distribution in the dynamic environment. Determination of the efficiency of the cluster formed is done with the help of the hybrid clustering algorithm, and the related system architecture is proposed. The clustering efficiency has been enhanced in the P2P environment using the distributed data network. The efficiency of the formed cluster was evaluated in terms of Jaccard index, *F*-measures, mutual information and rand measure. The performance of the system was analyzed using the experimental results and discussions, namely, error rate, accuracy and time. The multi-objective system helps in easing the difficulties in the implementation of P2P environment sensitive to initial solutions.

Keywords Peer-to-peer · Distributed mining systems · Map reduce · Greedy-based local approximation fuzzy clustering approach · Jaccard index · *F*-measure · Mutual information · Rand measure

1 Introduction

In recent days, peer-to-peer (P2P) is one of the most common technologies for processing the different types of data in the distributed environment. The method analyzes the networking data such as processing power and network bandwidth and divides these data into different parts between the peers (Ku and Zimmermann 2008). The shared data in the peer are used in various applications like content delivery, privacy and anonymity, file-sharing networks, medical data sharing. These applications utilize the large scale of the data processing and data analytic mechanisms for solving problems like scalability, communication, asynchronies, decentralization and fault tolerance considering the factors related to the future. In P2P computing, information is shared via direct exchange and computing resources. The most specific feature of P2P computing is the existence of symmetric communication between the peers with the implication of the ability of the peer to function as both a client and a server.

Communicated by A.K. Sangaiah, H. Pham, M.-Y. Chen, H. Lu, F. Mercaldo.

✉ S. Raju
rasakudil@gmail.com

M. Chandrasekaran
drmcs123@yahoo.com

[1] Mahendra Engineering College, Namakkal, Tamil Nadu, India

[2] Government College of Engineering, Bargur, Tamil Nadu, India

2 Motivation

The data analysis process is performed with the help of the data mining which analyzes the data and clusters similar data for making the efficient distribution (Nghiem et al. 2014). The peer-to-peer-based clustering process includes the characteristic ability to be scalable in the peer-to-peer technology, ability to perform the routerless network and willingness to perform the functions despite any changes in the node or peer. By using these characteristics, the similar data present in the network have been estimated using the neighborhood relationship clustered together. The data mining process computes the data in the dataset in terms of using exact local algorithm and approximate local algorithm. The exact local algorithm is more desirable, but it is difficult to identify the particular similarity data while the approximate algorithm analyzes the data according to the decision taken (Bhandari and Dabhi 2016). So, normally *K*-means clustering algorithm is utilized for efficiently overcoming the issues present in the exact and approximate algorithm (Rostami et al. 2018). The sample-distributed data in the peer-to-peer environment are shown in Fig. 1.

The *K*-means clustering algorithm (Chen and Ho 2006) shares the data by exchanging the message between the peers, thereby reducing the problems seen in the normal clustering process.

3 Methodology

Data clustering is one of the major data mining problems (Kant et al. 2018). One of the most commonly used clustering algorithms is the *K*-means algorithm. The goal of this algorithm is to partition a dataset into separate groups (clusters); each group is represented by its centroid. The portioning is based on the minimization of the sum of squared Euclidean distances between patterns and their corresponding cluster centers. Clustering is an unsupervised technique of data mining. It means grouping similar objects together and separating the dissimilar ones. Each object in the dataset is assigned a class label in the clustering process using a distance measure. The way to initialize the means was not specified. A novel and optimized method like hybrid map reducing concept based greedy local approximation fuzzy clustering approach is used.

4 Problem statement

The way to start the data selection is to randomly choose k of the samples. The results produced depend on the initial values for the means, and it frequently happens that suboptimal partitions are found. The standard solution is to try a number of different starting points. It can happen that the set of samples closest to m_i is empty, so that m_i cannot be updated. This is an annoyance that must be handled in an implementation, but that we shall ignore. The results depend on the metric used to measure $\|x - m_i\|$. A popular solution is to normalize each variable by its standard deviation, though this is not always desirable. The results depend on the value of k. The *K*-means clustering algorithm (Chen and Ho 2006) shares the data by exchanging the message between the peers, thereby reducing the problems seen in the normal clustering process. The algorithm makes adjustment in the network to enable easy adaptation to the changing and dynamic network environment leading to the formation of a better clustering process and minimizing the number of iterations. Despite the provision of an efficient data clustering mechanism, difficulties are still found like completely in computation and distributed clustering (Yang and Yang 2010).

5 Related works

This section discusses various opinions relating to the peer-to-peer data sharing concepts. Yang and Yang (2010) implemented the hybrid peer-to-peer structure which combines the structured peer-to-peer network and unstructured peer-to-peer network. The method analyzes the data in terms using the lookup table. Resilient features reduce the cost and improve the flexibility and efficiency of the shared data. In addition, the system organizes the structured and unstructured data with the help of a particular topology. Then, the efficiency of the system is evaluated using the experimental results and discussions which ensure the highest accuracy while sharing the data in the peer-to-peer network. Bhagat et al. (2016) have implemented the content- and threshold-based information sharing system in the peer-to-peer environment. Initially, the request message was transmitted to the network as files, documents, music and so on. After getting the reply message, the user shares and searches their files in the peer-to-peer environment without any need for centralized network. This system determines the resource utilization of the network based on the content of the file. Then, the efficiency of the system is evaluated with the help of the experimental results and discussions. Thus, the proposed system ensures lower transmission cost and also enhances the success rate when compared to the other traditional methods.

