IET Image Processing

Research Article

MapReduce framework based big data clustering using fractional integrated sparse fuzzy C means algorithm

ISSN 1751-9659
Received on 16th July 2019
Revised 16th December 2019
Accepted on 17th January 2020
E-First on 28th July 2020
doi: 10.1049/iet-ipr.2019.0899
www.ietdl.org

Omkaresh Kulkarni[1] ✉, *Sudarson Jena*[2], *V. Ravi Sankar*[3]

[1]*Research Scholar, GITAM University, Hyderabad, India*
[2]*SUIIT, Sambalpur University, India*
[3]*GITAM University, Hyderabad, India*
✉ *E-mail: omkaresh.kulkarni@mitwpu.edu.in*

Abstract: Big data analytics gain significant interest over the traditional data-processing methodologies that engage in extracting the hidden patterns and correlations from the massive data, termed as big data. With the aim of relieving the computational complexity the clustering method plays a significant role. With the knowledge of the clustering algorithms, the big data arriving from the distributed sources is processed using the MapReduce framework (MRF). The MRF possesses two functions, namely, map function and reduce function, such that the map function is based on the proposed Fractional Sparse Fuzzy C-Means (FrSparse FCM) algorithm and reduce function is based on particle swarm optimisation-based whale optimisation algorithm (P-Whale). Initially, the optimal centroids are computed using the proposed algorithm in the mapper phase that is optimally tuned in the reducer phase, and it is clear that the proposed FrSparse FCM-based MRF ensures the parallel processing of the big data. Experimentation is performed using the Skin data set and the localisation data set taken from the UCI machine learning repository, and the analysis is progressed using the metrics, such as accuracy and DB Index. The analysis proves that the proposed method acquired a maximum accuracy of 90.6012% and a minimum DB Index of 5.33.

1 Introduction

The advent of technology has caused the growth of big data imposing the companies to alter the strategy of gathering, storing, and analysing the data to extract the essential data for performance perfection [1]. The growth in technologies associated with personal computing and social websites, like Facebook and Twitter, resulted in the development of big data [2]. The data is heterogeneous such that the individual objects in big data are multi-modal [3]. The big data comprises of a number of interrelated objects, like texts, audios, and images, which result in high heterogeneity, mainly, in the structural form either as structured or unstructured data. On the other hand, various objects consist of information, even when they are interrelated [4, 5]. The organisations are on track to solve the issues associated with big data that accumulate huge space and takes large time to process the analysis using big data. On the other hand, the data processing using big data is a tedious process as big data is defined as 3 Vs, such as Volume, Velocity, and Variety [1, 6]. Data from different platforms represented as massive data carries the highly significant information, and this information is unable to be analysed because of the above computational issues such that the extraction of information seems to be a huge concern to the society [3]. The potential issues in big data processing and analysis are addressed by the researchers to tackle the diversity, volatility, efficiency, high value, and magnanimity of processing and analysis [4].

The aim of big data analysis relies on the extraction of the in-depth information from the massive data such that the knowledge extraction becomes advantageous [5], for which the data mining algorithm plays a significant role in dealing with big data [7]. In technical terms, data mining is the process of extracting the hidden and useful information from the massive, noisy, incomplete, fuzzy, disordered, and random data. Cluster analysis [8] is the functional area in data mining technology [9]. Clustering is the process in which the data points are grouped as groups or clusters based on the similarity of the data points. The data points in the cluster possess greater similarity among each other, whereas the data points between the clusters often exhibit less similarity [10]. The similarities and dissimilarities between the data points are established based on the attributes and more often, distance measure is employed for analysing the similarities and the dissimilarities. There are several algorithms for clustering that establishes different solutions for the same dataset. Clustering analysis is capable of establishing the hidden relationships available in the data [11]. Thus, the algorithms for clustering are concentrated to solve the big data problem [9]. Upon regressive developments in the literature, the clustering algorithms are classified as density-based clustering, hierarchical clustering, model-based clustering, grid-based clustering, and partitional clustering [12, 13].

Clustering algorithms resemble their style and methods in clustering the data, and one of the partitional clustering algorithms is the K-means that achieves greater clustering performance. It is well known that clustering is a Non-deterministic Polynomial-time (NP)-hard problem even when there are two clusters [14]. Since clustering is a hard problem, the metaheuristic approaches ensure an effective platform for addressing the big data issue [15]. Several methods, such as decision tree [16], support vector regression [17], optimisation algorithms [18], and so on have been introduced for big data clustering. There are various clustering algorithms based on meta-heuristics, such as tabu search [19], Genetic Algorithm (GA) [20], Ant Colony Optimisation (ACO) [21], Particle Swarm Optimisation (PSO) [22], and bee colony [23] such that these meta-heuristics employ clustering algorithm that is computationally efficient in order to determine the better results. K-means algorithm is a common approach that is employed with meta-heuristic approaches [24] for the effective clustering. K-means algorithm is simple and efficient, but it is affected with serious issues with sensitivity such that the optimisation is based on the initial position of cluster centers, presence of empty clusters, and convergence in local optima. Due to the aforementioned issues, k-means avoids converging to the global optimum. On the other hand, nature inspired algorithms, like PSO [25], ant colony-based method for unsupervised learning [26], human interactions [27], water cycle chaotic behaviour [28], and human body systems [29, 30] also contributed much in clustering. The size of the data and the disadvantages of a single machine are tackled using a natural solution that works parallel in a distributed computational environment. A programming framework, termed as MapReduce [31], is capable of dealing with large-scale datasets through the exploitation of parallelism among clusters. MapReduce is advantageous because of its simplicity, fault tolerance, flexibility, and scalability, and hence, it is employed for big data clustering [32–34]. Google Labs make use of MapReduce [35], for processing the massive data [9, 36].

Big data, an area of research, requires an effective method that relieves the complexity associated with data analysis. The proposed method of dealing with big data is handled using the MapReduce model that enables the parallel processing of the big data. The aim of the proposed model is to perform the data clustering using the proposed algorithm. Initially, the big data is spilt as sets of data and admitted to the mappers in the map-reduce framework. In the mapper phase, the optimal centroids are determined using the proposed algorithm Fractional Sparse Fuzzy C-Means (FrSparse FCM) that is the integration of the fractional theory in the Sparse FCM. The main aim of the proposed algorithm is to tune the optimal centroids effectively such that the standard centroid update equation in sparse FCM is modified with the fractional theory. At the same time, the reducers in the MapReduce framework (MRF) acts at performing the data clustering using the optimal centroids with the particle swarm optimisation-based whale optimisation algorithm (P-Whale) algorithm, which is based on Whale Optimisation Algorithm (WOA) and PSO. Thus, the data clustering using the proposed framework ensures an effective analysis of the data.

The main contributions of the work are:

Proposed FrSparse FCM algorithm: The developed FrSparse FCM algorithm is the integration of the fractional theory in the sparse FCM algorithm that aims at determining the optimal centroids to perform optimal clustering.

Proposed FrSparse FCM-based MRF: The aim of the proposed FrSparse FCM-based MRF is to perform the effective optimal clustering. The mapper phase in the MRF uses the proposed FrSparse FCM, for locating the optimal centroid and the reducer phase uses the P-Whale, for effective data clustering. The P-Whale algorithm is the integration of PSO and WOA for enhancing convergence and performance.

The organisation of the paper is: Section 1 outlines the idea behind the big data and insists on the need for data clustering through picturing various clustering algorithms in the literature. The traditional data clustering algorithms that aimed at performing the data clustering are listed in the literature section along with the merits and demerits of the methods portraying the challenges for the research. Section 3 formulates the structure of data clustering using the MRF, and Section 4 deliberates the results of the proposed framework. Finally, Section 5 summarises the research highlighting the achievements of the work.

2 Motivation

The significance of this section is to detail the existing works and present the challenges for the research.

2.1 Literature review

M. Sassi Hidri *et al.* [1] developed a method based on a divide-and-conquer strategy for analysing the massive data that used parallel algorithms and performed distributed clustering. The data is partitioned as portions and was processed at different machines of the MRF. The memory size is reduced so as to minimise the slowdowns, and the cluster selection was performed locally, missing out the global clustering.

