Optimizing clustering algorithm using Lévy Flight Empowered Whale Optimization Algorithm and MapReduce

Harsh D. Dwivedi, Pranav Saxena, Shashank Tripathi, Ashish Tripathi*

Abstract

In the era of Web 2.0, when the amount of data is increasing at high rate, effective data-analytics strategies need to be used in order to use this large pool of data for our advantage. The most popular method used for data-analysis is the Data clustering. Generally, algorithm applied for data-clustering is the K-Means algorithm. The algorithm positions the centroid of clusters using greedy approach, which lead to missing of global optima. To solve this problem, nature-inspired algorithm called Lévy Flight Empowered Whale Optimization Algorithm (LFEWOA) is introduced which hybridizes the hunting habits of Humpback Whale with Lévy flight steps which works to achieve global optima and hence a better clustering result. The proposed algorithm LFEWOA is verified on the seven UCI benchmark datasets and are mathched with already present algorithms namely, K-Means, Bat Algorithm, Gravitation Search Algorithm, Particle Swarm Optimization, and Grey-wolf Optimization in terms of intra-cluster distance. Nature-inspired algorithms fail on large dataset, as the time it might take would be large. For solving it, the LFEWOA algorithm is adapted in the MapReduce model of Apache-Hadoop framework and is titled MR-LFEWOA. Further, the clustering efficiency of MR-LFEWOA is weighed up on the basis of intra-cluster distance with popular Map-Reduce based algorithms namely, Parallel K-Means Algorithm, Parallel K-PSO Algorithm, Dynamic Frequency based parallel K-Bat (DFBPKB) Algorithm. The experimental results infer that proposed algorithm is a good choice for efficient large-scale data clustering.

Keywords: Clustering algorithm, MapReduce, Lévy Flight, Enhanced Whale Optimization Algorithm

1. INTRODUCTION

Clustering algorithm comes under unsupervised-learning category and is the most used technique in data-analytics. It has its applications in computer vision, information retrieval system, data mining [1], image-segmentation [3], anomaly detection [2]. Since the discovery of its application, several algorithms have been suggested to upsurge the efficacy of data clustering. One of the simplest and popular algorithms is K-Means algorithm [4]. But the K-Means algorithm is a greedy approach and is affected by the initial locations of cluster centroids and because of this the algorithm gets trapped in local-optima [5].

To solve the problem of local-optima, many nature-inspired meta-heuristic optimization algorithms have been developed and are integrated with clustering algorithm. Maulik et al. [16] first suggested a method to use nature-inspired meta-heuristic algorithms to find the position of centroid such that the intra-cluster distance is optimized. Hatamlou et al. [17] developed a clustering method which used K-Means algorithm to initialize the position of cluster-centroid and then Gravitational Search Algorithm was used to optimize the repositioning of the cluster-centroids. A bat algorithm based clustering method was proposed by Sharma et al. [18]. Pandey et al. [22] proposed an algorithm for clustering twitter data which was used for tweet's sentiment analysis using hybrid cuckoo search algorithm. Cura et el. [23] proposed a Particle Swarm Optimization based clustering method which is used to optimize the clustering of data fetched by sensors in real time. The mentioned methods work good with small datasets. However, with a big dataset, the algorithms may fail due to large time complexity and space complexity [15]. Therefore, Apache-Hadoop needs to be used which is an efficient tool for parallelization. The tool uses MapReduce [24] which is a novel approach to parallelize tasks on DataNodes [25].

Apache-Hadoop is an open-source state-of-art tool used for handling extremely large datasets, and to use the strengths of parallel computing. Hadoop makes use of its own file-system known as Hadoop Distributed File System (HDFS) [25] which has a master/slave architecture, in which there exist a master node called NameNode, and the other machines of cluster are called as DataNodes. Hadoop uses a programming model

called as MapReduce [24] for parallel processing. The model has a Map function and a Reduce function. In Map function, the (key, value) pairs are processed to produce intermediate (key, value) pairs. Reduce function combines all intermediate values of the intermediate keys generated by the Map function. MapReduce has solved many problems like count of URL access frequency, distributed grep, term-vector per host, distributed sort and inverted index. Thus, the research on meta-heuristic algorithms have started to explore the possibility of applying MapReduce in the meta-heuristic algorithm. A MapReduce-based Artificial Bee Colony algorithm (MR-ABC) was developed by Banharnsakun et al. [26] for clustering massive datasets. Wang et al. [18] proposed MapReduce-based K-PSO clustering method for clustering massive datasets. Dynamic Frequency based K-Bat algorithm (DFBPKBA) was proposed by Tripathi et al. [28] in which the large datasets was handled by MapReduce and dynamic changes to frequency of bat were done to improve the clustering process. Tripathi et al. [15] has proposed cluster algorithm which uses the Enhanced Grey Wolf Optimizer and MapReduce, in which the authors have introduced a novel approach which uses the strengths of Grey Wolf Optimizer [20] combined with the concepts of Lévy Flights [12] and Binomial Crossover, and has used MapReduce to cluster data of large dataset.

Whale Optimization Algorithm (WOA) [7] is a meta-heuristic optimization algorithm which is inspired by the hunting nature of a humpback whale. It's result have shown that it is better than many popular meta-heuristic algorithms namely, Fast Evolutionary Programing, Differential Evolution, Gravitational Search Algorithm, Particle Swarm Optimization. Mefarja et al. [27] have proposed a hybrid version of WOA which is combined with Simulated Annealing to enhance the solution of each iteration of WOA. Ling et al. [28] have proposed Lévy flight trajectory-based whale optimization algorithm (LWOA), in which Lévy Flights are used to help WOA against premature convergence. Khalil et al. [19] have used MapReduce along with WOA to tackle large datasets.

