JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 38, XXX-XXX (2022)
DOI: 10.6688/JISE.20220X_38(X).00XX

Soft Computing Based Multi-View High Dimensional Data Clustering Algorithm for Big Data

DR. CH. RAJA RAMESH*[1], DR. RAKESH NAYAK[2], DR. VEERARAGHAVAN JAGANNATHAN[3] AND DR. P MURALIDHAR[4]

[1]*Vignan Institute of Technology and Science,*
Vignan Hills, Deshmukhi Village, Pochampally (Mandal),
Yadadri Bhuvanagiri (District) - 508284.
[1]*scholar.rajaramesh213@gmail.com*
[2]*OP Jindal University,*
Asst. Dean and Head,
Department of Computer Science and Engineering, Raigarh.
[2]*drrakeshnayak62@gmail.com*
[3]*Associate Professor, Department of CSE,*
Sri Vishnu Engineering College for Women,
Vishnupur, Bhimavaram, 534202 Andhra Pradesh.
[3]*drveeraraghavanjagannathan@gmail.com*
[4]*Vignan Institute of Technology and Science,*
Vignan Hills, Deshmukhi Village, Pochampally (Mandal),
Yadadri Bhuvanagiri (District) - 508284.
[4]*vijayamuralisarma@gmail.com*

Clustering is the most popular approach to solving complex tasks in data mining. Numerous real-time applications are based on the clustering process, and the advancement of technology results in the utilization of Fast data sets in different formats. Though clustering algorithms are abundant in the existing literature, they are unsuitable for handling high-dimensional data. This article presents a soft computing-based multi-view high-dimensional data clustering algorithm that can handle large amounts of data. The proposed clustering algorithm employs a weighted k-means and Whale Optimization (WO) algorithm to cluster the multi-view high dimensional data. The performance of the proposed work is tested and compared with state-of-the-art algorithms such as GMM (Gaussian mixture model), k-means, fast search and FCAN-MOPSO (Fuzzy-Based Graph Clustering Algorithm for Complex Networks with Multiobjective Particle Swarm Optimization) on variant datasets. The experimental results prove that the proposed clustering algorithm performs well regarding F-measure, FMI, and RI for datasets such as Internet advertisement, spam base, Image segmentation, and Cora dataset. As a result, the proposed model is far better than GMM, K-means, and fast search algorithms, and it almost achieves equivalent F-Measure, FMI and RI compared to FCAN-MOPSO.

Keywords: Clustering, soft computing, high dimensionality, multi-view data.

1. INTRODUCTION

Today's world expels enormous amounts of data as its end or by-product in all domains, necessitating a means to analyze and study the data to form valuable patterns. Data mining is a technology that analyses voluminous data to extract beneficial patterns. Knowledge Discovery (KD) is the most crucial objective of the data mining process. Based on the kind of knowledge, the process has different faces, such as association rules, clustering, classification, evaluation analysis and so on [1]. Out of all these, Association

DOI: 10.6688/JISE.20220X_38(X).00XX

Rule Mining (ARM), clustering and classification are the three significant areas of data mining [2,3].

ARM is meant to provide allegations between the data entities; clustering involves unsupervised learning that clusters data entities based on some criteria under unknown classes, and classification is a supervised learning technique that discriminates the data entity based on predefined knowledge. Clustering [39] requires no predefined knowledge. Hence, it is more suitable for many real-time applications such as image processing, voice mining, web mining [35], bioinformatics, text mining, and image mining [4].

Due to the rapid growth of various technologies such as Artificial Intelligence (AI), Internet of Things (IoT) and remote monitoring, data growth is uncontrollable and unpredictable. This has sown the seeds of 'Big Data', which can deal with massive volumes of data, in addition to veracity, velocity, value and variety. However, it takes work to handle large-scale multi-view-based high-dimensional datasets [36]. Different feature spaces and structures represent these multi-view high dimensional datasets. For instance, the patient dataset can be viewed from different perspectives, such as blood investigation data, genetic data and medical images. Hence, data clustering with multi-view data is quite complicated; however, the demand for clustering algorithms for multi-view data is increasing.

As the datasets involve data with different views, the clustered outcomes cannot be consistent for the whole dataset. Besides, multi-view high-dimensional data clustering completely differs from classical data clustering problems. Additionally, the computational complexity of a multi-view high-dimensional dataset is higher due to its structure, data sources, and data size.

The paper aims to address the above-stated issue by presenting a clustering algorithm based on a metaheuristic algorithm in combination with a weighted k-means algorithm. Three significant steps are involved in the proposed work to achieve the clusters. The weighted k-means algorithm is initially applied, and the cluster centre point [43], weight of views and features are fed into the Whale Optimization Algorithm (WOA). Finally, the performance of the work is assessed by comparing the attained results with the existing approaches. The contributions of the proposed work are as follows:

- The employment of the metaheuristic algorithm 'WOA' helps in the choice of a better cluster centre point.
- This work focuses on clustering multi-view high-dimensional datasets, which is minimal in the existing literature.
- The experimentation of the work is carried out in
- computational platforms such as Apache Spark and single node.

The rest of the article is arranged in the following style: Section 2 discusses the related literature review, and section 3 elaborates on the proposed clustering algorithm. Section 4 discusses the results attained by the proposed work and compares them with the existing clustering algorithms. Section 5 concludes the work.