Datta et al. proposed the *K*-means clustering-based data sharing in a peer-to-peer environment. The method analyzes the asynchronous data present in the network. The process involves the analysis in terms of the specific

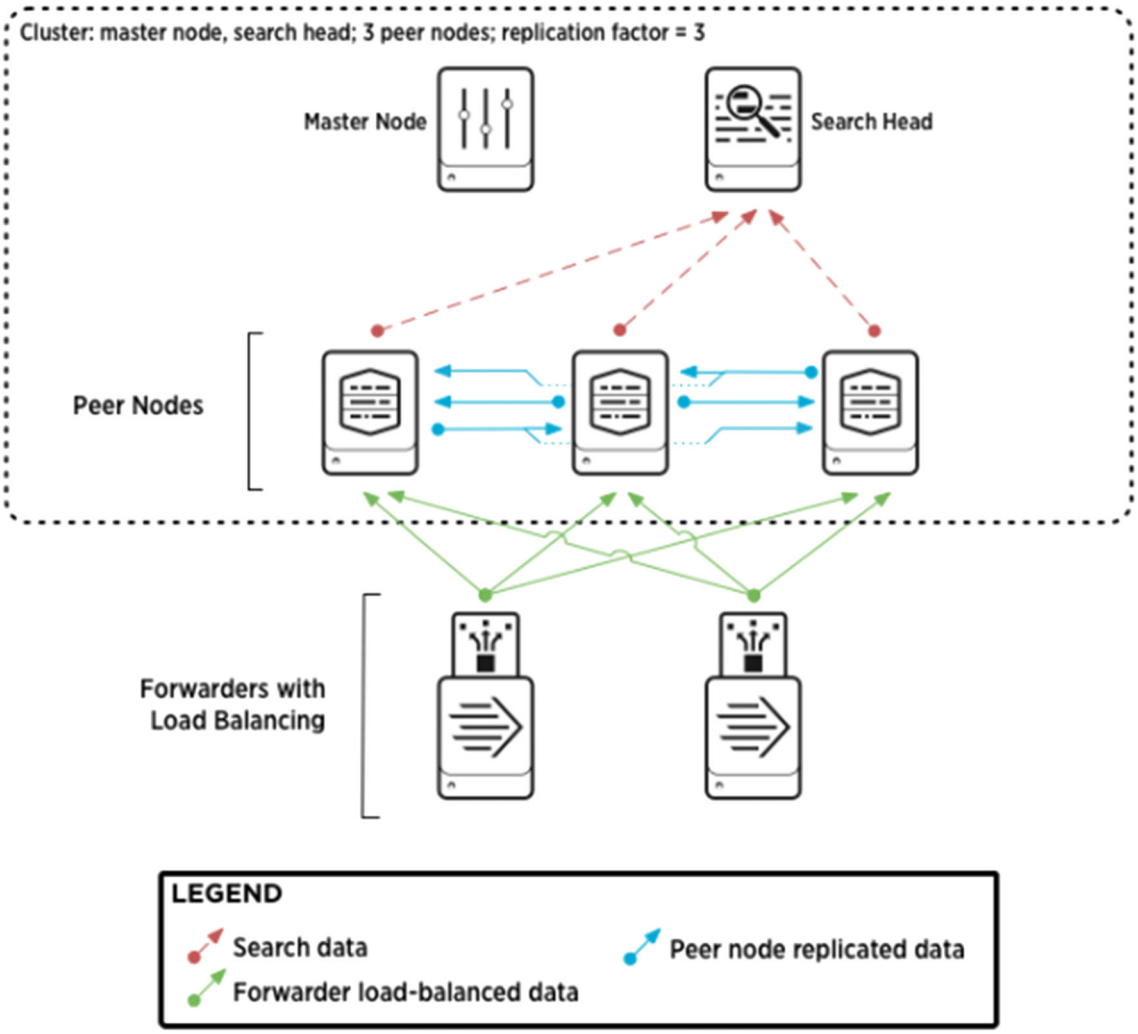

Fig. 1 Data clustering in peer-to-peer environment

topological manner with the estimation of the neighboring node. After estimating the neighboring node, the centroid node is estimated for each iteration. The centroid node is analyzed in the centralized data, thereby reducing the false clustering process. The formation of the cluster helps in efficient sharing of the data in the network. This is followed by the evaluation of the performance of the system using the computation complexity and accuracy of the cluster. Hammouda and Kamel (2014) have resolved the communication cost and privacy issues in the data sharing process using the locally optimized distributed clustering and globally optimized distributed clustering process that aim at providing the solution to the centralized problems during the clustering process. The method utilizes the collaborative clustering approach while analyzing the user request which organizes the user request information in the particular structure. This improves the overall data sharing efficiency when compared to the other methods. The performance of the system is evaluated using the experimental results according to the user recommendation with the specific cluster.