Q. Zhang *et al.* [5] developed a method, termed as Privacy-preserving High-order Possibilistic c-Means Algorithm, to deal with the heterogeneity of the big data. The method effectively dealt with the heterogeneous data with good scalability but fails in terms of accuracy.

J. Wu, Z *et al.* [37] designed a model, termed as Fuzzy Consensus Clustering (FCC), for performing the big data clustering based on fuzzy that offered better capability to deal with the incomplete data with better clarity, flexibility, and robustness. However, the method was not automatic in determining the number of clusters.

Q. Zhang *et al.* [38] performed the big data clustering using the secure weighted possibilistic c-means algorithm that operated based on the Brakerski-Gentry-Vaikuntanathan (BGV) encryption scheme. The method was capable of offering better scalability but failed to attain better classification accuracy.

L. H. Son and N. D. Tien [39] proposed a hybrid clustering algorithm using the incremental clustering and FCM that offered a better clustering accuracy with less computational time. The method was not applicable to greater dimensions.

M. H. Hajeer and D. Dasgupta [40] developed a data-aware module and the distributed encoding method using the GA. The method was capable of managing the data distribution and capable of dealing with a wide range of data types. The method permitted the parallel processing of the queries, but the computational overhead was high.

S. S. Ilango *et al.* [41] modelled a method to minimise the execution time of the traditional method, and the modelled method employed the distributed environment to perform the clustering with better time efficiency and accuracy. The performance of the method was tested with multi-node Hadoop cluster.

P. A. Traganitis *et al.* [42] developed a method, termed as sketch and-validate (SkeVa) based on the K-means clustering, and the method was computationally effective and minimised the complexity. The method was incapable of clustering the nonlinearly separable data.

Nikolaos Tsapanos *et al.* [43] described a plethora of theoretical and technical works concerning the application of the Kernel k-Means clustering algorithm to datasets of big data. They have analysed the merits and demerits of every framework and provided an informative suggestion on the steps one should follow for determining the best framework for big data clustering.

Puja Shrivastava *et al.* [44] introduced an Augmentation of K-Means (AKM) clustering algorithm for big data clustering. The AKM was the augmentation of the genetic K-Means clustering algorithm, which provides optimised clusters from the data set in less computing time.

Qingchen Zhang *et al.* [45] developed a secure weighted possibilistic c-means algorithm based on the BGV encryption scheme for big data clustering on the cloud. This method offered a high performance than the conventional weighted possiblistic c-means algorithm, and it obtained good scalability on the cloud for big data clustering.

2.2 Challenges

The challenges of the work are depicted as follows:

- The main challenge of the big data clustering is that the method is big, heterogeneous, and dynamic as they are gathered from various sources with no standard structure [40].
- As big data has huge volumes of data, there are a large number of data objects, and it requires a huge time to process. Most of the traditional methods require huge time to process and suffers from computational complexity [38].

This work considers the drawbacks of the existing big data clustering techniques as the motivation and developed a big data clustering technique using the FrSparse FCM-based MRF. The proposed method tries to solve the challenges in big data clustering.

3 Proposed method of big data clustering using the FrSparse FCM-based map-reduce framework

The ultimate aim of the research is to concentrate on the big data clustering as the traditional methods of data clustering seems to be unsuitable for clustering big data due to the time complexity and inability to extract the required patterns from big data. The proposed method employed the map-reduce framework to deal with the big data, and the optimal clustering is initiated using the optimisation algorithms for forming the optimal clusters. The

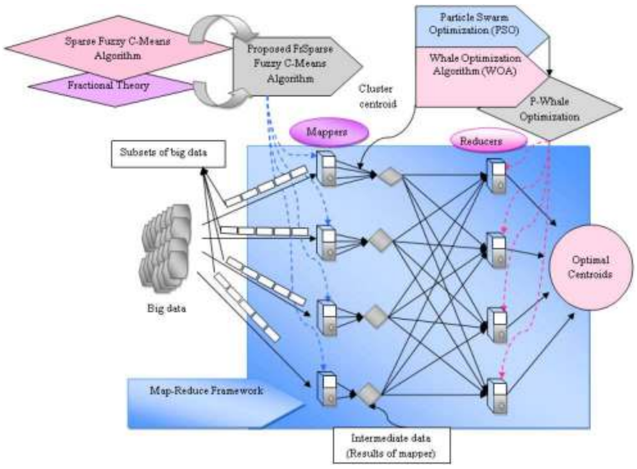

Sparse Fuzzy C-Means Algorithm
Fractional Theory
Proposed FrSparse Fuzzy C-Means Algorithm
Cluster centroid
Particle Swarm Optimization (PSO)
Whale Optimization Algorithm (WOA)
P-Whale Optimization
Mappers
Reducers
Subsets of big data
Big data
Optimal Centroids
Map-Reduce Framework
Intermediate data (Results of mapper)

Fig. 1 *Block diagram of big data clustering technique using the proposed FrSparse FCM based MRF*

modules in MRF operate on the optimisation algorithm, each of which assists the optimal clustering. The mapper phase is inbuilt with a mapper function that follows the proposed FrSparse FCM, which works on the concept of fuzzy and fractional theory, such that the optimal centroids are obtained. Using the generated centroids, the optimal clusters are formulated in the reducer phase that generates the optimal clusters based on the P-Whale algorithm. The proposed method of big data clustering minimises the time spend over the big data in analysing and extracting the interesting patterns of the data. Fig. 1 shows the block diagram of the proposed method of big data clustering.

3.1 MapReduce framework for data clustering using the optimisation algorithm

MRF [46] is a programming paradigm platform for the processing of massive data over several thousands of servers using a Hadoop cluster. MRF is simple and easy to understand, and it defines the two distinct functions of the Hadoop platform. The first function is the map function that transforms the set of input to a data and forms key pairs out of it. The reduce function in the reducer phase takes the input from the mapper and reduces the data to generate the required output. Thus, MapReduce plays a major role in data clustering. The MRF offers better scalability, the ability to store and distribute the big data over a large number of servers available in the platform operating in parallel. The number of servers increases the processing power of MRF offering cost-effective solutions with better storage capabilities. On the other hand, the flexibility and the time of processing are high as the big data is divided among the servers accordingly. Thus, the design of MRF is performed using two functions that are based on the optimisation algorithms.

MRF comprises of two phases, like the mapper and the reducer phases, that operate based on the proposed FrSparse FCM and P-Whale algorithm. The mapper phase uses the proposed FrSparse FCM method for optimally detecting the centroids, and the reducer function uses the P-Whale algorithm for obtaining the optimal clusters. Let us consider the database D that consists of a number of data with the attributes given as,

$$D = \{d_{ij}\} \tag{1}$$

where, d_{ij} corresponds to the data in the big data D or the jth attribute of the ith data present in the big data. Let us assume that there are a number of data points and b number of attributes in the database such that $(1 \le i \le a)$ and $(1 \le j \le b)$. The mapper phase consists of the mapper function that aims at transforming the subsets of data into key pairs, and the mapper phase comprises of a number of mappers based on the sub-sets of data. The map function uses the proposed FrSparse FCM that is the integration of the fractional concept in the sparse FCM algorithm, and the map function determines the optimal centroid. The data d_{ij} is partitioned into subsets in such a way that the total number of sub-sets equals to the total number of mappers in the mapper phase. The sub-sets of data obtained from d_{ij} are given as,

$$d_{ij} = \{J_k\}; \quad (1 \le k \le n) \tag{2}$$

where, n is the total number of sub-sets of data from d_{ij} and n is the total number of mappers in the mapper phase. Let us represent the mappers in the mapper phase given as,