2. RELATED LITERATURE

In [4], a distributed clustering approach is presented for High-Dimensional (HD) data based on local density subspace. This work is based on the belief that the local dense area

DOI: 10.6688/JISE.20220X_38(X).00XX

of the HD data is distributed in a low-dimensional space. The processing structure of this work considers both global and local structure, where the local dense areas are grouped and fitted by a subspace Gaussian model. Finally, all the clusters are merged by considering the broadcast from the global site.

In [5], density peak clustering is presented based on boundary detection. This work introduces an asymmetric measure that effectively detects the boundary points. This work utilizes two measures to segregate clusters with a better clustering degree. The authors claim this technique holds for both even and uneven data distributions.

In [6], a clustering approach is presented for spectral-embedded adaptive neighbours. This work presents a linear space-embedded clustering technique that uses adaptive neighbours that exclude the need for similarity matrix computation and low-dimensional data representation. The data representation is done in linear embedded spectral by linear regularization. The adaptive neighbours are employed for similarity matrix optimization, and the results are grouped.

Clustered data streams are processed by scalable distributed k-Nearest Neighbour (k-NN) in [7]. The high-dimensional k-NN queries are processed with a sliding window over data streams. This work is scalable for both users and data volume. A Dynamic Bounded Rings Index (DBRI) index structure is built on the data stream for indexing purposes, which detects the pivot initially and then allows the data points to the closest pivot for forming subsets.

A consensus Hilbert space and II-order neighbours are employed for performing multiple kernel subspace clustering in [8]. This work utilizes Hilbert space to deal with the non-linear subspace data, and the overlapping subspace problem is handled by II-order neighbours, which can optimize the sparseness along with better connectivity.

In [9], a multi-view clustering technique is presented on the basis of re-weighted discriminative embedded k-means algorithm. This work reduces the impact of outliers along with dimension reduction with the help of multi-view least absolute residual model. Iterative re-weighted least squares are incorporated in this work for solving least absolute residual and clustering indicator matrix formation.

A high dimensional clustering scheme with regularized Gaussian Mixture Model (GMM) in [10] is proposed. GMM is unsuitable for high dimensional clustering, so the regularized model is proposed. The component covariance matrices are formed by regularization, and the local feature correlations are measured [44,45,49]. The likelihood function of GMM is improved by the expectation-maximization algorithm through the M-step, which is based on the determinant maximization problem.

An ensemble clustering approach is introduced [11], [25] based on cumulative aggregations over random projections for high-dimensional data. The fuzzy partitions are aggregated [42] with the help of cumulative agreement and ranked with the help of external and internal indices of cluster validity. The best partition is declared the core partition, and the rest of the partitions provide inputs to the core partition.

[12] presents a high-dimensional static dataset with mixed attributes, 'CRAFTER,' with the help of a tree-ensemble clustering algorithm. This work can deal with categorical and numerical attributes. The indicative data points are detected by class probability estimates to perform a clustering operation.

In [13], a high-dimensional data clustering algorithm based on a fast adaptive k-means subspace clustering scheme is presented. Here, an adaptive loss function is introduced to

DOI: 10.6688/JISE.20220X_38(X).00XX

compute the cluster indicator. The optimal feature subset is built, and feature selection is performed by Eigenvalue decomposition.

A large-scale document categorization scheme is presented based on fuzzy clustering in [14]. This work considers both large-scale and high-dimensionality and employs three schemes: sampling extension, single pass, and divide ensemble. The drawbacks of the Fuzzy C Means (FCM) clustering algorithm are analyzed, and the hyperspherical FCM with fuzzy co-clustering is presented with scale-up schemes.

In [15], a high-dimensional clustering approach is presented based on Artificial Intelligence (AI) for incomplete mixed datasets. This work builds a phase space reconstruction by evaluating the mixed data, and the feature correlation value is detected by the correlation dimension. Standard deviation is then computed, and the features are extracted by computing the sparsity of sample features [41]. A repetitive clustering process concerns the correlation among the mixed datasets.

An approach for simultaneous subspace clustering along with a cluster count estimator is presented in [16], which is based on a triplet relationship. This work formulates the triplet relationship based on a hyper-correlation-based data structure. The correlations between triple entities are built by a self-representation matrix, and the data points are assigned to the clusters in an iterative manner. The cluster allocations are automatically carried out by an optimization technique, such that the similarity of triple entities between different clusters is maximized, and the relationship between the entities of the same cluster is minimized. Over-segmentation is avoided in this work by the automatic computation of cluster count.

In [17], a flexible high-dimensional clustering scheme that can handle missed data is proposed. This work employs Generalized Hyperbolic Factor Analyzers (MGHFA), a flexible modelling approach that can handle missing data. An expectation conditional maximization algorithm is presented to set the parameters of the MGHFA model by following various patterns of the missed data.

In [18], a cluster boundary detection algorithm based on multidimensional balance is proposed for high dimensional data. This work evaluates the high-dimensional neighbourhood space from various perspectives, and every dimension of the data point is simulated by kNN, where the lever theory is applied to compute the balance fulcrum of all dimensions. The distance between the data point coordinate and balance fulcrum of every dimension and DHBlan coefficient computes the balance of the neighbourhood space. Several works concerning feature selection are discussed in different works [19-27].

Though there is abundant literature on clustering, the clustering algorithms meant for multi-view high-dimensional data are scarce. The following section explains the proposed clustering algorithm.