Despite of its wide applications, the WOA algorithm has relaxed convergence speed and there exist a risk for achieving local optima [9]. To improve the clustering process of the WOA, the authors have proposed a novel approach using Lévy Flights along with WOA called Lévy Flight Empowered Whale Optimization Algorithm (LFEWOA) to prevent the achievement of local-optima, and further the authors have proposed MapReduce-based method called, MapReduce-based Lévy Flight Empowered Whale Optimization Algorithm (MR-LFEWOA), to tackle the clustering of large datasets. The proposed algorithm LFEWOA is matched with the already present algorithms namely, K-Means, Bat Algorithm, Gravitation Search Algorithm, Particle Swarm Optimization, and Grey-wolf Optimization by testing each on seven UCI benchmark datasets on the basis of intra-cluster distance. The benchmark datasets are Iris, Seeds, Glass, Cancer, Balance, Haberman, and Wine. Clustering efficiency of MR-LFEWOA is weighed up on the basis of intra-cluster distance with popular Map-Reduce based algorithms namely, Parallel K-Means Algorithm, Parallel K-PSO Algorithm, Dynamic Frequency based parallel K-Bat (DFBPKB) Algorithm by testing each on large datasets namely, Replicated Iris, Replicated CMC, Replicated Wine, Replicated Vowel.

The research paper has been organized as: Section 2 discusses the basics of data-clustering and Whale Optimization Algorithm (WOA). Section 3 discusses the clustering method of the proposed method LFEWOA, and further discusses the MapReduce-based proposed method MR-LFEWOA. Section 4 presents the experimental arrangements and results. And Section 5 gives the conclusion of the paper.

2. BACKGROUND

2.1 Data Clustering Approach

Clustering algorithm clusters similar looking data-points on the basis of resemblance. Clustering of *N* data-points with *t* attributes each into *K* clusters is an iterative process and is an algorithm of unsupervised learning category. There is not training required and the data-points are clustered on the how the data-points are structured, unlike the supervised-learning technique like Regression, or Classification. The evaluation of clustering is done on the basis of the summation of the intra-cluster distance of *K* clusters. Let $Z=\{\{z_{11},z_{12},\ldots,z_{1t}\},\{z_{21},z_{22},\ldots,z_{2t}\},\ldots,\{z_{n1},z_{n2},\ldots,z_{nt}\}\}$ be set of *N* data-points with *t* attributes each, where z_{ij}

indicates the value of j^{th} attribute of i^{th} data-point. The clustering algorithm runs iteratively to find the set of cluster-centroids, $C=\{\{c_{11},c_{12},\ldots,c_{1t}\},\{c_{21},c_{22},\ldots,c_{2t}\},\ldots,\{c_{k1},c_{k2},\ldots,c_{kt}\}\}$, where c_{ij} indicates the the value of j^{th} attribute of i^{th} centroid, such that the summation of intra-cluster distance is minimum. Clustering process should at-least satisfy the below conditions:

1. No cluster should not be empty of data-point, i.e., $C_i \neq \phi, \forall i \in \{1,2,3,\ldots,K\}$
2. Each data-point belongs to one cluster
3. No data-point can't belong to more than one cluster, i.e., $C_q \cap C_r = \phi, \forall q \neq r \; and \; q,r \in \{1,2,3,\ldots,K\}$

The evaluation of quality of clustering process uses one of the prominent method of summation of Euclidian distance [6] between centroids and data-points, which is calculated as equation 1:

$$f(Z,C) = \sum_{i=1}^{K} \sum_{Z_i \in C_i} (d(Z_i, C_i))^2 \quad (1)$$

Where $d(Z_i, C_i)$ is the Euclidian distance between C_i centroid and the Z_i data-point that comes under the cluster of C_i centroid, and $C_i=\{c_{i1},c_{i2},\ldots,c_{it}\}$. The Euclidian distance is calculated according to equation 2.

$$d(Z_i, C_i) = \sqrt{\sum_{j=1}^{t} (z_{ij} - c_{ij})^2} \quad (2)$$

2.2 Whale Optimization Algorithm

Whale Optimization Algorithm [7] is inspired by the hunting nature of humpback whales. The algorithm works in exploration phase and exploitation phase. The Exploitation phase is further divided into two different strategies, namely, Shrinking encircling mechanism and Spiral update mechanism. In Shrinking encircling mechanism, the whale moves toward the best whale of population in circular manner. It can be mathematically defined by equation 3.

$$\vec{X}(t+1) = \overrightarrow{X^*}(t) - \vec{A}.\vec{D} \quad (3)$$

$$\vec{D} = \left|C.\overrightarrow{X^*}(t) - \vec{X}(t)\right| \quad (4)$$

$$\vec{A} = 2.\vec{a}.\vec{r}.-\vec{a} \quad (5)$$

$$\vec{C} = 2.\vec{r} \quad (6)$$

where $\overrightarrow{X^*}(t)$ is the position of best whale in previous iteration, $\vec{X}(t)$ is a whale position and 't' indicates the current iteration. The mathematical Shrinking encircling behaviour is achieved by linearly decreasing the value of $|\overrightarrow{a}|$ in equation 4 from 2 to 0 over the course of iterations and $|\vec{r}|$ is a random number uniformly distributed in the range of [0,1]. In Spiral update mechanism, the distance between best whale position ($\overrightarrow{X^*}(t)$) and the position of whale ($\vec{X}(t)$). Helix-movement of the whale is then mathematically mimicked by the equation 7.

$$\vec{X}(t+1) = \overrightarrow{D'}.e^{bl}.\cos(2\pi l) + \overrightarrow{X^*}(t) \quad (7)$$

Where $\overrightarrow{D'} = |\overrightarrow{X^*}(\mathrm{t}) - \vec{X}(t)|$ is the distance between the whale and the best whale. b is a constant for defining the shape of logarithmic spiral, l is random number between [-1,1]. Both the strategies have 50% probability of being used. The mathematical model that is used by the humpback whale is given it equation 8.

$$\vec{X}(t+1) = \begin{cases} \overrightarrow{X^*}(t) - A.\vec{D} & , \ p < 0.5 \\ \overrightarrow{D'}.e^{bl}.\cos(2\pi l) + \overrightarrow{X^*}(t), & p \geq 0.5 \end{cases} \quad (8)$$

Where, p is a random number between [0,1].