$$F = \{F_1, F_2, \ldots, F_k, \ldots, F_n\}; \quad (1 \le k \le n) \tag{3}$$

Now, the input to the kth mapper is represented as,

$$J_k = \{g_{l,j}\}; \quad (1 \le l \le p); (1 \le j \le b) \tag{4}$$

where, $g_{l,j}$ is the data to the kth mapper, p is the total number of data in the qth mapper, and b represents the total number of the attributes. The individual mappers map the input data and form the centroids based on the user-defined centroid size, and establish an intermediate data. Thus, the mapper output forms the input to the reducer module, and the output from the n number of mappers is represented as,

$$H = [q_1|q_2|\ldots q_n|] \tag{5}$$

where, $[q_1|q_2|\ldots q_n|]$ corresponds to the output from the individual mappers and H is the intermediate data of the mappers. The reducer phase possesses the reducer function that determines the optimal cluster using the P-Whale algorithm for which it makes use of the centroids obtained using mappers (presented as intermediate data). The total number of reducers in the reducer phase is given as,

$$r = \{r_1, r_2, \ldots, r_t, \ldots, r_u\} \tag{6}$$

where, u is the total number of reducers in the reducer phase. Thus, the optimal clusters are generated from the reducer module of the MRF based on the P-Whale algorithm. The output from the reducer phase is given as,

$$w_t = \{w^t_{s,j}; (1 \le s \le K_t); (1 \le j \le b); (1 \le t \le u) \tag{7}$$

where, K_t signifies the cluster size defined for the t^{th} reducer and $w^t_{s,j}$ denotes the cluster in the t^{th} reducer.

3.1.1 Mapper phase with the proposed FrSparse FCM as map function: The mapper phase contributes to the big data clustering with the centroids for the big data d_{ij} that is formulated using the proposed FrSparse FCM algorithm. The developed FrSparse FCM algorithm inherits the advantages of fractional theory [47] and sparse FCM [48]. The Sparse FCM handles the high-dimensional data and possesses the ability to select the highly significant and effective cluster centroid. The fractional concept is capable of extracting the texture features effectively for the localisation of the optimal centroid. The fractional theory has the unique property of using the history of the variable as they possess the memory. The main goal of the proposed FrSparse-FCM is to compute the optimal centroid, and the proposed FrSparse-FCM algorithm is the map function in the mappers of the MRF. Let us denote the data matrix as, $J_k = g_{lj} \in \Re^{p \times b}$, in which p indicates the total number of data points in the data and b corresponds to the total number of attributes. The FrSparse-FCM algorithm employs the distance metric to compute the cluster centroid, and the total number of centroids is based on the user and is hence, predefined. The cluster centroids are given as,

$$w = \{w_1, w_2, \ldots, w_\rho, \ldots., w_A\} \tag{8}$$

where, A is the total number of cluster centroids. The cluster centroids are the highly significant data points that are highly essential to group the big data such that the processing and analysis using the big data become less complex and take less time. The algorithmic steps of the proposed FrSparse-FCM are given below.

Step 1: Initialisation: The main step in the proposed FrSparse-FCM is the initialisation of the data points in the big data that is given as,

$$\omega = \omega_1^o = \omega_2^o = \ldots = \omega_b^o = \frac{1}{\sqrt{b}} \tag{9}$$

Step 2: Update the partition matrix: Let us assume w as the cluster center and let us fix the attribute weights as, ω such that $\varepsilon(\mathfrak{R})$ is minimised if and only if,

$$P_{l\rho} = \begin{cases} \frac{1}{C_\rho}; & \text{if } \varepsilon_{l\rho} = 0 \text{ and } C_\rho = \text{card}\left\{w : \varepsilon_{l\rho} = 0\right\} \\ 0; & \text{if } \varepsilon_{l\rho} \neq 0 \text{ but } \varepsilon_{l\lambdabar} = 0 \text{ for some } \lambdabar, \lambdabar \neq \rho \\ \frac{1}{\sum_{\lambdabar=1}^{A}\left(\varepsilon_{l\rho}/\varepsilon_{\lambdabar\rho}\right)^{(1/(\alpha-1))}}; & \text{Otherwise} \end{cases} \tag{10}$$

where, card(A) specifies the cardinality of set A. The distance measure in the proposed FrSparse-FCM algorithm is computed between the individual data point g_{lj} in the big data and the cluster centroid $w_{\rho j}$, given as,

$$\varepsilon_{l\rho} = \sum_{\rho=1}^{A} \omega_j\left(g_{lj} - w_{\rho j}\right)^2 \tag{11}$$

where, g_{lj} is the data point, and $w_{\rho j}$ is the cluster centroid.

Step 3: Update the cluster centers: Let w and $\mathfrak{R}$ be fixed and $\varepsilon(c)$ is minimised if

$$w_{\rho j} = \begin{cases} 0; & \text{if } w_j = 0 \\ \frac{\sum_{l=1}^{p} P_{l,\rho}^{\alpha} \cdot g_{lj}}{\sum_{i=1}^{p} P_{l,\rho}^{\alpha}}; & \text{if } w_j \neq 0 \end{cases} \tag{12}$$

where, $l = 1, \ldots, p$, $j = 1, \ldots, b$, and α specifies the weighted exponent, which is responsible for handling the degree of membership sharing among the fuzzy clusters. The dissimilarity measure is indicated as, $\mathfrak{R}$. Equation (12) is the standard equation of the computation of cluster centroid, which is modified using the fractional theory. The proposed update rule of cluster centroid is derived as follows:

Let us assume the cluster centroid as $w_{\rho j}^{z+1} = N$. For including the fractional concept in the cluster centroid update equation, the assumption $w_{\rho j}^{z+1} = N$ is modified with the centroid obtained in the previous iteration as,

$$w_{\rho j}^{z+1} - w_{\rho j}^{z} = N - w_{\rho j}^{z} \tag{13}$$

$$\partial^{\alpha}\left[w_{\rho j}^{z+1} - w_{\rho j}^{z}\right] = N - w_{\rho j}^{z} \tag{14}$$

where, z is the iteration number. According to the fractional concept, (14) is given as,

$$\begin{aligned} & w_{\rho j}^{z+1} - \alpha w_{\rho j}^{z} - \frac{1}{2}\alpha \cdot w_{\rho j}^{z-1} - \frac{1}{6}(1-\alpha)w_{\rho j}^{z-2} \\ & - \frac{1}{24}\alpha(1-\alpha)(2-\alpha)w_{\rho j}^{z-3} = N - w_{\rho j}^{z} \end{aligned} \tag{15}$$

$$\begin{aligned} & w_{\rho j}^{z+1} = N - w_{\rho j}^{z}(1-\alpha) + \frac{1}{2}\alpha \cdot w_{\rho j}^{z-1} \\ & + \frac{1}{6}(1-\alpha)w_{\rho j}^{z-2} + \frac{1}{24}\alpha(1-\alpha)(2-\alpha)w_{\rho j}^{z-3} \end{aligned} \tag{16}$$

where, $w_{\rho j}^{z+1}$ is the cluster centroid in the $(z+1)^{\text{th}}$ iteration and α defines the fractional coefficient. $w_{\rho j}^{z-1}$, $w_{\rho j}^{z-2}$, and $w_{\rho j}^{z-3}$ are the cluster centroids in the previous iterations. Equation (16) conveys that the cluster centroid is updated based on the centroids in the previous iterations. The proposed update rule of the FrSparse FCM algorithm obtains the cluster centroids effectively with good accuracy, and it overcomes the demerits of the existing sparse FCM [48]. The proposed FrSparse FCM is applicable to the big data clustering that holds the ability to solve the data with variable features and missing data.

Step 4: Compute the class: The class label is estimated using the fixed clusters $w = \left\{w_1, w_2, \ldots, w_\rho, \ldots., w_A\right\}$ and the membership P. The class κ_j is computed based on the following objective, $\max_w \sum_{j=1}^{b} \omega_j \cdot \kappa_j$ such that $\| \omega \|_2^2 \leq 1$, $\| \omega \|_f^f \leq \tau$ to obtain ω^*.where, τ is the tuning parameter and $(0 \leq \tau \leq 1)$; $\| \omega \|_\hbar^\hbar = \sum_{j=1}^{b} \left|\omega_j\right|^\hbar$.