3. PROPOSED MULTI-VIEW HIGH DIMENSIONAL DATA CLUSTERING ALGORITHM BASED ON WEIGHTED K-MEANS AND LOA

The main goal of the clustering technique is to group related data entities under a cluster, and the unrelated data entities are placed in different clusters. Consider a group of data entities $DE = [de_{i,j}]_{C \times D}$, where C is the count of data entities, and the dimension of

DOI: 10.6688/JISE.20220X_38(X).00XX

the data object is represented by D, which means that the data entity has D features. The clustering process divides the data entities DE into k groups. Let the cluster centre points be represented by $CCP = \left[ccp_{k,j}\right]_{CP\times D}$ and the membership degree of data entities within a cluster is denoted by a fuzzy division matrix $FDM = \left[fdm_{i,k}\right]_{C\times CP}$.

3.1 Weighted k-means Algorithm

The k-means algorithm is the most significant clustering algorithm, and it is employed in numerous real-time applications based on image processing and data mining [28,29]. However, in the standard k-means algorithm, all the features are assigned equal weights, which is different for numerous real-time applications. The influence of features varies, so it is essential to allot varied weights to the features. This is attained by the weighted k-means algorithm [30], and the following equation presents the algorithm's objective function.

$$F(PM, CCP, FW) = \sum_{k=1}^{CP} \sum_{i=1}^{C} \sum_{j=1}^{D} b_{i,k}\, fw_j \left(de_{i,j} - ccp_{k,j}\right)^2 \qquad (1)$$

Where

$$b_{i,k} \in \{0,1\}; \sum_{k=1}^{cp} b_{i,k} = 1 \qquad (2)$$

$$\sum_{j=1}^{D} fw_j = 1, 0 \leq fw_j \leq 1 \qquad (3)$$

In the above equations, PM is the partition matrix with $c \times k$, FW is the feature weight, $b_{i,k}$ is a binary variable, $CCP = \{ccp_1, ccp_2, \dots, ccp_k\}$ indicate the vectors with the centre points of k clusters. $\left(de_{i,j} - ccp_{k,j}\right)^2$ is the distance between the i^{th} data entity and the k^{th} cluster centre on j^{th} variable.

3.2 WO ALGORITHM

WOA is a bio-inspired algorithm that mimics the behaviour of whales [29,47,48]. Usually, whales attack a group of fish by generating bubbles that encircle the fish. The whales attack fish in two steps: exploitation and exploration. The exploitation step encircles the fish, and the exploration step randomly looks out for the fish. The activities of whales are represented in the following way.

$$P = \left|C_1 . \overrightarrow{X^*}(i) - \vec{X}(i)\right| \qquad (4)$$

$$\vec{X}(i+1) = \overrightarrow{X^*}(i) - \overrightarrow{C_2} \cdot P \qquad (5)$$

In the above equations, i indicates the current iteration, X^* is the optimal result achieved, and X denotes the position vector. The pipe operator (||) indicates the absolute value, and the dot product $(\cdot)$ is carried out on all elements. The coefficient vectors are denoted by C_1 and C_2, which are calculated by the following equations.

$$\overrightarrow{\mathrm{C_2}} = 2\overrightarrow{\mathrm{c_2}} \cdot \overrightarrow{\mathrm{rd}} - \overrightarrow{\mathrm{c_2}} \qquad (6)$$

$$\overrightarrow{\mathrm{C_1}} = 2 \cdot \overrightarrow{\mathrm{rd}} \qquad (7)$$

In equations (6 and 7), the $\overrightarrow{c_2}$ decreases with increasing iterations, and $\overrightarrow{rd}$ is the random vector between 0 and 1. The whales tend to relocate themselves concerning the location of the food availability based on eqn. (6) the location is taken care of by $\overrightarrow{C_1}$ and $\overrightarrow{C_2}$. As stated earlier, the foodstuffs of whales is encircled by reducing the $\overrightarrow{c_2}$, as shown in eqn.(8).

$$c = 2 - i\frac{2}{Grt_i} \qquad (8)$$

In the above equation, i represents the total number of iterations, and Grt_i represents the most significant number. The location of the neighbouring whale is calculated by considering the distance between the whales a and b, as given in the following equation.

$$\vec{\mathrm{X}}(\mathrm{i}+1) = \mathrm{P}' \cdot \mathrm{d}^{\mathrm{csrv}} \cdot \cos(2\pi \mathrm{rv}) + \vec{\mathrm{X}} \times (\mathrm{i}) \qquad (9)$$

In eqn.(9), $P' = \left|\vec{X}^{*}(i) - \vec{X}(t)\right|$ indicates the distance between the n^{th} whale and the found optimal food source, cs is the constant that represents the curve, and rv is the random number that lies in the range from -1 to 1. The food source location and the path formation are measured using a probability (prb) of 0.5, as represented in the following equation.

$$\vec{X}(i+1) = \begin{cases} Encircling\ food\ source\ eqn.(2)\ when\ prb < 0.5 \\ \quad path\ formation\ eqn.(6) \quad when\ prb \geq 0.5 \end{cases} \qquad (10)$$

In equation (10), prb is a random value lies in between 0 and 1. In the exploration step, the whales are selected randomly to search for the food source. The vector $\overrightarrow{C_1}$ with random numbers explores the best possible neighbouring whale as denoted by the following equations.

$$\vec{P} = \left|\overrightarrow{C_1}.\overrightarrow{X_{rd}} - \vec{X}\right| \qquad (11)$$

$$\vec{X}(i+1) = \overrightarrow{X_{rd}} - \overrightarrow{C_2} \cdot \vec{P} \qquad (12)$$

$\overrightarrow{X_{rd}}$is the whale selected randomly.

Hence, this section explains the WOA algorithm and the following part describes the proposed clustering algorithm.

