Product Recommendation System Using Tournament-Selection Empowered Whale Optimization Algorithm Optimized K-Means and MapReduce

Ashish Tripathi*, Siddharth Gupta, Pranav Saxena

Abstract

In the era of Web 2.0, the data is growing on an immense scale. The E-commerce business require information retrieval tools like Recommender Systems, for assisting their costumers to select better choice for their shopping bucket. One of the prominent recommendation system approach is Collaborative filtering, that outputs best suggestion by finding similar items based on user's shopping history. K-Means prominent method is used to find clusters of similar items in Collaborative filtering. K-Means is a greedy approach and thus suffers from local minima problem. To mitigate this problem, nature-inspired algorithm named Tournament-Selection Empowered Whale Optimization Algorithm Optimized K-Means (TSWOAK) is introduced. The introduced algorithm TSWOAK, uses the hunting behaviour of Humpback Whales and Tournament-Selection concept to achieve global optima. The introduced algorithm is tested on seven benchmark UCI datasets, and the results are compared with the ones achieved by state-of-art algorithms namely, Bat algorithm, Particle Swarm Optimization, Grey-Wolf Optimization, K-Means on the basis of inter-cluster distance. The results showed that the K-Means has been optimized by the proposed method. The algorithm was then checked for its application in Recommender System. For checking its applicability, it was tested on MovieLens dataset. The results in term of Mean Absolute Error(MAE) were compared to the results of already present algorithms namely, K-Means, PCA-K-Means, ABC-KM, GAKM, PCA-GAKM, SOM, PCA-SOM, UPCC. The comparison shows that the TSWOAK has lesser MAE than other algorithms, which shows that Recommender System is a good application of TSWOAK. The problem which TSWOAK still face is that it will fail on large datasets, as the time complexity of nature-inspired algorithms is high. To solve this problem, the TSWOAK is adapted in the MapReduce model of Apache-Hadoop framework and is named MR-TSWOAK

Keywords: Clustering algorithm, MapReduce, Recommendation Systems, Tournament-Selection Whale Optimization Algorithm

1. INTRODUCTION

Recommendation Systems are one of the most important information retrieval tools used by online businesses to make the environment for customers more personal. Most promiment technique for Recommendation System is the Collaborative Filtering [1]. K-Means algorithm is one of the most used algorithms in Collaborative Filtering [4]. K-Means being a greedy approach suffers from trapping in local-optima [5].

To prevent K-Means from getting trapped into local-optima, many nature-inspired algorithms have been developed and are hybridized with the clustering algorithm. Maulik et al. [7] were first to propose methodology of using nature-inspired algorithm to optimise the positioning of centroids, in a way that the intra-cluster distance is optimized. Hatamlou et al. [6] have optimized K-Means algorithm using Gravitational Search Algorithm (GSA). The GSA was used to optimize the repositioning of cluster-centroids. Sharma et al. [8] have proposed to use bat-algorithm to optimize the positioning of cluster's centroids and have further extend their method on MapReduce. Pandey et al. [9] have proposed an algorithm

that uses Hybrid Cuckoo Search Algorithm to optimize clustering of twitter data being used in sentiment analysis of tweets. Cura et al. [10] have proposed a method to optimize the clustering of sensor's data using Particle Swarm Optimization. Though the above-mentioned methods perform good on small datasets, but fail on large datasets, as the nature-inspired algorithm-based clustering algorithms have high time complexity [5]. Most of the Recommendation Systems suffer from reduced scalability [2]. To solve these problems, Apache-Hadoop needs to be braught into application, as it is an efficient tool for parallel-computing. The tool uses MapReduce programming model [11] which helps to parallelize the tasks on DataNodes [12]

Apache-Hadoop is a state-of-art open-source tool for handling large datasets, and support parallel-computing. Hadoop uses its own file-system called as Hadoop Distributed File System (HDFS) [12]. The HDFS is a master-slave architecture, in which there exist one master node called as NameNode, which keeps the meta-data of all the slave nodes called as DataNodes. MapReduce programming model [11] is used by Hadoop for parallel-processing. The MapReduce model has a Map function and Reduce function. The Map function takes <key, value> pair as an input and outputs intermediate <key, value> pairs. The Reduce function takes all the intermediate <key, value> pairs and combine all values related to each unique key. As Hadoop simplifies parallel-computing, many meta-heuristic algorithms have been developed over Hadoop. MapReduce-based Artificial Bee Colony algorithm [13] optimised the clustering of large datasets. Dynamic Frequency based K-Bat algorithm (DFBPKBA) [14] showed that the dynamic change to frequency of bats optimised the clustering process and the MapReduce version of the proposed method was able to handle large datasets. Tripathi et al. [5] have proposed an algorithm that uses a hybrid of Grey-wolf Optimizer with Levy Flights and Crossover on MapReduce programming model.