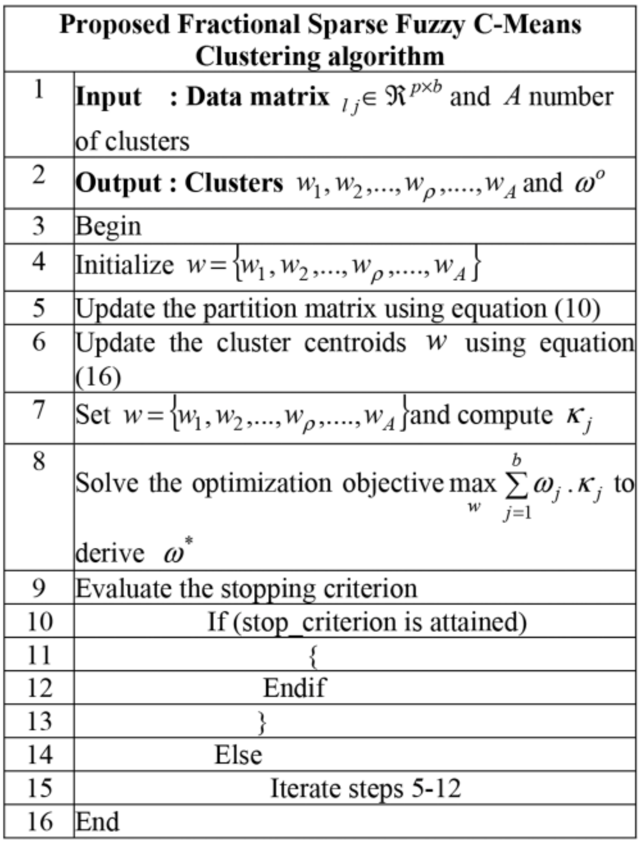
Step 5: Terminate: The iteration is repeated for the maximum number of iterations until the stopping criterion is reached. The stopping criterion is given as,

$$\frac{\sum_{j=1}^{b}\left|\omega^* - \omega_j^\hbar\right|}{\sum_{j=1}^{b}\left|\omega_j^\hbar\right|} < 10^{-4} \tag{17}$$

Thus, the cluster centroids are optimally tuned using the proposed FrSparse-FCM, and the number of the centroids is based on the user-set count, and the centroids group the clusters such that the data points in a cluster group exhibit similar characteristic, whereas the data points between the clusters exhibit dissimilar characteristics. The cluster centroids determined using FrSparse FCM in the individual mappers are combined to form the intermediate data that form the input to the reducers. The reducer phase uses the P-Whale algorithm to form the optimal clusters for the big data. Fig. 2 shows the algorithmic steps of the proposed FrSparse-FCM.

3.1.2 P-Whale algorithm for computing the optimal clusters in the reducer phase for big data clustering: The reduce function is based on the P-Whale algorithm that is developed using WOA [49]and PSO algorithm [50] such that the update rule in PSO is modified using the update rule in WOA. The P-Whale algorithm inherits the advantages of PSO and WOA, such that there is a proper balance in the merits and the demerits of the algorithms. WOA that exhibits the hunting mechanism of the humpback whales ensures the balance in the exploration and the exploitation phases. WOA is effective in dealing with real-time issues and capable of operating even with unknown search spaces. The algorithm exhibits unique features in employing the random or the best search in the solution space to chase the prey. The solution parameters are located effectively in three modes of search that includes the encircling phase, exploration, and exploitation phases.

On the other hand, PSO is based on the swarm behaviour that makes the possibility of the algorithm to search for the solution space effectively as the search space is large. The positions and velocities of the particle are updated based on the environmental change assuring proximity and quality. In other words, the swarm holds simple computations, consuming less time and could afford the changes in the quality associated with the environment. Moreover, the movement of the particles in PSO is not limited, but



Proposed Fractional Sparse Fuzzy C-Means Clustering algorithm	
1	**Input : Data matrix** $I_j \in \Re^{p\times b}$ and A number of clusters
2	**Output : Clusters** $w_1, w_2, ..., w_\rho, ..., w_A$ and ω^o
3	Begin
4	Initialize $w = \{w_1, w_2, ..., w_\rho, ..., w_A\}$
5	Update the partition matrix using equation (10)
6	Update the cluster centroids w using equation (16)
7	Set $w = \{w_1, w_2, ..., w_\rho, ..., w_A\}$ and compute κ_j
8	Solve the optimization objective $\max_{w} \sum_{j=1}^{b} \omega_j . \kappa_j$ to derive ω^*
9	Evaluate the stopping criterion
10	If (stop_criterion is attained)
11	{
12	Endif
13	}
14	Else
15	Iterate steps 5-12
16	End

Fig. 2 *Proposed FrSparse FCM in the MRF*

the optimal search is continued until the possible solution is obtained. Additionally, the behaviour of the particle never varies for all the changes in the environment, but adaptively changes the behaviour upon the worthy environmental variations. PSO offers excellent robustness, supports parallel computation, and converges easily to the optimal solution. However, the estimation of the algorithm parameters and the inability to deal with the discrete variables are serious issues of PSO. Moreover, there is no proper balance between the exploration and exploitation phases.

The P-Whale algorithm inherits the advantages and the disadvantages of the PSO and the WOA in determining the optimal clusters using the big data. The intermediate data obtained from the mapper phase is employed by the reducer that is represented as,

$$H = \{H^{v,j}\}; 1 \le v \le n^w \times n; 1 \le j \le b \tag{18}$$

where, n^w specifies the total number of the clusters and $H^{v,j}$ is the intermediate data belonging to the n number of mappers. The optimal cluster in the reducer phase is determined using the intermediate data H of the mapper. The cluster obtained in the reducer phase is,

$$r(H) = w_t \tag{19}$$

where, w_t specifies the clusters in the t^{th} reducer. Thus, the final output from the reducer is based on the following equation,

$$r_u = \{q_{e,j}; (1 \le k \le n); (1 \le e \le n_k^w); (1 \le j \le b) \tag{20}$$

where, n_k^w is the total number of the clusters obtained from the *k*th cluster such that the optimal cluster generation follows the objective function that is based on the DB index.

Solution vector: The solution is the optimal cluster obtained using the P-Whale algorithm-based MRF, and the solution is initially generated randomly based on the intermediate data generated using the mappers. The formulation of the optimal centroid is based on the objective measure, and the dimension is based on the total number of clusters and data.

Objective function: The objective function is based on the DB index [46] that evaluates the quality of the solution, and it measures the similarity between the clusters based on the distance measure. DB index is given as,

$$DB = \frac{1}{w_t} \sum_{h=1}^{w_t} G_h \tag{21}$$

G_h is the similarity measure for computing the similarity between the clusters that is given as,

$$G_h = \max_{x \ne h} \boldsymbol{I}_{h,x} \tag{22}$$

where, $\boldsymbol{I}_{h,x}$ is the Euclidean distance to measure the similarity between the clusters, given as,

$$\boldsymbol{I}_{h,x} = \frac{E_h + E_x}{B_{x,h}} \tag{23}$$

where, E_h and E_x are the scattering coefficients of two clusters, and $B_{x,h}$ denotes the Euclidean distance among two cluster matrices that signify the clustering performance. The lower values of $B_{x,h}$, between the cluster centroid and the data point constitute the better performance of the cluster. Euclidean distance $B_{x,h}$ is given as,

$$B_{x,h} = \| P_h - P_x \| \tag{24}$$

where, P_h and P_x denote the centroids of h^{th} and x^{th} clusters, respectively. On the other hand, the scattering coefficient of h^{th} cluster is given as,

$$E_h = \frac{1}{a^h} \sum_{h=1}^{a^h} \| H^{v,j} - P_h \| \tag{25}$$

where, a^h specifies the available number of data with respect to P_h. It is to be noted that the distance between the data point and its corresponding cluster centroid is minimum, but the distance between the cluster centroids should be in maximum.