3.3 Proposed Multi-view High Dimensional Clustering Algorithm (MHDCA)

The proposed clustering algorithm handles the high dimensional data with multiple views by considering the weights of features and views. The WO algorithm helps MHDCA in choosing the best cluster centre points. The population of whales is initialized, and the fitness value of the whale is computed using equation (13). The location of the whale is updated, and the search process continues with the objective of finding the best whale. The WO algorithm is employed due to its degree of exploration with respect to changes in location and convergence level. The fitness function of the WO algorithm is computed as follows.

$$FF(PM, CCP, VW, FW) = \frac{\sum_{k=1}^{CP}\sum_{i=1}^{C}\sum_{p=1}^{P}\sum_{j\epsilon view_p} b_{i,k} vw_p fw_j \left(de_{i,j} - ccp_{k,j}\right)^2}{\sum_{k=1}^{CP}\sum_{p=1} P \sum_{j\epsilon view_p} vw_t fw_j \left(ccp_{k,j} - q_j\right)^2} \qquad (13)$$

Where

$$\sum_{k=1}^{CP} b_{i,k} = 1, 1 \leq i \leq C, b_{i,k} \in [0,1] \qquad (14)$$

$$\sum_{p=1}^{P} vw_p = 1, 0 \leq vw_{Op} \leq 1 \qquad (15)$$

$$\sum_{j\in view_p} fw_j = 1, 0 \leq fw_j \leq 1, 0 \leq p \leq P \qquad (16)$$

$$oq_j = \sum_{k=1}^{CP} ccp_{k,j} / CP \qquad (17)$$

The fitness function of this work considers the view and feature weights for clustering the input data into clusters. In the above-given equations, $PM = \left[pm_{i,k}\right]_{C\times CP}$ and its entities are binary, such that $pm_{i,k} = 1$, when the data entity i is assigned to the cluster k.$CCP = \left[ccp_{k,j}\right]_{CP\times D}$is a $C \times CP$ matrix and the entities of $ccp_{k,j}$ indicate the j^{th} feature of the k^{th} cluster. $VW = \left[vw_p\right]_P$are the corresponding weights of P views.$FW = \left[fw_j\right]_j \in View_p$represents the feature weights concerning the $view_p$. $vw_p fw_j \left(de_{i,j} - ccp_{k,j}\right)^2$ is the weighted distance on j^{th} feature between the object i and cluster centre point k. $vw_t fw_j \left(ccp_{k,j} - q_j\right)^2$is the weighted distance of feature j between the cluster k and the mean cluster centre. $q_j$is the mean cluster centre of CP clusters. The outcome of this measure presents the coupling degree between the clusters and the greater the value, the more is the dissimilarity. The overall algorithm of the proposed work is presented as follows.

```
Proposed Clustering Algorithm based on weighted k-means and WOA
Input: Multi-view high dimensional data
Output: Data clusters
Begin
  Initialize the population of whales $A_i (i = 1,2,\dots,n)$;
  Compute the fitness value of $A_i$ by eqn. (13);
  Let $A *$ be the best whale;
  While (t<max_iter)
    For every whale
     If ($prb < 0.5$)
      Update location and the search by eqn (18);
     Apply weighted k-means;
     Else
       Update the location of whale by eqn (19);
     End if;
   End for;
Calculate the fitness of all whales;
Update $A *$ when better solution exists;
Apply weighted k-means;
t+=1;
End while;
```

```
Return A*;
```

In the entire population of whales, each whale i represents a solution in the D-dimensional solution space, as represented by the following equations.

$$P^d = \left| C_1 . \overrightarrow{X^{d^*}}(i) - \overrightarrow{X^d}(i) \right| \tag{18}$$

$$\overrightarrow{X^d}(i+1) = \overrightarrow{X^{d^*}}(i) - \overrightarrow{C_2} \cdot P^d \tag{19}$$

Here, d denotes the dimension of the search space. The whales are encoded to extract the best solution, which is done by encoding the initial cluster centre point, weights of views and features. The whales are encoded by a real number vector computed by $FC \times CP + P + FC$, where FC is the feature count of objects in the clustering process. The following equation can represent the encoded ith whale.

$$WL_i = \begin{bmatrix} wl_i^{1,1}, wl_i^{1,2}, \dots, wl_i^{1,FC}, \dots wl_i^{CP,1}, wl_i^{CP,2}, \dots, wl_i^{CP,FC} \\ vw_i^1, \dots, vw_i^P, fw_i^1, \dots, fw_i^{FC} \end{bmatrix} \tag{20}$$

The WO algorithm was chosen for its better exploration, and the location update results in detecting the best solution. The proposed work's performance is analyzed, and the results are discussed in the next section.

4. RESULTS AND DISCUSSION

Apache Spark is a cluster-based framework that offers a programming interface for performing clustering processes with parallel processing and fault tolerance [31]. This work utilizes the Apache Spark environment, which is meant for big data applications. The proposed algorithm is tested in environments such as Apache Spark and single node. A system with 8 GB RAM with $a\ core\ i5$ processor is employed in a single node. The Apache Spark environment is equipped with ten worker nodes with 500 GB cloud disk and 16 GB Double Data Rate III (DDRIII), and the master node is equipped with 1 TB cloud disk and 64 GB DDRIII. The machine learning [34] library of Apache Spark and Resilient Distributed Datasets (RDD) supports the proposed MHDCA to suit big data [38,40] applications.