Papapetrou et al. (2010) have succeeded in reducing the cost utilization during the text sharing in the peer-to-peer environment using the probabilistic clustering process, which analyzes the scalability of the data present in the network and the similar texts are clustered together. The similarity concept reduces the overall data loss while sharing the text in the peer-to-peer network within an efficient manner. The method which analyzes the probability value of each text provides guarantees to each document present in the cluster.

6 Objectives

The estimated clusters are distributed in the peer-to-peer environment which entails the minimum cost when compared to the other existing methods. The proposed work

clearly demonstrates the need for the clustering process by data sharing in the peer-to-peer environment. The reduction is done by the network in the data distribution, time and accuracy of the data cluster during data distribution. So, this paper proposes a novel and optimized method like hybrid map reducing concept with the greedy local approximation fuzzy clustering approach for reducing the issues discussed.

7 Proposed system

This section discusses the proposed big data and data mining technique for sharing the data in the peer-to-peer environment. The data have been distributed using the hybrid map reducing concept which analyzes the large volume of data by performing filtering and sorting. In addition, the map reducing technique (He et al. 2014) reduces the parallelizable problems effectively. Then, the selected data are formed together by using the greedy-based local approximation fuzzy clustering approach and the clusters are distributed in the peer-to-peer environment. Determination of the efficiency of the cluster formed is done with the help of the hybrid clustering algorithm and the related system architecture proposed. This is shown in Fig. 2.

7.1 Clustering data in the peer-to-peer environment using the hybrid map reduce and greedy-based local approximation fuzzy clustering approach

The data are distributed using the map reducing concept considering its ability to analyze the large amount of data parallel using the map () and reduce () procedure (Panthong and Srivihok 2015). The map reducing function analyzes the data and forms the efficient cluster which is distributed in the peer-to-peer environment for making the efficient future processing. Initially, map step is applied to each of the data and the output of the data is stored in the temporal memory for reducing redundancy and fault tolerance. The mapping function arranges the entire data in the sorting order and picks the optimized data using the greedy approach (Lee et al. 2015). The selected data are used for reshuffling and redistributing the data in the environment. The greedy method is applied to the data, considering the method analyzes each of the data at least once and computes the importance of each of the data. Greedy algorithm selects the data from the candidate dataset using the selection, feasibility, objective and solution functions. The selection method fetches the data from the dataset which is analyzed for justifying its ability to produce the solution to the specific problem. Then, the objective function is applied to the particular data which are either maximum or minimum value. The chosen data are compared with the threshold value 0.5. If the value is minimum to the threshold value, it is considered as a feasible solution or else it is rejected during the clustering process. The selected data are fed into the clustering process which groups similar data before distributing them in the peer-to-peer environment.

K-means algorithm can be considered as a route for an elegant synchronization technique for clustering distributed data. The fundamental idea with regard to this algorithm is that every node runs a single *K*-means iteration over its local data, and then, the resultant centroids can synchronize the clustering results with the neighboring nodes. Centroids are sent by every node to their neighbors. The centroids of certain cluster are received at all neighboring nodes for certain clusters obtained at neighboring nodes where centroid is modified for each node for that particular cluster to be the weighted average of the received centroid and the current local centroid. Then, next iteration is started at every node with the use of average centroids obtained until the stopping criteria are reached. The behavior of dynamic networks when there is a change in the network structure or data is also included. Higher accuracy is achieved by this algorithm when compared to the classical centralized *K*-