Algorithmic steps of P-Whale algorithm: The optimal clusters are computed using the P-Whale algorithm, and the detailed steps are discussed as follows. The P-Whale algorithm determines the personal and the global best solutions based on the interaction of the particles in search space. PSO suffers from premature convergence, which is resolved using WOA that inherits good convergence behaviour without leaning on the local optima. The algorithmic steps of the proposed P-Whale algorithm are:

(a) Initialisation

The swarm population is initialised randomly, and the solutions in the search space are given as,

$$L = \{L_1, L_2, \ldots, L_f, \ldots, L_y\}; 1 \le f \le y \tag{26}$$

where, L_f denotes the position of f^{th} solution, whose dimension is $[1 \times M]$. The total number of swarm particles in search space is y.

(b) Evaluation of objective function

The fitness of the solutions given in (26) is evaluated based on the objective function shown in (21) such that the solution acquiring minimum fitness value becomes the best solution. Finally, global and personal best solutions are obtained using the algorithm.

(c) Update rule of P-Whale algorithm

The update rule of WOA is merged in the update rule of PSO that is based on the position and velocity of the particles in the search space.

$$L_f(z+1) = \frac{1}{[1 - R_2 S_2]}[L_f(z) + U\,V_t(z) + R_1 S_1\big(Y - L_f(z)\big) \\ - R_2 S_2\big(Z'\,e^{c\,N}\cos(2\pi N) + L_f(z)\big)] \tag{27}$$

where, z symbolises the iteration number, U denotes the inertia weight, R_1, R_2 signify the acceleration rates, S_1, S_2 notate the random numbers in [0, 1], $V_t(z)$ denotes the velocity of the particle at z^{th} iteration, and $L_f(z)$ signifies the position of f^{th} particle at z^{th} iteration. Z' is the distance measured based on the global best and is given as,

$$Z' = \left|Y - L_f(z)\right| \tag{28}$$

where, Y is the global best solution and X is the personal best solution. It is noted from (28) that the distance measure is computed as the distance between the global best solution and the position of f^{th} particle at z^{th} iteration. The global best solution is given as,

$$Y = L_f(z+1) - Z'\,e^{c\,N}\cos(2\pi N) \tag{29}$$

where, c is a constant and N is a random number on the range [−1, 1]. Equation (27) is the update rule of P-Whale algorithm employed for determining the optimal cluster.

(d) Locating the best solution

Once the position of the particles in search space is updated, the fitness of the updated solution is evaluated based on the objective function. The objective function uses the DB index, and the solution with the minimum value of the fitness is taken as the optimal solution that replaces the existing solution. The solution is the cluster centroid that is formulated for individual iteration to replace the existing centroid if the fitness of the new cluster centroid is better than the fitness of the existing centroid.

(e) Termination

The steps are iterated for the maximum number of iterations, and the optimal cluster centroid is determined to perform the optimal clustering of the big data.

4 Results and discussion

The section deliberates the results and discussion of the proposed method of big data clustering to prove the effectiveness of the proposed method.

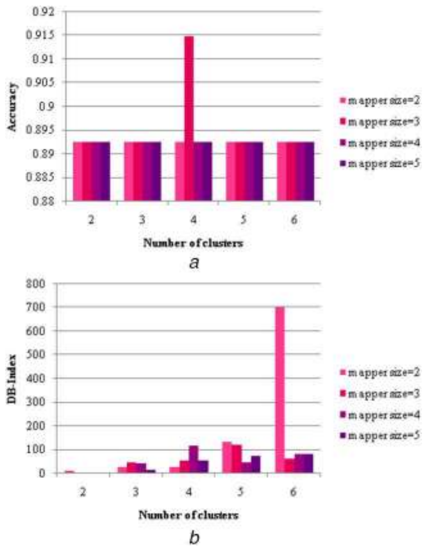


Fig. 3 *Performance analysis using skin dataset*
(a) Accuracy, ***(b)*** DB-index

4.1 Experimental setup

The implementation of the proposed work is performed in JAVA using two datasets, localisation, and skin segmentation datasets.

4.1.1 Dataset description: The datasets utilised for the experimentation are localisation dataset (Dataset 1) [51], skin segmentation dataset (Dataset 2) [51] (taken from UCI Machine Learning Repository), and Arrhythmia Data Set (Dataset 3) [52]. The localisation dataset has various activities of five different people wearing four tags, ankle left, ankle right, chest, and belt. There are a total of 164,860 instances in the localisation dataset with every instance constituting a localisation data for an individual tag that could be identified using 8 attributes. The skin segmentation dataset is developed through sampling the R, G, B values possessing four attributes and 245,057 instances to yield 50,859 skin samples and 194,198 non-skin samples.

4.2 Performance metrics

The performance of the proposed method of big data clustering is evaluated using two metrics, clustering accuracy, and DB-index. DB index is computed as in (21), and the classification accuracy is based on the number of classes. The effective method offers a maximum value of classification accuracy and DB index.

4.3 Comparative methods

The comparative methods of the big data clustering are performed using the methods, like Multiple Kernel and a Swarm-Based Map-Reduce Framework (MKS-MRF) [53], K-Means [54], FCM [55], KFCM [56], FPWhale-MRF [obtained by integrating fractional theory into TSK clustering algorithm, and PSO with WOA, and Sparse FCM [48].

4.4 Performance analysis

The section deliberates the performance analysis of the proposed big data clustering methods using the skin dataset and localisation dataset.

4.4.1 Using skin data: The analysis is progressed using the skin dataset that is depicted in Fig. 3, and the analysis is progressed for various mapper sizes. Fig. 3*a* depicts the accuracy of the proposed method for various mappers. When the mapper size is 3, the accuracy of the proposed method is 0.8924, 0.8924, 0.9148, 0.8924, and 0.8924, respectively, for the cluster size 2, 3, 4, 5, and 6, respectively. Fig. 3*b* depicts the DB-index of the proposed method for various mappers. When the mapper size is 3, the DB-Index of the proposed method is 3.4049, 48.0146, 54.6878, 122.7196, and 61.8339, respectively, for the cluster size 2, 3, 4, 5, and 6, respectively.

4.4.2 Using localisation data: The analysis is progressed using the localisation dataset that is depicted in Fig. 4, and the analysis is progressed for various mapper sizes. Fig. 4*a* depicts the accuracy of the proposed method for various mappers. When the mapper size is 3, the accuracy of the proposed method is 0.8559, 0.8559, 0.8559, 0.8559, and 0.8621, respectively, for the cluster size 2, 3, 4, 5, and 6, respectively. The accuracy of the proposed method increases with the cluster size. Fig. 4*b* depicts the DB-index of the proposed method for various mappers. When the mapper size is 3, the DB-Index of the proposed method is 5.0224, 12.5303, 97.1403, 135.343, and 85.586, respectively, for the cluster size 2, 3, 4, 5, and 6, respectively.