The proposed work's performance is compared with the existing approaches, such as the GMM model [10], k-means [13] and fast search [32], in terms of precision, recall, F-measure, Fowles Mallow Index (FMI) and Rand Index (RI). The multi-view high-dimensional datasets utilized for testing the proposed work are Internet advertisement, spam base, and image segmentation. The population size is set at 30. A summary of the datasets is presented in Table 1.

The Internet advertisement dataset is comprised of 3279 images categorized as 'advertisement' and 'no advertisement'. All the entities are presented in six views, which include the image geometrical information, page URL, image URL, target URL, anchors, and text.

Table 1. Dataset Information

Data set/Features	Internet Advertisement	Spambase	Image Segmentation	Cora Dataset
Number of data items	2359	4601	2310	2708
Class count	2	2	7	7
Feature Count	1557	57	19	1433
ViewCount	6	3	2	158

The spam base dataset includes a collection of spam mail containing 4601 objects categorized under 'spam' and 'non-spam'. 57 features and three views denote all the objects. The image segmentation dataset contains 2310 objects with 19 features and 2 views [33]. The performance measures are discussed as follows.

Cora dataset includes 2708 objects categorized into 7 classes. All objects contain 1433 features and 158 views.

4.1 F-measure

This measure determines the accurate mapping of data points to the cluster. The F-measure is computed using the following equation.

$$F_m = \frac{2\times P\times R}{P+R} \qquad (21)$$

Here, P and R denote the precision and recall rates, respectively and are denoted as follows.

$$P = \frac{T_P}{T_P+F_N}\times 100 \qquad (22)$$

$$R = \frac{T_N}{T_N+F_P}\times 100 \qquad (23)$$

This indirectly implies that the F-measure score of a clustering algorithm depends on the precision and recall rates.

4.2 Fowles Mallow Index (FMI)

This external assessment technique is employed to balance the similarity of two clustering operations. Greater FMI indicates that there is more similarity between the clusters and the ground truth of the benchmark clusters, which is computed by

$$\mathrm{FMI} = \sqrt{\frac{\mathrm{T_P}}{\mathrm{T_P+F_P}}\cdot\frac{\mathrm{T_P}}{\mathrm{T_P+F_N}}} \qquad (24)$$

4.3 Rand Index (RI)

This measure computes the percentage of correct clustering activity between the data entities and their corresponding classes. It considers both the F_P and F_N rates equally and is computed as follows.

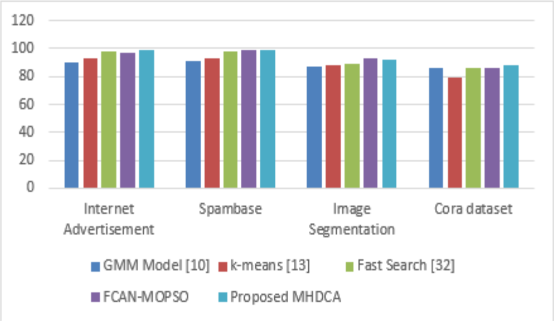
$$RI = \frac{T_P+T_N}{T_P+F_P+T_N+F_N} \qquad (25)$$

The results of the proposed approach over four different datasets concerning $P, R, F_m, FMI\ and\ RI$ are presented and compared with the existing approaches, as follows. As the F-measure rate relies on $P$and R, the results are tabulated in Table 2.

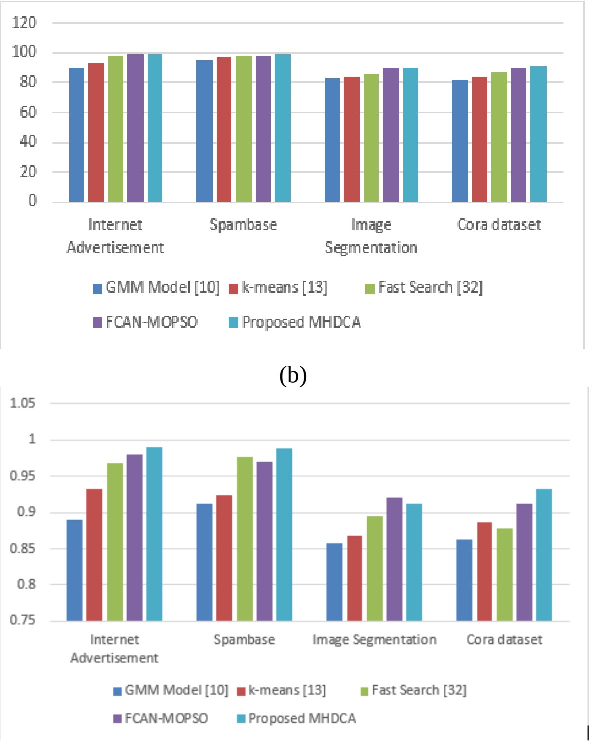
Table 2. P, R and F-measure (%) Results

Precision Rate Analysis (%)					
Techniques/Dataset	GMM Model [10]	k-means [13]	Fast Search [32]	FCAN-MOPSO [46]	Proposed MHDCA
Internet Advertisement	89.98	93.41	97.82	97.40	98.94
Spambase	91.28	92.78	97.84	98.65	98.79
Image Segmentation	86.98	88.17	89.37	93.52	91.87
Cora dataset	85.88	79.12	86.25	86.53	88.25
	Recall Rate Analysis (%)				
Internet Advertisement	89.92	93.67	98.28	99.02	99.09
Spambase	95.64	97.24	98.63	98.56	99.34
Image Segmentation	82.94	84.32	86.48	90.15	89.97
Cora dataset	81.58	83.58	86.88	90.15	91.25
	F-measure Analysis (%)				
Internet Advertisement	89.94	93.53	98.04	95.58	99.01
Spambase	93.4	94.95	98.23	98.23	99.06
Image Segmentation	84.91	86.2	87.9	91.25	90.91
Cora dataset	83.67	86.26	86.56	88.30	89.72

Table 2 shows the precision, recall, and F-measure rates attained by the proposed clustering algorithm, and the graphical results are depicted below.