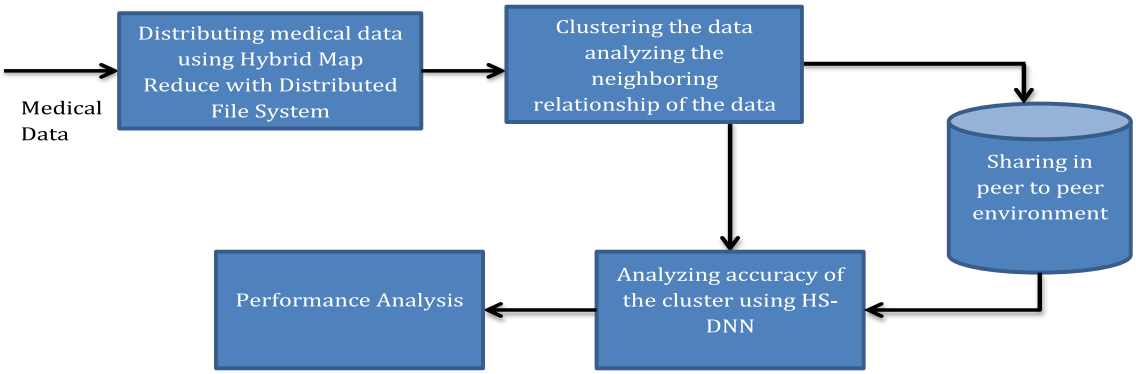

Fig. 2 Proposed system architecture

means and followed by the occurrence of the demonstration of the communication efficiency (Elgohary and Ismail 2011). To classify the database of people with disease, noninvasive-based methods such as machine learning are reliable and efficient. In the proposed study, we developed a machine learning-based diagnosis system for disease prediction by using disease dataset. We used popular machine learning algorithms, feature selection algorithms, the cross-validation method and classifiers performance evaluation metrics such as classification accuracy, sensitivity, coefficient and execution time. The proposed system can easily identify and classify people with disease from healthy people (Samuel et al. 2017).

In Collaborative Distributed Fuzzy *C*-Means Clustering algorithm, the collaborative clustering and its essence explore the structures of every peer using peer exchanges. The Genetic Algorithm (GA) technique is used for special diagnosis,which is automatically evolve optimum connection weights needed to efficiently train a ANFIS model (Asogbon et al. 2016). There are two phases which do the clustering at the individual peer in an interaction among the neighbor peers through exchange of the findings. They also interleave and occur within a fixed sequence. The collaborative fuzzy clustering algorithm approximates the centralized clustering solution and performs distributed clustering at each peer with the collaboration of other peers. The communication links are recognized for the cluster prototype and attribute the weight. The neighboring peers alone communicate the information. The ideal distribution of attribute weights is done on the basis of the attribute–weight–entropy regularization that produces better clustering results. The features that influence the efficiency are mined out for high-dimensional clustering techniques.

The proposed system uses a combination of distance and energy which are critical for selecting the initial cluster point. Hence, accuracy is high and communication cost is low.

7.1.1 Clustering the selected data

Clustering is done with the help of the local approximation-based fuzzy clustering approach which analyzes the density of the complete dataset. During the clustering process, the neighborhood relationship between the data is analyzed. Initially, the *K*-nearest neighboring method (Zheng et al. 2015) is applied to the dataset for estimating the density and structure of the dataset. It is divided into three different clusters, namely, cluster supporting object, cluster outliers rest of the cluster on the basis of the structure and density of the data. After deciding the number of cluster, the cluster centroid is chosen as follows.

7.1.2 Functions of *K*-means

Input: centroids input.//The input is centroid of the cluster and (value of the input) data points.
Output: The output is the nearest cluster of object and value of the object.
1: nxtCentroid ⇓ null, nxtDist ⇓ ∝
2: for each c ε Centroids do
3: dist ⋄ Distance (input. Value, c);
4: if nxtCentroid == null ‖ dist < nxtDist then
5: nxtCentroid ⇓ c, nxtDist ⇓ dist;
6: end if
7: end for
8: output. Collect (nxtCentroid, object)

$$\text{Centroid} = \frac{\sum_x w_k(x)^m x}{\sum_x w_k(x)^m} \tag{1}$$

where x is the set of coefficients in the kth cluster and w_k is the weighted value of the kth cluster of the element x.

The centroid of the each cluster is chosen on the basis of Eq. (1), and the membership value of the each object is defined using the local approximation algorithm and calculated as follows.

$$E(\{P\}) = \sum_{x \in X} \left\| p(x) - \sum_{y \in N(x)} w_{xy} p(y) \right\|^2 \tag{2}$$

where x is the set of objects present in the cluster. $p(x)$ is the fuzzy membership value of the data x. $N(X)$ is the nearest neighbors of the data x. w_{xy} is the coefficients reflecting the nearest neighboring data.

Then, the cluster membership is decided according to the fuzzy membership value. The data which have full membership value belong to the cluster supporting objects. Finally, the clusters are formed in terms of the one-to-one cluster membership and one-to-multiple object-based clustering which reduces the error rate in the cluster formation. If the estimated data membership value is smaller than the threshold value, it is assigned as one-to-one cluster or else the data features belong to the one-to-multiple cluster. The error rate is minimized through a continuous updating process of the weight value of the data. The updating process is done with the help of the following Eq. (3)

$$w_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}} \tag{3}$$

This process is repeated until the better cluster is formed by utilizing the selected data from the user database. After the application of the shuffling step to the data for redistribution, the remaining data are clustered for improving

the cluster formation process to follow. In addition, it reduces the step process in all the data parallel for ensuring efficient data distribution in the peer-to-peer environment. A description of the hybrid map reducing-based local approximation fuzzy clustering approach algorithm is provided as Fig. 3.

Data are clustered and distributed in the peer-to-peer environment using the above algorithm. This is followed by the evaluation of the cluster formed using the optimization search-based distributed network.