Whale Optimization Algorithm [15] is a meta-heuristic nature inspired algorithm inspired by the hunting behaviour of Humpback whales. The algorithm has shown better results than many state of art algorithms such as Particle-swarm Optimization. Grey-wolf Optimizer, Differential Evolution, Gravitational Search Algorithm, Fast Evolutionary Programming.

Though Whale Optimization Algorithm has proved to give better results than other state of the art meta-heuristic nature inspired algorithms, but it has risk to get into local optima [16]. Whale Optimization Algorithm discard the (called as whales) with bad fitness values, and there is a chance that the whale having bad fitness value might be near global-optima [16]. To overcome this shortcoming, the authors have proposed a method that uses the concept of Tournament Selection along with WOA, called as Tournament Selection Empowered Whale Optimization Algorithm optimized K-Means (TSWOAK). To handle extremely large dataset, the authors have further proposed a MapReduce variant of TSWOAK called MapReduce-based Tournament Selection Empowered Whale Optimization Algorithm optimized K-Means (MR-TSWOAK). The proposed method TSWOAK has been tested on seven UCI datasets and the results have been compared with respect to intra-cluster distance with state of art algorithm namely. K-Means, Bat Algorithm, Gravitation Search Algorithm, Particle Swarm Optimization, and Grey-wolf Optimizer. The UCI datasets used are Iris, Seeds, Glass, Cancer, Balance, Haberman, and Wine. The MapReduce based method MR-TSWOAK was tested on large datasets namely, Replicated Iris, Replicated CMC, Replicated Wine, Replicated Vowel, and the results were compared with respect to intra-cluster distance with popular MapReduce based meta-heuristic algorithms namely, Parallel K-Means Algorithm, Parallel K-PSO Algorithm, Dynamic Frequency based parallel K-Bat (DFBPKB) Algorithm. Further the authors, tested the proposed method for Recommendation System application. For this test, MovieLens dataset was used. The result of the test was compared with respect to Mean Absolute Error with state of art collaborative filtering algorithms namely, PCA-GAKM, PCA-SOM, SOM, UPCC, K-Means, PCA-K-Means, GAKM, ABC-KM.

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 TSWOAK, and further discusses the MapReduce-based proposed method MR-TSWOAK. Section 4 presents the experimental arrangements and results. And Section 5 gives the conclusion of the paper.

2. BACKGROUND

2.1 Data Clustering Approach

Data clustering is an unsupervised machine learning approach. Clustering algorithm checks for similar-looking data-points and group them in same cluster. Clustering in K clusters of N data-points is an iterative process. There is no training required unlike supervised approaches like regression and classification etc. Clustering is done in such a way that the sum of Euclidean distances of all data-points with thier corresponding centroid is minimum. 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 a set of 'N' data-points having 't' features each, where the value of i^{th} data-point's j^{th} attribute is denoted by 'z_{ij}'. The clustering algorithm runs iteratively and finds 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 the value of j^{th} attribute of i^{th} centroid is denoted by 'c_{ij}'. The set $C_i = \{c_{i1},c_{i2},\ldots,c_{it}\}$ is position vector of i^{th} cluster-centroid. The set C is found such that the intra-cluster distance is minimum. Any clustering algorithm should satisfy the given conditions:

1. Each cluster should contain at least one data-point, i.e, $Ci\neq\phi$, $\forall i \in \{1,2,3,\ldots,K\}$
2. No data-point can be outside of all clusters
3. There should be no data-point belonging to more than one cluster, i.e, $Cq\cap Cr=\phi$, $\forall q\neq r$ *and* $q,r \in\{1,2,3,\ldots,K\}$

Clustering algorithm are tested on the basis of intra-cluster distance. The intra-cluster distance is the summation of Euclidean distance of all data-points with their corresponding cluster-centroid. The intra-cluster distance is calculated as given in equation 1.

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

Where Z_i is a data-point belonging to C_i. $d(Z_i,C_i)$ is called as Euclidean distance between Z_i and its corresponding cluster-centroid C_i . The Euclidean distance is calculated as given in 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 [15] 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. It is shown in equation 9 and 10

$$\vec{D} = |\vec{C} \cdot \overrightarrow{X_{rand}} - \vec{X}| \quad (9)$$

$$\vec{X}(t+1) = \overrightarrow{X_{rand}} - \vec{A}.\vec{D} \quad (10)$$

3. PROPOSED METHOD

Whale Optimization Algorithm discards the solution (called whale in WOA) with bad fitness value. But there could be a case that global-optima would be near that solution. This could lead WOA to get into the trap of local-optima. The authors have therefore proposed a method called as Tournament Selection empowered Whale Optimization Algorithm optimized K-Means. Nature-inspired meta-heuristic algorithms are iterative and thus have high time complexity. Therefore, the nature-inspired algorithms cannot be applied on extremely large datasets. Therefore, the authors have further proposed the method over MapReduce programming model called as MapReduce-based Tournament Selection empowered Whale Optimized Algorithm optimized K-Means (MR-TSWOAK).