(a)



(b)

(c)

Fig.1. (a) Precision rate analysis (b) Recall rate analysis (c) F-measure rate analysis

The greater precision and recall rate indicates that the data entities are correctly assigned to the corresponding clusters. Better precision and recall rates indicate that the clustering algorithm involves minimal false negative and false positive rates respectively. These false rates increase when the data entities are allotted to inappropriate clusters, which when compared with the ground truth. The proposed work shows decent precision and recall rates and so is the F-measure rate. Greater F-measure rate proves the correctness of the clustering algorithm. The following section shows the FMI results.

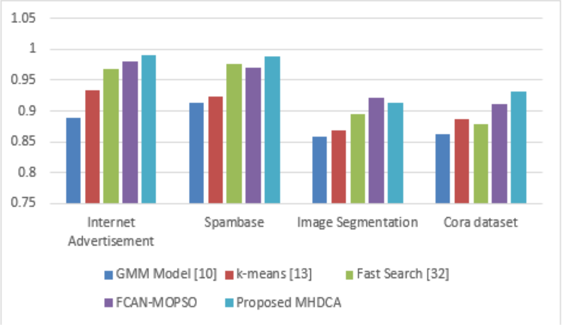


Fig.2. FMI analysis

DOI: 10.6688/JISE.20220X_38(X).00XX

Fig.2 shows that the proposed clustering algorithm shows better FMI rates than the existing approaches. This means that the proposed algorithm ideally assigns data entities to the clusters, which matches the ground truth of the utilized datasets. The following figure shows the rand index attained by the proposed algorithm compared to the existing works.

Table 3. Rand Index analysis

Techniques/Dataset	GMM Model [10]	k-means [13]	Fast Search [32]	FCAN-MOPSO [46]	Proposed MHDCA
Internet Advertisement	0.7386	0.8126	0.9143	0.9210	0.9587
Spambase	0.7957	0.8396	0.8964	0.9245	0.9238
Image Segmentation with Apache Spark	0.5362	0.8110	0.8256	0.9182	0.8846
Cora Data set	0.6582	0.7815	0.8821	0.9345	0.9246

The experimental results prove that the proposed work performs better than the existing work. Hence, the proposed clustering algorithm works well for multi-view high-dimensional datasets on both Apache Spark and single nodes. The results presented above are the mean outcome of ten runs.

5. CONCLUSIONS

This article proposes a novel clustering algorithm based on weighted k-means and Whale Optimization (WO) for multi-view high dimensional datasets. This article deals with the weight of features while building the objective function. The WO algorithm chooses the initial cluster centre points. The performance of the clustering algorithm is evaluated with performance measures such as precision, recall, F-measure, FMI and RI over variant datasets such as Internet advertisement, spam base, Image segmentation and Cora dataset. Our proposed algorithm outperforms compared to GMM, K-means and fast search on all four datasets. Compared with FCAN-MOPSO, the proposed algorithm results better with Internet advertisement, spam base, and Cora datasets. However, it results in slight variations in F-measure, FMI, and RI with image segmentation. Enhancing the algorithm for data with contaminated noise and outliers into clusters with unequal sizes as a future direction

REFERENCES

1. S. C. Pandey and G. C. Nandi. "Convergence of knowledge, nature and computations: A review. Soft computing—A fusion of foundations," *Methodologies and Applications*, Vol. 20, 2016, pp. 319–342.
2. R. Agrawal and R. Srikant. "Fast algorithms for mining association rules," *Proceedings of the 20th VLDB Conference* 1994, pp. 487–499.

DOI: 10.6688/JISE.20220X_38(X).00XX

3. J. Han and M. Kamber. "Data mining: Concepts and techniques," *Massachusetts, MA: Morgan Kaufmann Publishers,* 2006.

4. Y. A. Geng, Q. Li, M. Liang, C. Y. Chi, J. Tan, and H. Huang. "Local-Density Subspace Distributed Clustering for High-Dimensional Data," *IEEE Transactions on Parallel and Distributed Systems*, Vol. 31, No. 8, 2020, pp. 1799-1814.

5. D. Qiao, Y. Liang and L. Jiao. "Boundary detection-based density peaks clustering," *IEEE Access*, Vol. 7, 2019, pp. 152755-152765.

6. Q. Wang, Z. Qin, F. Nie and X. Li. "Spectral embedded adaptive neighbors clustering," *IEEE transactions on neural networks and learning systems*, Vol. 30, No. 4, 2018, pp.1265-1271.

7. M. Yang, Y. Zuo, M. Chen and X. Yu. "Scalable distributed kNN processing on clustered data streams," *IEEE Access*, Vol. 7, 2019, pp. 103198-103208.

8. Z. Wang and J. Liu. "Multiple Kernel Subspace Clustering Based on Consensus Hilbert Space and Second-Order Neighbors," *IEEE Access*, Vol. 8, 2020, pp. 124633-124645.