In exploration phase, the whale chooses a random position $\overrightarrow{X_{rand}}$, and move towards it. This phase is triggered when the value of $|\vec{A}| > 1$ in equation 5 of Shrinking encircling mechanism.

3. PROPOSED METHOD

Whale Optimization Algorithm doesn't work properly for large datasets, as it is an iterative process and takes a lot of time to compute. The authors have proposed a MapReduce-based method, called MapReduce-based Lévy Flight Empowered Whale Optimization Algorithm (MR-LFEWOA). MR-LFEWOA uses the advantages of Lévy Flight Empowered Whale Optimization Algorithm (LFEWOA), for having efficacious clustering of data-points, which is a novel variant of Whale Optimization Algorithm. The section discusses the details of Enhanced Whale Optimization Algorithm (LFEWOA), and then discusses the MapReduce variant of LFEWOA.

3.1 Lévy Flight Empowered Whale Optimization Algorithm (LFEWOA)

The equilibrium between exploration and exploitation decides the success of any nature-inspired meta-heuristic algorithm [8]. The WOA algorithm has a risk of achieving local optima and slow convergence speed [9]. These problems could be solved to a large extent by improving the reposition of whales using Lévy Flights [10].

3.1.1 Repositioning using Lévy Flights

The convergence speed is slow in WOA as the whale's repositioning is depended only on the position of best whale. To increase the efficiency of exploitation process, the proposed LFEWOA uses the strengths of Lévy Flights to reposition each whale. Lévy flight uses Lévy distribution [11] to define steps of random lengths. The authors have used the Magenta algorithm [10] to generate such steps. The calculation of step length(z) defined by Magenta algorithm is given in equation 9.

$$z = \left[\frac{r}{|s|^{1/\beta}}\right] \quad (9)$$

where $\beta>0$ and $\beta<=2$, is lévy index and r and s are variables following normal distribution of $N(0,\sigma_r^2)$ and $N(0,\sigma_s^2)$, respectively. The σ_r is calculated by equation (10) while σ_s is always 1.

$$\sigma_r = \left[\frac{\Gamma(1+\beta)\sin(\pi\beta/2)}{\beta\Gamma\left[\frac{1+\beta}{2}\right]2^{[(\beta-1)/2]}}\right]^{1/\beta} \quad (10)$$

where, Γ() is called Gamma function [13].

In the proposed method, the whale takes lévy flight for the search of the prey and updates its position using equation (11).

$$\vec{X}(t+1) = \vec{X}(t) + z * \left|\overrightarrow{X^*}(t) - \vec{X}(t)\right| \quad (11)$$

Where $\vec{X}(t)$ represent the position of whale for i^{th} iteration and $\overrightarrow{X^*}(t)$ represents the best position of whale for i^{th} iteration.

3.1.2 LFEWOA based clustering

The clustering problem can be solved by applying the LFEWOA algorithm. In the proposed LFEWOA based clustering, each whale's position $\vec{X}$ is a set a set of cluster centroid. $\vec{X}$=={{c_{11},c_{12},…,c_{1t}},{c_{21},c_{22},…,c_{2t}},…,{c_{k1},c_{k2},…,c_{kt}}}, where c_{ij} indicates the the value of j^{th} attribute of i^{th} centroid. The clustering is evaluated in terms of summation intra-cluster distance of each centroid. The whale whose summation of intra-cluster distance is taken as the best whale. The summation of intra-cluster distance is evaluated according to equation 1. The algorithm for the proposed method is given in Algorithm 1. The time-complexity of the algorithm is O(P*N*K*t), where P is the population of whales used, N is the number of data-points, K is the number of clusters we are having study on, t is the number of attributes each data-point has.

whale population X_i (i = {1, 2, ..., P}) initialization
$\overrightarrow{X^*}$ =the best whale of 0^{th} iteration
while(t< maximum iteration possible)
    for each whale
        Update a, A, C, l, and p
        if1(p<0.5)
            if2($|\vec{A}|$<1)
                Position of the current whale is updated by the Equation 2
            else2
                Select a random whale($\overrightarrow{X_{rand}}$)
                Position of the current whale is updated by the Equation 8
            end if2
        else1
            Position of the current whale is updated by the Equation 5
        end if1
        Re-update the position of whale according to Equation 11
    end for
    if available better solution Update $\overrightarrow{X^*}$
    t=t+1
end while
return $\overrightarrow{X^*}$

Algorithm 1: Lévy Flight Empowered Whale Optimization Algorithm based clustering

3.3 Parallelization of LFEWOA using MapReduce architecture

The time complexity O(P*N*K*t) shows that it is directly proportional to the size of collection of data-points. If the data-points are in large number, it would lead to large time-run of the algorithm. In real-time, the sensors fetch trillions of data-points, we require a parallel version of the proposed algorithm. The authors hence propose MapReduce-based Lévy Flight Empowered Whale Optimization Algorithm (MR-LFEWOA) to make the algorithm work on parallel architecture. MR-LFEWOA works in two phases: LFEWOA-Map and LFEWOA-Reduce. The proposed method initially divides the collection of data-points in smaller collection and distribute them uniformly along Hadoop Distributed File System (HDFS) [14] datanodes. LFEWOA-Map processes data-point and finds cluster of some whale which would include the given data-point and calculate the Euclidian distance between the cluster's centroid and the given data-point. The output of this phase would be {key:(whaleID,centID),value:minDistance}. The pseudo-code for LFEWOA-Map is given in Algorithm 2. LFEWOA-Reduce phase merges the outputs given by Map phase, and then calculate the intra-cluster distance of each key given by Map phase. The output of this phase is in the form {key:(whaleID,centID),value:intra-cluster-distance}.The pseudo-code of this phase is given in