Distributed nodes with similar characteristics are clustered with the active nodes formed as cluster with the help of the nearest nodes. Every node in the cluster sends a message to the nearest node to exchange the information about the current status of the node. Global search is performed for avoiding the algorithm getting trapped in local optima. Information relating to the convergence of the solution is propagated here. The neighboring nodes get this information, and the solution is modified on the basis of the information, thus escaping local minima. In this work, (Yang 2014) the multi-objective optimization has been hybridized using a *K*-means clustering and the AFSA through the cooperation of neighbors. This multi-objective system helps in easing the difficulties in the implementation of P2P environment. This work introduces an AFSA multi-objective system that promises the neighbor cooperation in the application of P2P network.

7.2 Estimating the accuracy of cluster using harmonic search

The efficiency of the formed cluster through the use of the harmonic search that analyzes the features present in the cluster and distributed cluster is described. The learning convolution neural networks work on the basis of the concept of the supervised learning method which includes four layers, namely, convolution layer, pooling layer, rectified unit layer and lose layer. Every layer performs a specific function, and the achieved output is compared to the neural networks due to the ability to process the noise data. In the convolution layer (Nallakannu and Thiagarajan 2016), the bulk of input is accepted from the clustering process which is analyzing the different directions in terms of measurements of the three different parameters, namely, depth, side and zero padding. After analyzing these parameters, pooling layer analyzes the maximum pooling value of each feature which is fed into the rectified unit value which calculates each feature value by applying the activation function. The membership functions of fuzzy logic systems which are quantitatively defining the linguistic labels employed by such systems usually take longer time to design and tune to accommodate new situations. Since neural network has the capabilities of self-learning and self-tuning, it can be used to automatically generate membership functions for fuzzy logic systems (Oluwarotimi et al. 2013). So, the activation or learning function determines the speed and accuracy of a cluster with particular features with the minimum error rate, which is estimated after the application of the activation function, by comparing the actual node with the related expected value. The weight and bias value are updated on a continuous basis, when changes occur, using the reactive optimization method and considering that it reduces the system error effectively.

At the time of output estimation process, every layer input is multiplied by its related weight value and bias value that needs addition. The output calculation is done using the following equation:

$$\text{Net output} = \sum_{i=1}^{N} x_i * w_i + b \tag{4}$$

Then, updating the weights and bias is done with the help of the optimization method considering the ease in analyzing the features in every direction and an objective manner. The feature weight and bias are estimated using the harmonic searching algorithm. The functioning of algorithm is analogous to the work of a musician playing harmony. Initially, the random vectors like weighted value are chosen from the network and the probability of each value is calculated and used for updating the weighted and bias value during the training process. The probability value which is greater than the random weighted value is replaced by the new value. Likewise, the bias value is also changed. Continuous update of weights with previous values is done according to the above optimization method. This is followed by the training of the features with the help of the sigmoid and Gaussian function. Then, the network analyzes the inputs present in the work and makes effective classification of the accuracy in clustering the features in a specific group with minimum error rate. The efficiency of the proposed system is evaluated using the following experimental results and discussions.

8 Performance analysis

This section deals with the efficiency of the proposed data distributing method in the peer-to-peer environment. Evaluation of the efficiency of the cluster formed is done using the Jaccard index, F-measure, mutual information and rand measure metrics. The metrics determine how the proposed hybrid map reducing concept with greedy-based local approximation clustering approach forms the cluster in an efficient manner. The high performance of the cluster indicates the successful distribution of the data by the

proposed system in the peer-to-peer environment. Then, the metrics are evaluated with the help of the skin segmentation dataset and adult dataset for determining the clustering accuracy.

8.1 Skin segmentation dataset

Feature of the skin segmentation dataset (https://archive.ics.uci.edu/ml/datasets/Skin+Segmentation) is collected from persons in various age groups and race bunches. The collected data consist of B, G and R values which are gathered from the pictures of the faces. The sexual oriented features were acquired from the FERET and PAL database (Asogbon et al. 2016). The size of the testing data base was 245,057 in which 50,859 were skin tests and remaining 94,198 were considered as the nonskin test set.

8.2 Adult dataset

The next dataset is adult dataset (https://archive.ics.uci.edu/ml/datasets/Adult) which was collected from the UCI machine learning store. The dataset contains 30,162 features collected from the various trainings, sex, work classes, conjugal status and age groups. The features of dataset are age, work class, fn weight, education, education num, occupation, transportation, relationship, house per week, race, capital gain and so on.

Table 1 shows the parameters for skin segmentation dataset used in this work which are specified and described.