9. J. Xu, J. Han, F. Nie and X. Li. "Re-weighted discriminatively embedded $ k $-means for multi-view clustering," *IEEE Transactions on Image Processing*, Vol. 26, No. 6, 2017, pp. 3016-3027.

10. Y. Zhao, A. K. Shrivastava, and K. L. Tsui. "Regularized Gaussian mixture model for high-dimensional clustering," *IEEE transactions on cybernetics*, Vol. 49, No. 10, 2018, pp. 3677-3688.

11. P. Rathore, J. C. Bezdek, S. M. Erfani, S. Rajasegarar, and M. Palaniswami. "Ensemble fuzzy clustering using cumulative aggregation on random projections," *IEEE Transactions on Fuzzy Systems*, Vol. 26, No. 3, 2017, pp. 1510-1524.

12. S. Lin, B. Azarnoush, and G. C. Runger. "Crafter: a tree-ensemble clustering algorithm for static datasets with mixed attributes and high dimensionality," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 30, No. 9, 2018, pp. 1686-1696.

13. X. D. Wang, R. C. Chen, F. Yan, Z. Q. Zeng and C. Q. Hong. "Fast adaptive K-means subspace clustering for high-dimensional data," *IEEE Access*, Vol. 7, 2019, pp. 42639-42651.

14. J. P. Mei, Y. Wang, L. Chen and C. Miao. "Large scale document categorization with fuzzy clustering," *IEEE Transactions on Fuzzy Systems*, Vol. 25, No. 5, 2016, pp. 1239-1251.

15. M. Li, X. Li and J. Li. "High-Dimensional Clustering for Incomplete Mixed Dataset Using Artificial Intelligence," *IEEE Access*, Vol. 8, 2020, pp. 69629-69638.

16. J. Liang, J. Yang, M. M. Cheng, P. L. Rosin and L. Wang. "Simultaneous subspace clustering and cluster number estimating based on triplet relationship," *IEEE Transactions on Image Processing*, Vol. 28, No. 8, 2019, pp. 3973-3985.

17. Y.Wei, Y. Tang and P. D. Mc Nicholas. "Flexible High-Dimensional Unsupervised Learning with Missing Data," *IEEE transactions on pattern analysis and machine intelligence*, 2018.

18. X. Cao, B. Qiu, X. Li, Z. Shi, G. Xu and J. Xu. "Multidimensional Balance-Based Cluster Boundary Detection for High-Dimensional Data," *IEEE transactions on neural networks and learning systems*, Vol. 30, No. 6, 2018, pp. 1867-1880.

19. S. P. Potharaju and M. Sreedevi. “A Novel Cluster of Quarter Feature Selection Based on Symmetrical Uncertainty,” *Gazi University Journal of Science*, Vol. 31, No. 2, 2018.

20. K. R. Nirmal and K. V. V. Satyanarayana, “REDIC K–Prototype Clustering Algorithm for Mixed Data (Numerical and Categorical Data),” *International Journal of Recent Technology and Engineering*, Vol. 7, No. 6, pp. 1-6

21. R. B. Vallabhaneni and V. Rajesh, “Brain tumour detection using mean shift clustering and GLCM features with edge adaptive total variation denoising technique,” *Alexandria engineering journal*, Vol. 57, No. 4, 2018, pp. 2387-2392.

22. S. P. Potharaju and M. Sreedevi. “Correlation Coefficient Based Feature Selection Framework Using Graph Construction,” *Gazi University Journal of Science*, Vol. 31, No. 3, 2018.

23. A. KousarNikhath and K. Subrahmanyam, “Feature selection, optimization and clustering strategies of text documents,” *International Journal of Electrical and Computer Engineering*, Vol. 9, No. 2, 2019, pp. 1313-1320.

24. K. VaradaRajkumar, AdimulamYesubabu and K. Subrahmanyam, “Fuzzy clustering and Fuzzy C-Means partition cluster analysis and validation studies on a subset of CiteScore dataset,” *International Journal of Electrical and Computer Engineering*, Vol. 9, No. 4, 2019, pp. 2760-2770.

25. S. P. Potharaju, M. Sreedevi and S. S. Amiripalli. “An Ensemble Feature Selection Framework of Sonar Targets Using Symmetrical Uncertainty and Multi-Layer Perceptron (SU-MLP),” *In Cognitive Informatics and Soft Computing* 2019, pp. 247-256. Springer, Singapore.

26. S. P. Potharaju and M. Sreedevi. “Distributed feature selection (DFS) strategy for microarray gene expression data to improve the classification performance,” *Clinical Epidemiology and Global Health*, Vol. 7, No. 2, 2019, pp. 171-176.

27. A. Manikandan, N. Danapaquiame, R. Gayathri, E. Kodhai and J. Amudhavel. “A Novel Clustering Algorithm for Big Data: K-Means-Fuzzy C Means,” *BIOSCIENCE BIOTECHNOLOGY RESEARCH COMMUNICATIONS*, Vol. 11, No. 1, 2018, pp. 85-93.

28. F. Khan, “An initial seed selection algorithm for k-means clustering of geo-referenced data to improve replicability of cluster assignments for mapping application,” *Appl. Soft Comput*. Vol. 12, No. 11, 2012, pp. 3698–3700.

29. H. Li, H. He, and Y. Wen, “Dynamic particle swarm optimization and k-means clus- tering algorithm for image segmentation,” *Opt.-Int. J. Light Electron Opt*. Vol. 126, No. 24, 2015, pp. 4817–4822.