Algorithm 3. The whales' position is then update according to Algorithm 1 which uses the output given by LEEWOA-Reduce phase. The complete architecture of MR-LFEWOA is briefed in Figure 1.

```
for each whale in population
        whaleId=ID_of_whale
        centroidArray=retrieve_centroid_array_of_given_whale(whaleID)
        minDistance=INT_MAX
        for each centroid in centroidArray
            distance=Euclidian_distance(data-point,centroid)
            if(distance<minDistance)
                centId=ID_of_centroid
                minDistance=distance
            end if
        end for
        key=(whaleID,centID)
end for
return {key:(whaleID,centroidID),value:minDistance}
```

Algorithm 2: MR-LFEWOA-Map

```
Input:{key:(whaleID,centID),value-list:distances}
intra-cluster-distance=0
for each distance in distances
            intra-cluster-distance+=distance
end for
return {key:(whaleID,centID),value:intra-cluster-distance}
```

Algorithm 3: MR-LFEWOA-Reduce

4. EXPERIMENTAL RESULTS

Evaluation of the proposed method is done in two phases. In the initial phase, Lévy Flight Empowered Whale Optimization Algorithm (LFEWOA) is evaluated by being weighed up with five prominent meta-heuristic algorithms namely, PSO, K-Means, GSA, BA, and GWO, by being run on seven UCI benchmark datasets namely, Iris, Seeds, Glass, Cancer, Balance, Haberman, and Wine, in terms of intra-cluster distance. In the next phase, the efficacy of MapReduce-based Lévy Flight Empowered Whale Optimization Algorithm (MR-LFEWOA) is measured in terms of F-Measure and weighed up with four state-of-art meta-heuristic algorithms namely, parallel K-means (PK-Means), parallel K-PSO [18], MR-ABC [26] and DFBPKBA [28]. The number of nodes in Hadoop cluster were increased for understanding the behaviour of MR-LFEWOA.

4.1 Performance analysis of LFEWOA based clustering

The proposed method LFEWOA is weighed up with five prominent meta-heuristic algorithms namely, PSO, K-Means, GSA, BA, and GWO based on intra-cluster distance found on running on seven UCI benchmark datasets namely, Iris, Seeds, Glass, Cancer, Balance, Haberman, and Wine. The experiment is done using single machine with the specification as 2.5 GHz Intel core i5 processer, 8 GB of RAM and 1TB hard-disk. The seven benchmark datasets considered are summarized on Table 1. Table 2 describes the setting of experimentation for each meta-heuristic algorithm. Table 3 shows the intra-cluster distance found on running each meta-heuristic algorithm on each dataset. It can be observed from Table 3 that Lévy

Flight Empowered Whale Optimization Algorithm (LFEWOA) outputs a lesser intra-cluster distance as weighed up to other five nature-inspired algorithms for each dataset.