4.4.3 Using arrhythmia data set: The analysis is progressed using the Arrhythmia dataset that is depicted in Fig. 5, and the analysis is progressed for various mapper sizes. Fig. 5*a* depicts the accuracy of the proposed method for various mappers. When the mapper size is 3, the accuracy of the proposed method is 0.8195, 0.8247, 0.8469, 0.8147, and 0.8479, respectively, for the cluster size 2, 3, 4, 5, and 6, respectively. The accuracy of the proposed

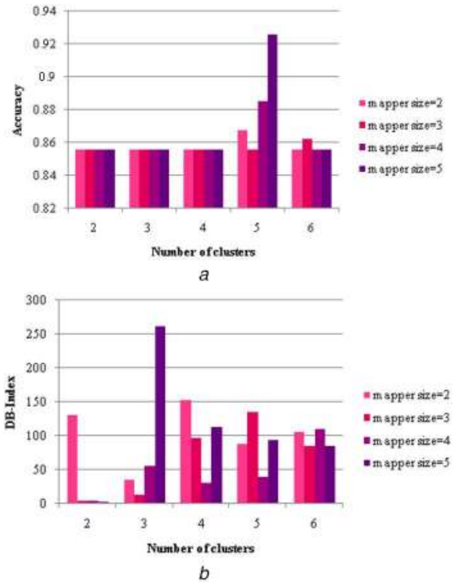


Fig. 4 *Performance analysis using localisation dataset*
(a) Accuracy, ***(b)*** DB-index

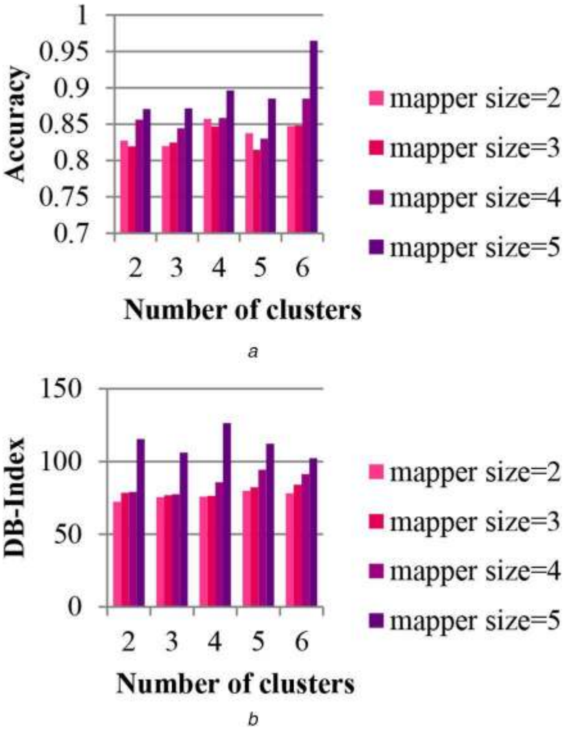


Fig. 5 *Performance analysis using arrhythmia data set*
(a) Accuracy, ***(b)*** DB-index

method increases with the cluster size. Fig. 5*b* depicts the DB-index of the proposed method for various mappers. When the mapper size is 3, the DB-Index of the proposed method is 78.5099, 76.7216, 76.2629, 82.3171, and 83.9616, respectively, for the cluster size 2, 3, 4, 5, and 6, respectively.

4.5 Comparative analysis

The section deliberates the comparative analysis of the big data clustering algorithms using two datasets, such as the skin and localisation dataset.

4.5.1 Using skin dataset: Fig. 6 shows the comparative analysis of the big data classification algorithms using the skin dataset. Fig. 6*a* deliberates the analysis based on accuracy. The accuracy is analysed using the cluster size, and at cluster size 2, the accuracy of the methods, MKS-MRF, K-Means, FCM, KFCM, FPWhale-MRF, Sparse FCM, and Proposed FrSparse FCM-based MRF is 75.58, 75.58, 75.58, 79.52, 87.91, 79.24, and 89.24%. The accuracy of the methods increases with the cluster size, and the proposed method acquired the maximum value of accuracy.

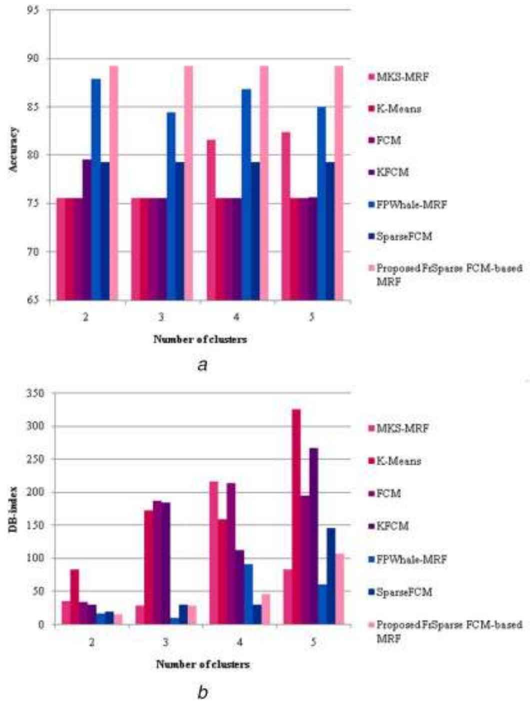


Fig. 6 *Analysis using skin dataset*
(a) Accuracy, ***(b)*** DB-index

Fig. 6*b* deliberates the analysis based on DB-Index. The DB-Index is analysed by varying the cluster size and at cluster size 2, the DB-Index of the methods, MKS-MRF, K-Means, FCM, KFCM, FPWhale-MRF, SparseFCM, and Proposed FrSparse FCM-based MRF is 36.05, 83.2, 34.04, 30.27, 16.93, 19.79, and 16.3. The proposed method acquired the minimum value of DB-Index than that of the existing methods.

4.5.2 Using localisation dataset: Fig. 7 shows the comparative analysis of the big data classification algorithms using the localisation dataset. Fig. 7*a* deliberates the analysis based on accuracy. The accuracy is analysed using the cluster size, and at cluster size 2, the accuracy of MKS-MRF, K-Means, FCM, KFCM, FPWhale-MRF, Sparse FCM, and Proposed FrSparse FCM-based MRF is 79.24, 78.57, 79.24, 79.24, 90, 75.599, and 90.60%. The accuracy of the methods increases with the cluster size, and the proposed method acquired the maximum value of accuracy.

Fig. 7*b* deliberates the analysis based on DB-Index. The DB-Index is analysed using the cluster size, and at cluster size 2, the DB-Index of MKS-MRF, K-Means, FCM, KFCM, FPWhale-MRF, Sparse FCM, and Proposed FrSparse FCM-based MRF is 12.01, 22.25, 88.17, 23.07, 7.73, 9.081, and 5.335. The DB-Index of the methods increases with cluster size, and the proposed method acquired the minimum value of DB-Index.

4.5.3 Using arrhythmia data set: Fig. 8 shows the comparative analysis of the big data classification algorithms using the Arrhythmia dataset. Fig. 8*a* deliberates the analysis based on accuracy. The accuracy is analysed using the cluster size, and at cluster size 2, the accuracy of MKS-MRF, K-Means, FCM, KFCM, FPWhale-MRF, Sparse FCM, and Proposed FrSparse FCM-based MRF is 75, 78.03, 79.73, 77.09, 78.73, 75.97, and 85.97%, respectively. The accuracy of the methods increases with the cluster size, and the proposed method acquired the maximum value of accuracy.

Fig. 8*b* deliberates the analysis based on DB-Index. The DB-Index is analysed using the cluster size, and at cluster size 2, the DB-Index of MKS-MRF, K-Means, FCM, KFCM, FPWhale-MRF, Sparse FCM, and Proposed FrSparse FCM-based MRF is