30. J.Z. Huang, M.K. Ng, H. Rong, and Z. Li, “Automated variable weighting in k-means type clustering,” *IEEE Trans. Pattern Anal. Mach. Intell*. Vol. 27, No. 5, 2005, pp. 657–668.

31. J. Chen, K. Li, Z. Tang, K. Bilal, S. Yu, C. Weng, and K. Li, “A parallel random forest algorithm for big data in a spark cloud computing environment,” *IEEE Trans. Parallel Distrib. Syst*. Vol. 1, 2017, pp. 1.

32. R. Mehmood, G. Zhang, R. Bie, H. Dawood, and H. Ahmad, “Clustering by fast search and find of density peaks via heat diffusion,” *Neurocomputing* Vol. 208, 2016, pp. 210–217.

DOI: 10.6688/JISE.20220X_38(X).00XX

33. A. Frank, and A. Asuncion, “Uci machine learning repository [http://archive.ics.uci. edu/ml],” *irvine, ca: university of california, Sch. Inf. Comput. Sci*. Vol. 213, 2010.

34. [34] K. R. Rao and B. M. Josephine, “Exploring the Impact of Optimal Clusters on Cluster Purity,” *In 2018 3rd International Conference on Communication and Electronics Systems (ICCES)* 2018, Oct, pp. 754-757. IEEE

35. S. K. Satapathy, S. Mishra, P. K. Mallick, L. Badiginchala, R. R. Gudur, and S. C. Guttha, “Classification of Features for detecting Phishing Web Sites based on Machine Learning Techniques,” *International Journal of Innovative Technology and Exploring Engineering,* Vol8, 2018, pp. 425-430.

36. K. Lavanya, L. S. S. Reddy, and B. E. Reddy, “Distributed based serial regression multiple imputation for high dimensional multivariate data in multicore environment of cloud,” *International Journal of Ambient Computing and Intelligence (IJACI),* Vol. 10, No. 2, 2018, pp. 63-79.

37. N. Danapaquiame, V. Balaji, R. Gayathri, E. Kodhai, and G. Sambasivam, “Frequent Item set Using Abundant Data on Hadoop Clusters in Big Data,” *BIOSCIENCE BIOTECHNOLOGY RESEARCH COMMUNICATIONS*, Vol. 11, No. 1, 2018, pp. 104-112.

38. N. Alange, and A. Mathur, “Small Sized File Storage Problems in Hadoop Distributed File System,” In *2019 International Conference on Smart Systems and Inventive Technology (ICSSIT)* 2019. Nov, pp. 1202-1206. IEEE.

39. D. K. Anguraj, and S. Smys, “Trust-based intrusion detection and clustering approach for wireless body area networks,” *Wireless Personal Communications*, Vol. 104, No. 1, 2019, pp. 1-20.

40. C. Banchhor, and N. Srinivasu, “Holoentropy based Correlative Naive Bayes classifier and Map Reduce model for classifying the big data,” *Evolutionary Intelligence*, pp. 1-14

41. C. Z. Basha, G. K. J. Simha, and Y. V. Krishna, “An efficient and robust fracture detection in femur bones,” *International Journal of Innovative Technology and Exploring Engineering*, Vol. 9, No. 1, 2019, pp. 1954-1957

42. M. Buvanesvari, J. Uthayakumar and J. Amudhavel, “Fuzzy based clustering to maximize network lifetime in wireless mobile sensor networks”.

43. Y. Mallikarjuna Rao, M. V. Subramanyam, and K. Satya Prasad, “Cluster-based mobility management algorithms for wireless mesh networks,” *International Journal of Communication Systems*, Vol. 31, No. 11, 2018, pp. e3595.

44. A. K. Nikhath, and K. Subrahmanyam, “Feature selection, optimization and clustering strategies of text documents,” *International Journal of Electrical & Computer Engineering (2088-8708),* Vol. 9, No. 2, 2019.

45. A. K. Nikhath, and K. Subrahmanyam, “Hybrid Approach to Explore Efficient Document Clustering Using Multi Objective Attributes,” *Journal of Computational and Theoretical Nanoscience,* Vol. 16, No. 5-6, 2019, pp. 2204-2209.

46. L. Hu, Y. Yang, Z. Tang, Y. He and X. Luo, "FCAN-MOPSO: An Improved Fuzzy-Based Graph Clustering Algorithm for Complex Networks with Multi Objective Particle Swarm Optimization," *In IEEE Transactions on Fuzzy Systems*, Vol. 31, No. 10, Oct. 2023, pp. 3470-3484, doi: 10.1109/TFUZZ.2023.3259726.

DOI: 10.6688/JISE.20220X_38(X).00XX

47. Q. V. Pham, S. Mirjalili, N. Kumar, M. Alazab, and W. J. Hwang, "Whale optimization algorithm with applications to resource allocation in wireless networks," *IEEE Transactions on Vehicular Technology,* Vol. 69, No. 4, 2020, pp. 4285-4297.

48. M. Braik, M. A. Awadallah, M. A. Al-Betar, Z. A. A. Alyasseri, A. Sheta, and S. Mirjalili, "Hybrid whale optimization algorithm for enhancing K-means clustering technique," *In Handbook of Whale Optimization Algorithm*, 2024, pp. 387-409. Academic Press.

49. M. Premkumar, G. Sinha, M. D. Ramasamy, S. Sahu, C. B. Subramanyam, R. Sowmya, and B. Derebew, "Augmented weighted K-means grey wolf optimizer: An enhanced metaheuristic algorithm for data clustering problems," *Scientific reports*, Vol. 14, No. 1, 2024, pp. 5434.