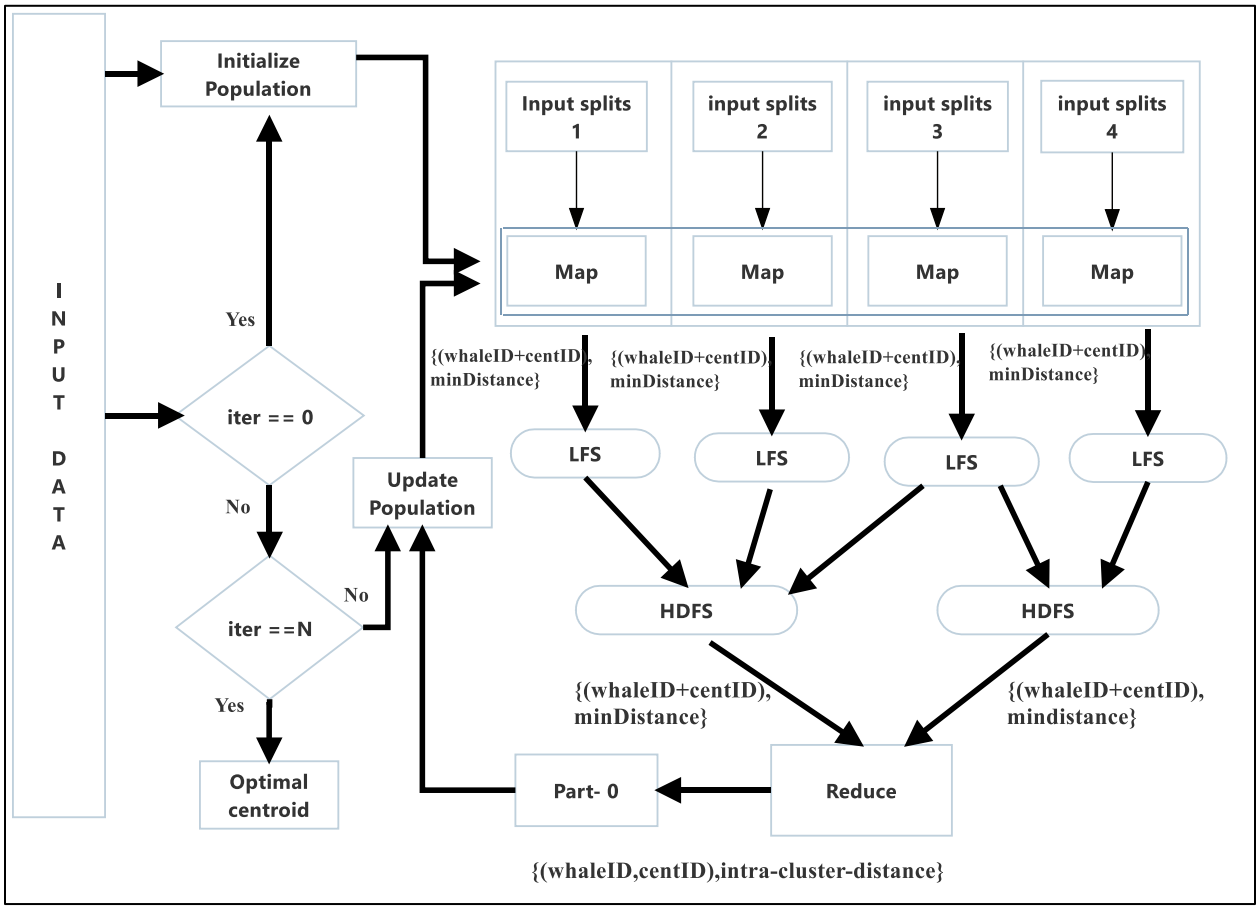

Figure 1: Architecture of MR-LFEWOA

Table 1: Dataset description

Dataset	Number of data-points	Number of attributes	Number of clusters
Haberman	306	3	2
Cancer	638	9	2
Iris	150	4	3
Glass	214	9	6
Wine	178	13	3
Balance	625	4	3
Seeds	210	7	3

Table 2: Setting parameters of each meta-heuristic algorithm for experimentation

Parameter name	K-Means	BA	PSO	GWO	GSA	LFEWOA
Population size	-	40	40	40	40	40
Cognitive Constant (c1)	-	-	1	-	-	-
r0	-	9	-	-	-	-
fmin	-	0	-	-	-	-

fmax	-	2	-	-	-	-
Inertial Constant (w)	-	-	0.5	-	-	-
Social Constant (c2)	-	-	1	-	-	-
Gamma ()	-	9	-	-	-	-
Alpha (α)	20	2	-	2	9	-
G-constant (G0)	-	-	-	-	20	-
Number of iterations	500	500	500	500	500	500

Table 3: Mean intra-cluster distance for 30 iteration of each algorithm on each dataset

Dataset	K-Means	BA	PSO	GWO	GSA	LFEWOA
Iris	97.34084	96.65552	96.78998	96.65826	96.65548	**96.65548**
Seeds	587.31957	311.79816	312.68370	311.88200	311.79804	**311.79804**
Haberman	30507.0207	2566.98889	2566.99548	2567.02562	2566.98989	**2566.98889**
Glass	292.75724	243.70331	238.51144	265.81420	286.11855	**214.44399**
Wine	2370689.68700	16371.05448	16298.98906	16307.09242	17038.59226	**16292.18465**
Cancer	19323.1738	2964.38718	2969.23958	2964.390179	2970.17834	**2964.38697**
Balance	3472.32142	1424.04307	1423.96787	1423.82106	1423.82042	**1423.82040**

To show that the converge speed of the LFEWOA is decreased, a graph for dataset Wine and Seeds is given in Figure 2 and 3, in which the intra-cluster distance of each algorithm is plotted against different number of iterations. The Y-axis of the graph represent the intra-cluster distance, and the X-axis represents the number of iterations. The plot of LFEWOA shows that the value of intra-cluster distance instantly decreases after a point, and then the plot is almost a straight line parallel to X-axis for increased number of iterations. The plot of algorithms decreases gradually. This shows that the LFEWOA achieves optimum intra-cluster distance in lesser iterations.