8.3 Performance metrics

8.3.1 Jaccard index

Jaccard index (Andreas de Ruiter) is the one of the important performance metrics used in the analysis of the accuracy of the proposed system in the retrieval of the elements common to the two different datasets. In most cases, the Jaccard index value lies between 0 and 1. The retrieved common data were clustered and distributed in the peer-to-peer network. Then, the Jaccard index value is calculated as follows.

$$J(A,B) = \frac{\text{True positive}}{\text{True positive} + \text{False positive} + \text{False negative}} \tag{5}$$

Table 1 Parameters for skin segmentation dataset

Dataset characteristics	Univariate
Attribute characteristics	Real
Associated tasks	Classification
Number of instances	245,057
Number of attributes	4
Missing values	N/A
Area	Computer
Number of web hits	106,563

8.3.2 *F*-measure

F-measure is the metric (Andreas de Ruiter; Gosain and Dahiya 2016) which is used for the determination of the efficiency of the contribution of the present system and also the identification of similar items from the dataset. In addition, the F-measure analyzes the precision and recalls values that indicate the accuracy in the formation of the cluster in the proposed systems. The F-measure is calculated as follows.

$$\text{Precision} = \frac{\text{True positive}}{\text{True positive} + \text{False positive}} \tag{6}$$

$$\text{Recall} = \frac{\text{True positive}}{\text{True positive} + \text{False negative}} \tag{7}$$

By using the precision and recall value, the F-measure is estimated as follows:

$$F_\beta = \frac{(\beta^2 + 1) \cdot P \cdot R}{\beta^2 \cdot P + R} \tag{8}$$

where β represents the weight factor of the particular data in the recall. P and R represents the precision and recall value.

8.3.3 Mutual information

The mutual information metric analyzes the magnitude of the information shared and distributed at the time of clustering which reduces the error rate and increases the cluster accuracy.

8.3.4 Rand measure

Rand measure (Ganeshkumar et al. 2016) analyzed the formation of similar cluster while distributing the data in the network. The rand measure is analyzed as follows:

$$\text{Rand measure} = \frac{\text{True positive} + \text{True negative}}{\text{True positive} + \text{True negative} + \text{False positive} + \text{False negative}} \tag{9}$$

The above metrics are used for the analysis of the accuracy. For improving the accuracy the hybrid map with greedy-based local approximation clustering technique

Algorithm for hybrid distance reducing based local approximation fuzzy clustering

Step 1: Initialize the candidate set

Step 2: Applying the map () function which sorting the data and the greedy algorithm selects the optimize data from the database.

Step 3: The optimized data are selected according to the min and max criteria.

Step 4: The selected data has been set into the clustering process.

Step 5: Number of clusters (3) has been decided before forming the cluster.

Step 6: Calculated the cluster centroid value as follows.

$$\text{Centroid} = \frac{\sum_{x} w_k (x)^m x}{\sum_{x} w_k (x)^m}$$

Step 7: After that the membership value of each data is estimated using the fuzzy clustering which analyze the nearest neighborhood value

$$E(\{P\}) = \sum_{x \in X} \left\| p(x) - \sum_{y \in N(x)} w_{xy} p(y) \right\|^2$$

Step 8: Each membership value is compared with the threshold value if the value is greater than threshold value which is assigned to the one to multiple criteria else one to one Cluster.

Step 9: Then the error rate has been reduced by continuous updating process of the weight

$$w_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}}$$

Step 10: Then apply the shuffling step, which redistribute the remaining data to form the cluster.

Step 11: At last the reduce step is applied to minimize the data also produce the output from whole dataset.

Step 12: Then the formed clusters are distributed in the peer to peer environment successfully.

Fig. 3 Algorithm for clustering process

Table 2 Jaccard index value

Data cluster	Skin segmentation		Adult database	
Methods	Jaccard measure	Rand measure	Jaccard measure	Rand measure
K-means	0.67	0.54	0.73	0.62
MEKPFCM	0.87	0.84	0.85	0.81
Proposed method	0.93	0.91	0.95	0.93

used to form the cluster in the peer-to-peer environment. Table 2 and Fig. 4 explain the rand measure and Jaccard index value.

Figure 4 clearly demonstrates the successful analysis of the similarity data from the two medical datasets and the formation of clusters which are formed with high success rate. The success rate of the proposed cluster approach is further analyzed using the following F-measure value because it decides the efficiency of the cluster in terms of the precision and recall value. The resultant value of the precision and recall is shown in Table 3.

Figure 5 depicts the proposed system ensuring high precision and recall values which indicate the formed clusters as accurate and the clusters containing similar information which enhance the success in the further searching or other process.

Further data of the clusters provide more accurate information which is used in the research process. In addition, the accuracy of the cluster is evaluated with the help of the error rate as the minimum error rate indicates the high accuracy of the system. The error rate of the proposed system is evaluated using Table 4 and Fig. 6.