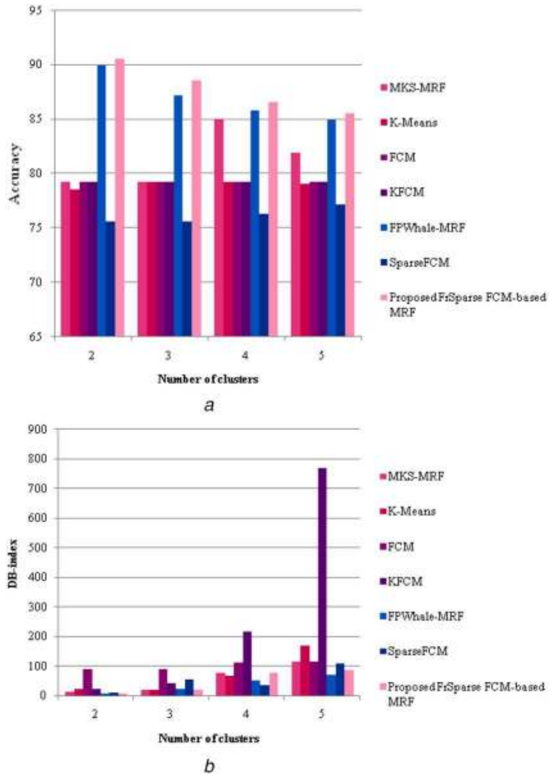


Fig. 7 *Analysis using localisation dataset*
(a) Accuracy, ***(b)*** DB-index

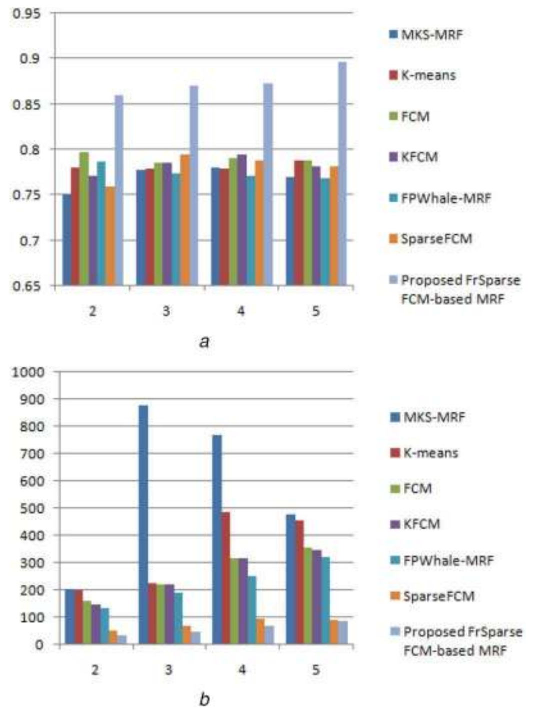


Fig. 8 *Analysis using arrhythmia data set*
(a) Accuracy, ***(b)*** DB-index

203.2028, 198.7232, 158.9375, 149.4166, 136.4192, 51.2953, and 36.0497, respectively. The DB-Index of the methods increases with cluster size, and the proposed method acquired the minimum value of DB-Index.

4.6 Comparative discussion

The comparative analysis of the big data classification methods is deliberated in Table 1. The accuracy of the methods, like MKS-MRF, K-Means, FCM, KFCM, FPWhale-MRF, Sparse FCM, and Proposed FrSparse FCM-based MRF is 82.43, 79.24, 79.73, 79.52, 90, 79.47, and 90.6012%. The proposed method acquired a maximum value of accuracy, with the minimum value of DB Index is 5.33. The existing methods, like MKS-MRF, K-Means, FCM, KFCM, FPWhale-MRF, and Sparse FCM attained a DB Index greater than the proposed method. The analysis reveals that the proposed method acquired a maximum value of accuracy and minimum value of the DB Index. The computational time of the proposed method is 4.8 s, which is minimum than the computational time of the existing methods, such as MKS-MRF, K-Means, FCM, KFCM, FPWhale-MRF, and Sparse FCM. An algorithm is said to be feasible if it requires minimum computational time. Here, the proposed algorithm requires minimum time for computation. Hence, the proposed algorithm is feasible.

Table 1 Comparative analysis

Methods	Accuracy, %	DB-index	Computational time, s
MKS-MRF	82.43	12.01	6.5
K-Means	79.24	18.11	8
FCM	79.73	34.04	7
KFCM	79.52	23.07	6
FPWhale-MRF	90	7.73	5
SparseFCM	79.47	9.08	6
proposed FrSparse FCM-based MRF	90.6012	5.33	4.8

5 Conclusion

This research work focuses on the big data clustering using the MRF based on the proposed FrSparse FCM algorithm. Initially, the mappers in the mapper phase compute the optimal centroid using the proposed FrSparse FCM algorithm, and the reducers in the reducer phase perform the classification using the P-Whale optimisation algorithm. The developed FrSparse FCM is the integration of the fractional concept in the sparse FCM such that the developed algorithm performs the big data clustering with better classification accuracy. The optimal centroids determined using the proposed FrSparse FCM in the mapper phase are optimally tuned in the reducer phase to present the better centroids in the reducer phase. The importance of the proposed method is that it can deal with big data arriving from the distributed sources. Experimentation is performed using the Skin dataset and localisation dataset taken from the UCI machine learning repository, and the analysis is progressed using the metrics, such as accuracy and DB Index. The analysis proves that the proposed method acquired a maximum accuracy of 90.6012% and a minimum DB Index of 5.33.

6 References

[1] Hidri, M.S., Zoghlami, M.A., Ayed, R.B.: 'Speeding up the large-scale consensus fuzzy clustering for handling big data', *Fuzzy Sets Syst.*, 2017, **1**, pp. 1–25

[2] Wu, X., Zhu, X., Wu, G.Q., *et al.*: 'Data mining with big data', *IEEE Trans. Knowl. Data Eng.*, 2014, **26**, (1), pp. 97–107

[3] Ermiş, B., Acar, E., Cemgil, A.T.: 'Link prediction in heterogeneous data via generalized coupled tensor factorization', *Data Min. Knowl. Discov.*, 2013, **29**, (1), pp. 203–236

[4] Zhang, Q., Yang, L.T., Chen, Z.: 'Deep computation model for unsupervised feature learning on big data', *IEEE Trans. Serv. Comput.*, 2016, **9**, (1), pp. 161–171

[5] Zhang, Q., Yang, L.T., Chen, Z., *et al.*: 'PPHOPCM: privacy-preserving high-order possibilistic c-means algorithm for big data clustering with cloud computing', *IEEE Trans. Big Data*, 2017, **7790**, pp. 1–1

[6] Rehioui, H., Idrissi, A., Abourezq, M., *et al.*: 'DENCLUE-IM: a new approach for big data clustering', *Proc. Comput. Sci.*, 2016, **83**, pp. 560–567

[7] Zhang, C., Hao, L., Fan, L.: 'Optimization and improvement of data mining algorithm based on efficient incremental kernel fuzzy clustering for large data', *Cluster Comput.*, 2019, **22**, pp. 3001–3010

[8] Wu, X., Kumar, V., Ross Quinlan, J., *et al.*: 'Top 10 algorithms in data mining', *Knowl. Inf. Syst.*, 2008, **14**, (1), pp. 1–37

[9] Fan, T.: 'Research and implementation of user clustering based on MapReduce in multimedia big data', *Multimed. Tools Appl.*, 2018, **77**, pp. 10017–10031

[10] Tan, P.N., Steinbach, M., Kumar, V.: 'Chap 8: cluster analysis: basic concepts and algorithms', Introd. to Data Min., 2005, Chapter 8

[11] Luna-Romera, J., García-Gutiérrez, M., Martínez-Ballesteros, M., *et al.*: 'An approach to validity indices for clustering techniques in big data', *Prog. Artif. Intell.*, 2018, **7**, pp. 81–94

[12] Madhulatha, T.S.: 'An overview on clustering methods', *IOSR J. Eng.*, 2012, **2**, (4), pp. 719–725

[13] Xu, R.: 'Survey of clustering algorithms for MANET', *IEEE Trans. Neural Netw.*, 2005, **16**, (3), pp. 645–678

[14] Drineas, P., Frieze, A., Kannan, R., *et al.*: 'Clustering large graphs via the singular value decomposition', *Mach. Learn.*, 2004, **56**, (1–3), pp. 9–33

[15] Welch, W.J.: 'Algorithmic complexity: three NP-hard problems in computational statistics', *J. Stat. Comput. Simul.*, 1982, **15**, (1), pp. 17–25

[16] Daga, B.S., Bhute, A.N.: 'Predicting recurrence pattern in breast cancer using decision tree', 2009

[17] Ghuge, C.A., Ruikar, S.D., Chandra Prakash, V.: 'Support vector regression and extended nearest neighbor for video object retrieval', *Evol. Intell.*, 2018, pp. 1–14

[18] Remmiya, R., Abisha, C.: 'Artifacts removal in EEG signal using a NARX model based CS learning algorithm', *Multimed. Res.*, 2018, **1**, (1), pp. 1–8

[19] Al-Sultan, K.: 'A tabu search approach to the clustering problem', *Pattern Recognit.*, 1995, **28**, (9), pp. 1443–1451