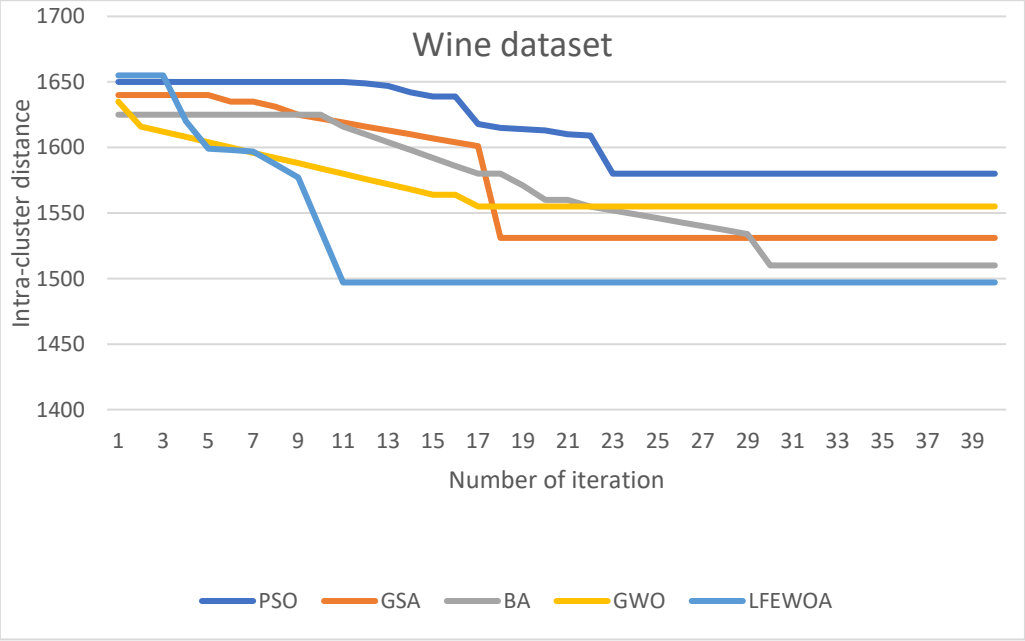

Figure 2: Convergence graph of Wine

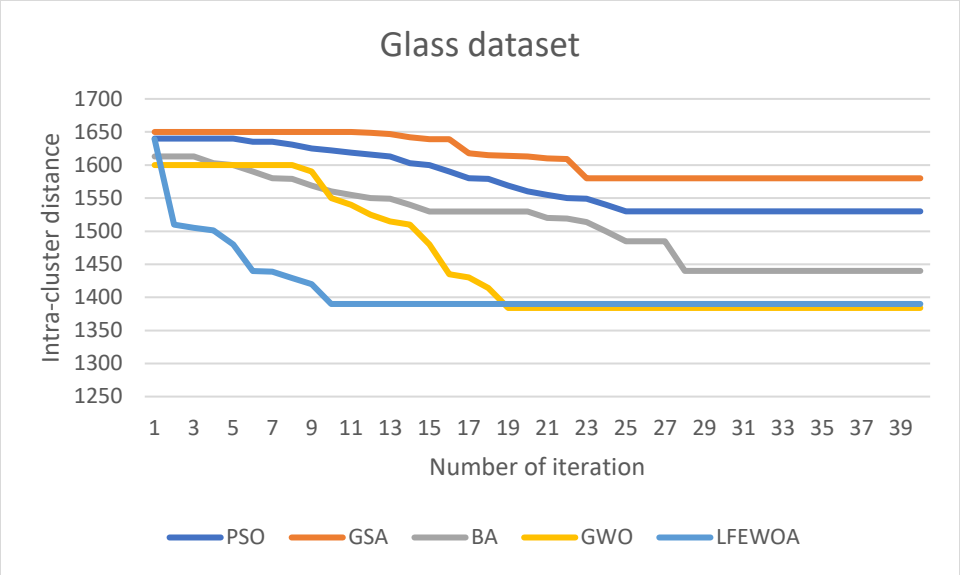

Figure 3: Convergence graph of Glass

4.2 Performance analysis of MapReduce-based LFEWOA (MR-LFEWOA)

Section 4.1 has shown that the LFEWOA has given better results than the state-of-art meta-heuristic algorithm. Hence, LFEWOA has been parallelized over Hadoop cluster, is called as MapReduce-based Lévy Flight Empowered Whale Optimization Algorithm. The MR-LFEWOA is analyzed using four researcher-made large datasets, which have been developed by replicating the original data-points 10^7 times. Table 4 describes the dataset used. For experimentation, a Hadoop cluster was utilized with the maximum capacity of 5 nodes. Each node's specifications are: Intel core i5 processor, 2.5GHz frequency, 8GB RAM, 1TB secondary-storage, openjdk version "1.8.0_191", Apache Hadoop version 2.6.0. MR-LFEWOA was run on each researcher-developed dataset, and F-measure and computation time was calculated. The results of the experiments are given in table 5.

Table 4: Description of large datasets

Dataset	Number of data-points	Number of attributes	Number of clusters
Reproduced Vowel	1,025,010	10	10
Reproduced CMC	10,000,197	9	3
Reproduced Iris	10,000,050	7	3
Reproduced Wine	5,000,000	18	2

Table 5: F-measure and computation time for each algorithm on each dataset on 30 iterations

Dataset	Criteria	Parallel K-PSO	DFBPKBA	Parallel K-Means	MR-ABC	MR-LFEWOA
Reproduced CMC	Computation time	10.33E + 04	10.34E + 04	**8.24E + 04**	10.33E + 04	10.32E + 04
	F-Measure	0.324	0.378	0.298	0.387	**0.391**
Reproduced Vowel	Computation time	13.22E + 04	13.23E + 04	**10.50E + 04**	12.21E + 04	13.21E + 04
	F-Measure	0.627	0.622	0.586	0.634	**0.635**

Reproduced Iris	Computation time	9.23E + 04	9.24E + 04	**8.05E + 04**	9.26E + 04	9.22E + 02
	F-Measure	0.785	0.790	0.667	0.842	**0.846**
Reproduced Wine	Computation time	13.22E + 04	13.23E + 04	**11.20E + 04**	12.21E + 04	13.21E + 04
	F-Measure	0.627	0.622	0.586	0.634	**0.635**

Table 5 shows that MR-LFEWOA gives better results than the state-of-art MapReduce-based meta-heuristic algorithms, while the parallel-KMeans gives the poorest result in the selected algorithms. But, parallel KMeans has the minimum computation-time, weighed up to other selected algorithms

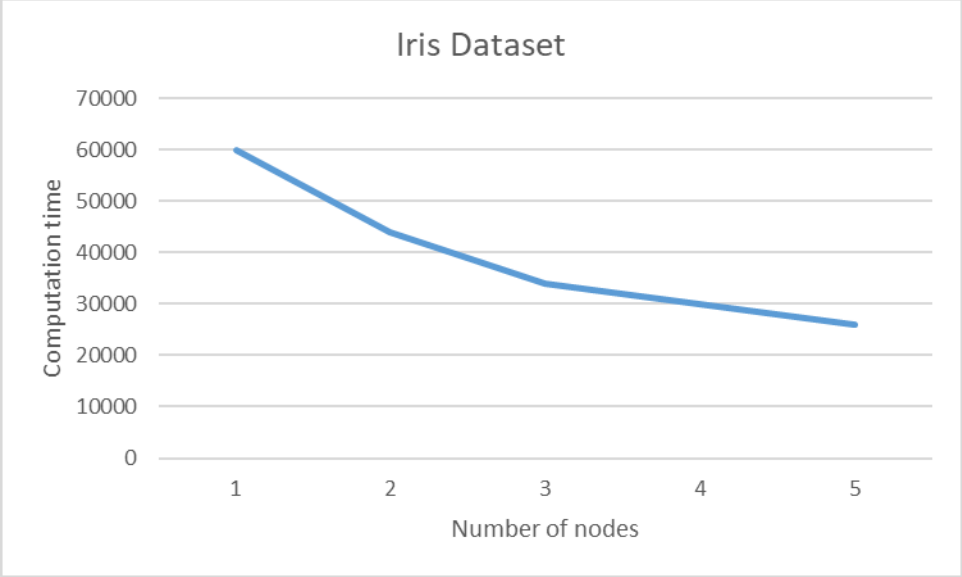

Figure 4: Speedup graph of Replicated-Iris dataset

Speedup-performance of MR-LFEWOA is studied on CMC dataset and Iris datasets. The calculation of speedup-performance is done according to equation 12.