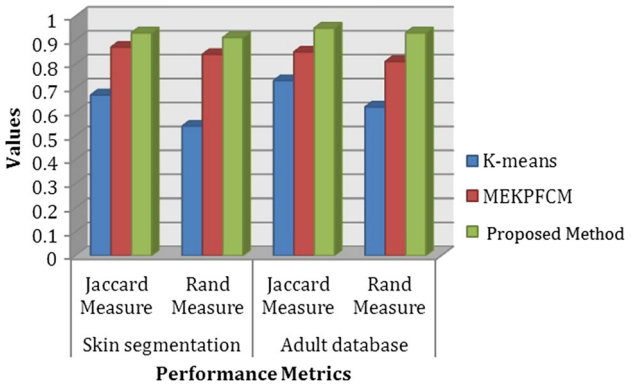

Fig. 4 Jaccard and rand measure

Table 3 Precision and recall

Data cluster	Skin segmentation		Adult database	
Method	Precision	Recall	Precision	Recall
K-means	0.75	0.61	0.77	0.65
MEKPFCM	0.87	0.82	0.89	0.84
Proposed method	0.94	0.92	0.96	0.91

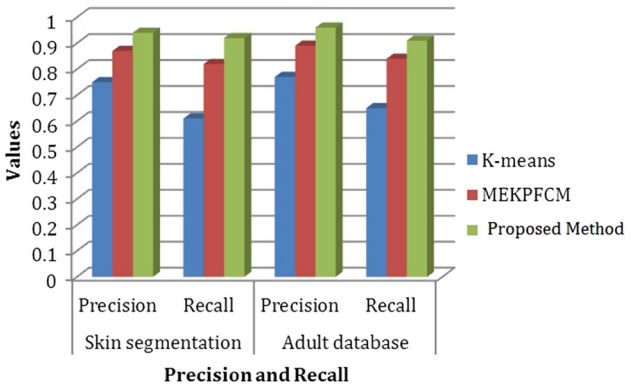

Fig. 5 Precision and recall

Table 4 Error rate

Method	Skin segmentation	Adult dataset
K-means	0.45	0.42
MEKPFCM	0.21	0.19
Proposed method	0.07	0.04

Figure 6 clearly shows the proposed system having a minimum error rate. The error rate analyzed from the features indicates the distributed cluster data seen in both the skin segmentation and the adult datasets. The minimum error rate leads to increase in the accuracy which is shown in Table 5 and Fig. 7.

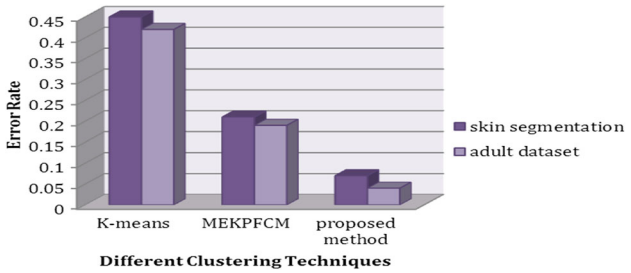

Fig. 6 Error rate

Table 5 Accuracy rate

Method	Skin segmentation	Adult dataset
K-means	79	81
MEKPFCM	89	92
Proposed method	96	97

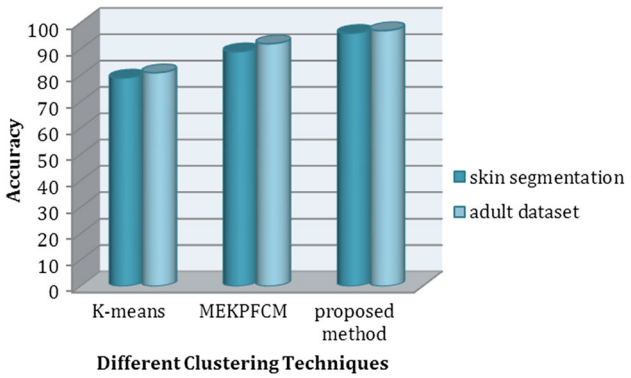

Fig. 7 Accuracy

Table 6 Time

Methods	Skin segmentation	Adult dataset
K-means	3.8	2.9
MEKPFCM	2.4	1.7
Proposed method	1.4	1.0

Figure 7 depicts the proposed system providing a high clustering accuracy which represents the distributed data analyzed in efficient manner when compared to the other existing method. The accuracy indicates that the proposed system solves the problems like redundancy and sparsity with efficient manner. Distributed data require suitable management to depute the proposed system for analysis of the data in the peer-to-peer environment with accuracy.

Then, the time consumption of the proposed system is evaluated using Table 6 and Fig. 8.

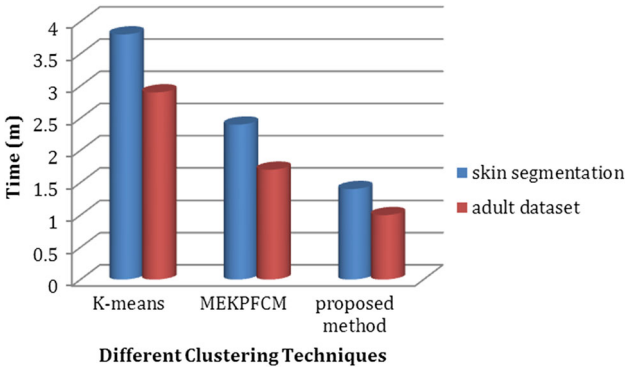

Fig. 8 Time

Thus, the proposed system distributes the data in the peer-to-peer environment in minimum time when compared to the other methods. The method also analyzes data accuracy which reduces the drawbacks in the entire distributing system.