[20] Krishna, K., Murty, M.N.: 'Genetic K-means algorithm', *IEEE Trans. Syst. Man Cybern. B Cybern.*, 1999, **29**, (3), pp. 433–439

[21] Shelokar, P.S., Jayaraman, V.K., Kulkarni, B.D.: 'An ant colony approach for clustering', *Anal. Chim. Acta*, 2004, **509**, (2), pp. 187–195

[22] Cura, T.: 'A particle swarm optimization approach to clustering', *Expert Syst.*, 2012, **39**, (1), pp. 1582–1588

[23] Zhang, C., Ouyang, D., Ning, J.: 'An artificial bee colony approach for clustering', *Expert Syst.*, 2010, **37**, (7), pp. 4761–4767

[24] Kao, Y.T., Zahara, E., Kao, I.W.: 'A hybridized approach to data clustering', *Expert Syst.*, 2008, **34**, (3), pp. 1754–1762

[25] Jordehi, A.R.: 'Enhanced leader PSO (ELPSO): a new PSO variant for solving global optimisation problems', *Appl. Soft. Comput. J.*, 2015, **26**, pp. 401–417

[26] Tabakhi, S., Moradi, P., Akhlaghian, F.: 'An unsupervised feature selection algorithm based on ant colony optimization', *Eng. Appl. Artif. Intell.*, 2014, **32**, pp. 112–123

[27] Jordehi, A.R.: 'Brainstorm optimisation algorithm (BSOA): an efficient algorithm for finding optimal location and setting of FACTS devices in electric power systems', *Int. J. Electr. Power Energy Syst.*, 2015, **69**, pp. 48–57

[28] Heidari, A.A., Abbaspour, R.A., Jordehi, A.R.: 'An efficient chaotic water cycle algorithm for optimization tasks', *Neural Comput. Appl.*, 2017, **28**, (1), pp. 57–85

[29] Jordehi, A.R.: 'A chaotic artificial immune system optimisation algorithm for solving global continuous optimisation problems', *Neural Comput. Appl.*, 2015, **26**, (4), pp. 827–833

[30] Bijari, K., Zare, H., Veisi, H., *et al.*: 'Memory-enriched big bang–big crunch optimization algorithm for data clustering', *Neural Comput.*, 2018, **29**, pp. 111–121

[31] Dean, J., Ghemawat, S.: 'Mapreduce: simplified data processing on large clusters'. Proc. 6th Symp. on Operating Systems Design and Implementation, San Francisco, California, USA, 2004, pp. 137–149

[32] Ekanayake, J., Pallickara, S., Fox, G.: 'Mapreduce for data intensive scientific analyses'. Proc. – 4th IEEE Int. Conf. eScience, Indianapolis, IN, USA, 2008, pp. 277–284

[33] Liu, T., Rosenberg, C., Rowley, H.A.: 'Clustering billions of images with large scale nearest neighbor search'. Proc. – IEEE Workshop on Applications of Computer Vision (WACV), Austin, TX, USA, 2007

[34] Cui, X., Zhu, P., Yang, X., *et al.*: 'Optimized big data K-means clustering using MapReduce', *J. Supercomput.*, 2014, **70**, (3), pp. 1249–1259

[35] Schölkopf, B., Platt, J., Hofmann, T.: 'Map-reduce for machine learning on multicore', *Adv. Neural Inf. Process. Syst.*, 2007, **19**, pp. 281–288

[36] Polo, J., Carrera, D., Becerra, Y., *et al.*: 'Performance-driven task co-scheduling for mapreduce environments'. Proc. 2010 IEEE/IFIP Network Operations and Management Symp. (NOMS), Osaka, Japan, 2010, pp. 373–380

[37] Wu, J., Wu, Z., Cao, J., *et al.*: 'Fuzzy consensus clustering with applications on big data', *IEEE Trans. Fuzzy Syst.*, 2017, **25**, (6), pp. 1430–1445

[38] Zhang, Q., Yang, L.T., Castiglione, A., *et al.*: 'Secure weighted possibilistic c-means algorithm on cloud for clustering big data', *Inf. Sci. (Ny)*, 2019, **479**, pp. 515–525

[39] Son, L.H., Tien, N.D.: 'Tune up fuzzy C-means for big data: some novel hybrid clustering algorithms based on initial selection and incremental clustering', *Int. J. Fuzzy Syst.*, 2017, **19**, (5), pp. 1585–1602

[40] Hajeer, M.H., Dasgupta, D.: 'Handling big data using a data-aware HDFS and evolutionary clustering technique', *IEEE Trans. Big Data*, 2017, **7790**, (c), pp. 1–1

[41] Ilango, S.S., Vimal, S., Kaliappan, M., *et al.*: 'Optimization using artificial bee colony based clustering approach for big data', *Cluster Comput.*, 2019, **22**, pp. 12169–12177

[42] Traganitis, P.A., Slavakis, K., Giannakis, G.B.: 'Sketch and validate for big data clustering', *IEEE J. Sel. Top. Signal Process.*, 2015, **9**, (4), pp. 678–690

[43] Tsapanos, N., Tefas, A., Nikolaidis, N., *et al.*: 'Big data clustering with kernel k-means: resources, time and performance', *Int. J. Artif. Intell. Tools*, 2018, **27**, (4), pp. 1–18

[44] Shrivastava, P., Sahoo, L., Pandey, M., *et al.*: 'AKM – augmentation of K-means clustering algorithm for big data', *Intell. Eng. Informat.*, 2018, pp. 103–109

[45] Zhang, Q., Yang, L.T., Castiglione, A., *et al.*: 'Secure weighted possibilistic c-means algorithm on cloud for clustering big data', *Inf. Sci.*, 2019, **479**, pp. 515–525

[46] Kulkarni, O., Jena, S.: 'MKS-MRF: A multiple kernel and a swarm-based map reduce framework for big data clustering', *Int. Rev. Comput. Softw.*, 2016, **11**, (11), pp. 997–1006

[47] Bhaladhare, P.R., Jinwala, D.C.: 'A clustering approach for the *l* -diversity model in privacy preserving data mining using fractional Calculus-bacterial foraging optimization algorithm', *Adv. Comput. Eng.*,2014, **2014**, pp. 10–13

[48] Chang, X., Wang, Q., Liu, Y., *et al.*: 'Sparse regularization in fuzzy c -means for high-dimensional data clustering', *IEEE Trans. Cybern.*, 2016, **47**, (9), pp. 2616–2627

[49] Mirjalili, S., Lewis, A.: 'The whale optimization algorithm', *Adv. Eng. Softw.*, 2016, **95**, pp. 51–67

[50] Wang, D., Tan, D., Liu, L.: 'Particle swarm optimization algorithm: an overview', *Soft Comput.*, 2018, **22**, (2), pp. 387–408

[51] 'UCI Machine Learning Repository'. Available at https://archive.ics.uci.edu/ml/datasets/Localization+Data+for+Person+Activity, accessed: 18 April 2018

[52] Arrhythmia Data Set. Available at https://archive.ics.uci.edu/ml/datasets/arrhythmia

[53] Kulkarni, O., Jena, S.: 'MKS-MRF: a multiple kernel and a swarm-based map reduce framework for big data clustering', *Int. Rev. Comput. Softw.*, 2016, **11**, (11)

[54] Xia, D., Wang, B., Li, Y., *et al.*: 'An efficient MapReduce-based parallel clustering algorithm for distributed traffic subarea division', *Discret. Dyn. Nat. Soc.*, 2015, Article ID 793010, p. 18

[55] Yu, Q., Ding, Z.: 'An improved fuzzy C-means algorithm based on MapReduce'. Proc. of 8th Int. Conf. on Biomedical Engineering and Informatics (BMEI), Shenyang, China, 2015

[56] Zhu, H., Guo, Y., Niu, M., *et al.*: 'Distributed SAR image change detection based on spark'. Proc. of IEEE Int. Conf. on Geoscience and Remote Sensing Symp. (IGARSS), Milan, Italy, 2015