$$S = T_{base}/T_N \quad (12)$$

Where T_{base} is the time taken by algorithm on one machine, and T_N is the time taken by algorithm on a parallel-processing architecture with N machines. So, the speedup-performance is the ratio between time taken by one machine and time taken by N machines working in parallel to run an algorithm. Figure 4 is the graph of MR-LFEWOA running on Iris dataset, which is a plot between time and number of nodes. Figure 5 is the graph of MR-LFEWOA running on CMC dataset, which is a plot between time and number of nodes. Both figure 4 and 5 shows that the time taken to run the algorithm decreases with an increase in the number of nodes in Hadoop cluster. The method has achieved speedup measure of 4.4576 and 4.7543 on Iris and CMC datasets respectively when 5 nodes are taken in Hadoop cluster. This shows that MR-LFEWO can be used to cluster large datasets.

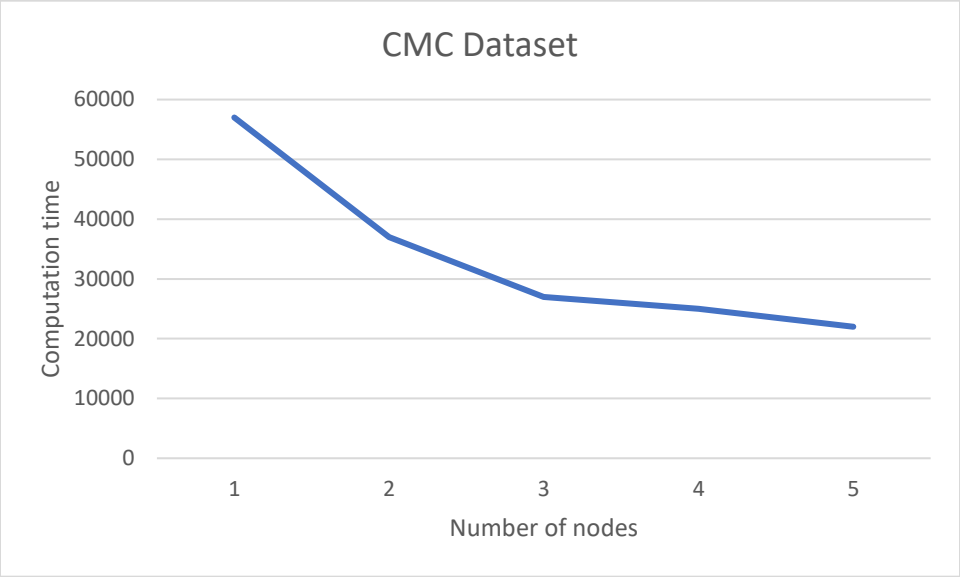

Figure 5: Speedup graph of Replicated-CMC dataset

5. CONCLUSION

In data-analytics, clustering algorithms are most popular tool. K-Means is a prominent and easy to implement clustering algorithm present. The algorithm though has some strengths, but is a greedy approach algorithm and hence has the risk of achieving local-optima. Therefore, many state-of-art nature-inspired meta-heuristic algorithms are used for optimizing the task of repositioning of centroids. Whale Optimization Algorithm (WOA) is newest meta-heuristic algorithm in research and has proved to be superior than other result-proven meta-heuristic algorithms. However, it still has risk of achieving local-optima and has slow convergence speed [9], and takes a lot of computational time for clustering large dataset. Hence, the authors present a MapReduce-based clustering method called MapReduce-based Lévy Flight Empowered Whale Optimization Algorithm (MR-LFEWOA). MR-LFEWOA is based on a novel variant of WOA called Lévy Flight Empowered Whale Optimization Algorithm (LFEWOA), which uses the strengths of WOA and the concept of Lévy Flight. LFEWOA is weighed up with five prominent meta-heuristic algorithms namely, PSO, K-Means, BA, GSA, GWO, on the basis of intra-cluster distance found on running on seven UCI benchmark datasets namely, Iris, Seeds, Glass, Cancer, Balance, Haberman, and Wine. The comparison showed that the intra-cluster-distance found by LFEWOA is minimum in the intra-cluster distances found by other algorithms. To deal with large dataset, MapReduce-based method MR-LFEWOA is proposed. MR-LFEWOA is weighted up with other MapReduce-based nature-inspired algorithms namely, Parallel K-Means Algorithm, Parallel K-PSO Algorithm, Dynamic Frequency based parallel K-Bat (DFBPKB) Algorithm by testing each on large datasets namely, Replicated Iris, Replicated CMC, Replicated Wine, Replicated Vowel on the basis of computation time and F-Measure. The comparison showed that MR-LFEWOA took lesser computational time than other algorithms, and has more F-Measure value than other algorithms. The comparison shows that the MR-LFEWOA is better clustering method than other MapReduce-based natured-inspired algorithms in term of F-Measure and computation time. MR-LFEWOA's Speedup performance is calculated for Replicated Iris and Replicated CMC, whose value came out to be 4.4576 and 4.7532 respectively when number of nodes in Hadoop cluster were five. Speedup performance shows that the MR-LFEWOA is apt for clustering large-dataset. Therefore, it is deduced that MR-LFEWOA is a competitive method to analyze extremely big datasets.

In future, the LFEWOA shall be implemented over Apache-Spark to further reduce the computation time. Furthermore, the LFEWOA shall be applied on some large-scale datasets which require clustering like, twitter data, satellite image data.