9 Conclusion

In this paper, an efficient algorithm for clustering distributed databases is proposed. The algorithm employs iterative optimization to formulate an efficient algorithm for clustering distributed databases in the form of a peer-to-peer network. Experimental results reported in this paper show the superiority of the proposed methodology over a recently proposed algorithm based on a distributed version of the well-known *K*-means algorithm. Initially, the mapping concept is applied to the dataset and the optimized data have been selected using the greedy algorithm, which processes all the data at least once. The selected data have been clustered with the help of the local approximation optimized base fuzzy clustering. Then, the clusters are distributed in the peer-to-peer environment. It is envisaged that the new proposed algorithm will find extensive applications of distributed clustering where efficient solutions are required. The efficiency of the proposed system is evaluated in terms of F-measure, Jaccard measure, precision, recall, accuracy and time.

The focus of future work will be on evaluation of the proposed algorithm also applied in the other fields in the industry for the prediction of results, optimization resources and the recognition of patterns.

Compliance with ethical standards

Conflict of interest The authors declare that there has no conflict of interest.

References

Andreas de Ruiter, Performance measures in Azure ML: Accuracy, Precision, Recall and F1 Score

Asogbon MG et al (2016) Enhanced neuro-fuzzy system based on genetic algorithm for medical diagnosis. J Med Diagn Meth 5(205):2

Bhagat A, Chaudhari R, Dongre K (2016) Content-based file sharing in peer-to-peer networks using threshold. Procedia Comput Sci 79:53–60

Bhandari RD, Dabhi DP (2016) Extensive survey on *K*-means clustering using MapReduce in datamining. In: International conference on electronics and communication systems (ICECS), May 2016

Chen AP, Ho TH (2006) Location aided mobile peer-to-peer system. J Parallel Distrib Comput 2(3):300–312

Datta S et al. *K*-means clustering over peer-to-peer networks

Elgohary A, Ismail MA (2011) Efficient data clustering over peer-to-peer networks. In: 2011 11th international IEEE conference on intelligent systems design and applications (ISDA), pp 208–212

Ganeshkumar P, Anandaraj M, Arun SR (2016) An effective framework for improving performance of P2P network using efficient group formation mechanism. In: 2016 IEEE international conference on advanced communication control and computing technologies (ICACCCT), pp 758–765

Gosain A, Dahiya S (2016) Performance analysis of various fuzzy clustering algorithms: a review. Procedia Comput Sci 79:100–111

Hammouda KM, Kamel MS (2014) Models of distributed data clustering in peer-to-peer environments. Knowl Inf Syst 38(2):303–329

He Y, Tan H, Luo W, Feng S, Fan J (2014) MR-DBSCAN: a scalable MapReduce-based DBSCAN algorithm for heavily skewed data. Front Comput Sci 8(1):83–99

Kant S, Mahara T, Jain VK, Jain DK, Sangaiah AK (2018) LeaderRank based *K*-means clustering initialization method for collaborative filtering. Comput Electr Eng 69:598–609

Ku WS, Zimmermann R (2008) Nearest neighbor queries with peer-to-peer data sharing in mobile environments. Pervasive Mob Comput 4(5):775–788

Lee EY, Cho HJ, Chung TS, Ryu KY (2015) Moving range *k* nearest neighbor queries with quality guarantee over uncertain moving objects. Inf Sci 325(20)

Nallakannu SM, Thiagarajan R (2016) PSO-based optimal peer selection approach for highly secure and trusted P2P system. Secur Commun Netw 9(13):2186–2199

Nghiem TP, Maulana K, Nguyen K, Green D, Waluyo AB, Taniar D (2014) Peer-to-peer bichromatic reverse nearest neighbours in mobile ad-hoc networks. J Parallel Distrib Comput 74(11):3128–3140

Oluwarotimi WS et al (2013) A web based decision support system driven by fuzzy logic for the diagnosis of typhoid fever. Expert Syst Appl (ESWA) 40(10):4164–4171

Panthong R, Srivihok A (2015) Wrapper feature subset selection for dimension reduction based on ensemble learning algorithm. Procedia Comput Sci 72:162–169

Papapetrou O, Siberski W, Fuhr N (2010) Text clustering for peer-to-peer networks with probabilistic guarantees. In: European conference on information retrieval, pp 293–305

Rostami AS, Badkoobe M, Mohanna F, Hosseinabadi AAR, Sangaiah AK (2018) Survey on clustering in heterogeneous and homogeneous wireless sensor networks. J Supercomput 74(1):277–323

Samuel OW et al (2017) An integrated decision support system based on ANN and Fuzzy_AHP for heart failure risk prediction. Expert Syst Appl 68:163–172

Yang S (2014) Nature-inspired optimization algorithms. Elsevier, Amsterdam

Yang M, Yang Y (2010) An efficient hybrid peer-to-peer system for distributed data sharing. IEEE Trans Comput 59(9):1158–1171

Zheng L, Diao R, Shen Q (2015) Self-adjusting harmony search-based feature selection. Soft Comput 19(6):1567–1579