References

[1] Fayyad, Usama M., Andreas Wierse, and Georges G. Grinstein, eds. Information visualization in data mining and knowledge discovery. Morgan Kaufmann, 2002.

[2] Friedman, Menahem, Mark Last, Yaniv Makover, and Abraham Kandel. "Anomaly detection in web documents using crisp and fuzzy-based cosine clustering methodology." Information sciences 177, no. 2 (2007): 467-475.

[3] Liao, Liang, Tusheng Lin, and Bi Li. "MRI brain image segmentation and bias field correction based on fast spatially constrained kernel clustering approach." Pattern Recognition Letters 29, no. 10 (2008): 1580-1588.

[4] Hartigan, John A., and Manchek A. Wong. "Algorithm AS 136: A k-means clustering algorithm." Journal of the Royal Statistical Society. Series C (Applied Statistics) 28, no. 1 (1979): 100-108.

[5] Kao, Yi-Tung, Erwie Zahara, and I-Wei Kao. "A hybridized approach to data clustering." Expert Systems with Applications 34, no. 3 (2008): 1754-1762.

[6] Yang, Shuyuan, RuiXia Wu, Min Wang, and Licheng Jiao. "Evolutionary clustering based vector quantization and SPIHT coding for image compression." Pattern Recognition Letters 31, no. 13 (2010): 1773-1780.

[7] Mirjalili, Seyedali, and Andrew Lewis. "The whale optimization algorithm." Advances in Engineering Software 95 (2016): 51-67.

[8] Vitaliy, Feoktistov. "Differential evolution–in search of solutions." (2006).

[9] Kaveh, A., and M. Ilchi Ghazaan. "Enhanced whale optimization algorithm for sizing optimization of skeletal structures." Mechanics Based Design of Structures and Machines 45, no. 3 (2017): 345-362.

[10] Yang, Xin-She, and Suash Deb. "Eagle strategy using Lévy walk and firefly algorithms for stochastic optimization." In Nature Inspired Cooperative Strategies for Optimization (NICSO 2010), pp. 101-111. Springer, Berlin, Heidelberg, 2010.

[11] Fister Jr, Iztok, Xin-She Yang, Iztok Fister, Janez Brest, and Dušan Fister. "A brief review of nature-inspired algorithms for optimization." arXiv preprint arXiv:1307.4186 (2013).

[12] Shlesinger, Micheal F., George M. Zaslavsky, and Uriel Frisch. "Lévy flights and related topics in physics." In Levy flights and related topics in Physics, vol. 450. 1995.

[13] Stacy, Edney W. "A generalization of the gamma distribution." The Annals of mathematical statistics 33, no. 3 (1962): 1187-1192.

[14] Shvachko, Konstantin, Hairong Kuang, Sanjay Radia, and Robert Chansler. "The hadoop distributed file system." In Mass storage systems and technologies (MSST), 2010 IEEE 26th symposium on, pp. 1-10. Ieee, 2010.

[15] Tripathi, Ashish Kumar, Kapil Sharma, and Manju Bala. "A Novel Clustering Method Using Enhanced Grey Wolf Optimizer and MapReduce." Big Data Research (2018).

[16] Maulik, Ujjwal, and Sanghamitra Bandyopadhyay. "Genetic algorithm-based clustering technique." Pattern recognition 33, no. 9 (2000): 1455-1465.

[17] Hatamlou, Abdolreza, Salwani Abdullah, and Hossein Nezamabadi-Pour. "Application of gravitational search algorithm on data clustering." In International Conference on Rough Sets and Knowledge Technology, pp. 337-346. Springer, Berlin, Heidelberg, 2011.

[18] Ashish, Tripathi, Sharma Kapil, and Bala Manju. "Parallel Bat Algorithm-Based Clustering Using MapReduce." In Networking Communication and Data Knowledge Engineering, pp. 73-82. Springer, Singapore, 2018.

[19] Khalil, Yasser, Mohammad Alshayeji, and Imtiaz Ahmad. "Distributed Whale Optimization Algorithm based on MapReduce." Concurrency and Computation: Practice and Experience: e4872.

[20] Mirjalili, Seyedali, Seyed Mohammad Mirjalili, and Andrew Lewis. "Grey wolf optimizer." Advances in engineering software 69 (2014): 46-61.

[21] Yang, Xin-She. Nature-inspired metaheuristic algorithms. Luniver press, 2010.

[22] Pandey, Avinash Chandra, Dharmveer Singh Rajpoot, and Mukesh Saraswat. "Twitter sentiment analysis using hybrid cuckoo search method." Information Processing & Management 53, no. 4 (2017): 764-779.

[23] Alam, Shafiq, Gillian Dobbie, and Patricia Riddle. "Particle swarm optimization based clustering of web usage data." In Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 03, pp. 451-454. IEEE Computer Society, 2008.

[24] Dean, Jeffrey, and Sanjay Ghemawat. "MapReduce: simplified data processing on large clusters." Communications of the ACM 51, no. 1 (2008): 107-113.

[25] Shvachko, Konstantin, Hairong Kuang, Sanjay Radia, and Robert Chansler. "The hadoop distributed file system." In Mass storage systems and technologies (MSST), 2010 IEEE 26th symposium on, pp. 1-10. Ieee, 2010.

[26] Banharnsakun, Anan. "A MapReduce-based artificial bee colony for large-scale data clustering." Pattern Recognition Letters 93 (2017): 78-84.

[27] Mafarja, Majdi M., and Seyedali Mirjalili. "Hybrid Whale Optimization Algorithm with simulated annealing for feature selection." Neurocomputing 260 (2017): 302-312.

[28] Ling, Ying, Yongquan Zhou, and Qifang Luo. "Lévy Flight Trajectory-Based Whale Optimization Algorithm for Global Optimization." IEEE access 5, no. 99 (2017): 6168-6186.