3.1 Tournament Selection empowered WOA optimized K-Means (TSWOAK)

There is a chance for WOA to get into local-optima as it just discards the solution with lower fitness value. Therefore, the authors have used the concept of Tournament Selection [17] while choosing a random solution for equation 9 and 10. Due to the concept of Tournament Selection, each and every solution get equal chance to compete with other solution to get selected.

3.2 TSWOAK based clustering

The problem of getting trapped in local-optima of clustering algorithm can be solved by using Tournament Selection empowered Whale Optimization Algorithm optimized K-Means (TSWOAK). In the proposed method TSWOAK, the position vector $\vec{X}$ of each whale is the position of centroid represented by the whale. $\vec{X}$ ={{c_{11},c_{12},…,c_{1t}} {c_{21},c_{22},…,c_{2t}},…,{c_{k1},c_{k2},…,c_{kt}}, where c_{ij} represents the centroid position for j^{th} feature of i^{th} cluster. Clustering performance is evaluated in terms of intra-cluster distance. The intra-cluster distance is calculated using formula given in equation 1. The whale with least intra-cluster distance is taken as best-whale. The algorithm for proposed method is given in Algorithm 1. The running time-complexity of the proposed method is O(P*N*K*t), where t is the number of iterations for which the algorithm will run. K is the number of clusters for which the clustering is done. N is the number of data-points. P is the number of whales used in algorithm.

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 3
            else2
                Select a random whale ($\overrightarrow{X_{rand}}$) by Tournament Selection
                Position of the current whale is updated by the Equation 10
            end if2
        else1
            Position of the current whale is updated by the Equation 7
        end if1
    end for
    if available better solution Update $\overrightarrow{X^*}$
    t=t+1
end while
return $\overrightarrow{X^*}$

Algorithm1: Tournament Selection empowered Whale Optimization Algorithm

3.3 Parallelization of TSWOAK using MapReduce architecture

The running time-complexity of the algorithm TSWOAK is O(P*N*K*t). The time-complexity shows that it is directly proportional to number of data-points involved. For large number of data-points, the time-complexity will be become very high. For performing clustering on large datasets, parallel-computing is required. For parallel-computing, the authors have developed TSWOAK over MapReduce-programming model. The parallel-computing version is called as MapReduce-based TSWOAK (MR-TSWOAK). MR-TSWOAK works in two phases, MR-TSWOAK-Map and MR-TSWOAK-Reduce. The data-points are distributed uniformly among Hadoop Distributed File System (HDFS) [12] DataNodes. MR-TSWOA-Map processes data-point and finds cluster in each whale, which would have the least Euclidean distance between the data-point and found centroid. The output of the Map phase is {key:(whaleId, cenId), value:minDistance}, where ‘whaleId’ is the identification of whale of whose clusters are being matched with the data-point.,‘cenId’ is the identification of cluster-centroid whose Euclidean distance is minimum

with the data-point. ‘minDistance’ is the Euclidean distance between data-point and the centroid with identification ‘cenId’. The pseudo-code of Map phase is given in Algorithm 2. MR-TSWOAK-Reduce phase processes the distance given by Map phase, and calculate intra-cluster distance for each centroid of each whale. The output of this phase is of the form {key:(whaleId, cenId), value:intra_cluster_distance}. The psudo-code of Reduce phase is given in Algorithm 3. The architecture of MR-TSWOAK is given 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-TSWOAK-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-TSWOAK-Reduce

3.4 Application of MR-TSWOAK on Product Recommendation System

The recommendation matrix is stored on the Hadoop Distributed File System (HDFS). The recommendation matrix has dimension of (users, product). The number of rows of matrix is equal to number of users. The number of columns is equal to number of products. Each row of matrix is a datapoint to MR-TSWOAK. The MR-TSWOAK is implemented over the recommendation matrix, and it output the desired solution of centroid position vector. Further this position vector is fine-tuned by running K-Means over it. The system is tested by calculating the Mean Absolute Error.

4. EXPERIMENTAL RESULTS

Evaluation of proposed method is done in three phases. In the first phase, TSWOAK is tested on seven UCI datasets and the results are compared on the basis of intra-cluster distance with five state of the art meta-heuristic algorithms namely PSO, K-Means, GSA, BA, GWO. The seven UCI datasets are Iris, Seeds, Glass, Cancer, Balance, Haberman, and Wine. In the second phase, the MapReduce based method MR-TSWOAK is tested on four extremely large datasets namely Replicated CMC, Replicated Vowel,

Replicated Iris, Replicated Wine, and the results are compared to Parallel K-PSO, DFBKBA, Parallel K-Means, MR-ABC on the basis of F-measure and computation time. In the third phase, the proposed method MR-TSWOAK is tested for the applicability in recommender systems. For this it is tested on MovieLens dataset, and its results will be compared on the basis of Mean Absolute Error (MAE) with the results of already present algorithms namely, K-Means, PCA-K-Means, ABC-KM, GAKM, PCA-GAKM, SOM, PCA-SOM, UPCC.

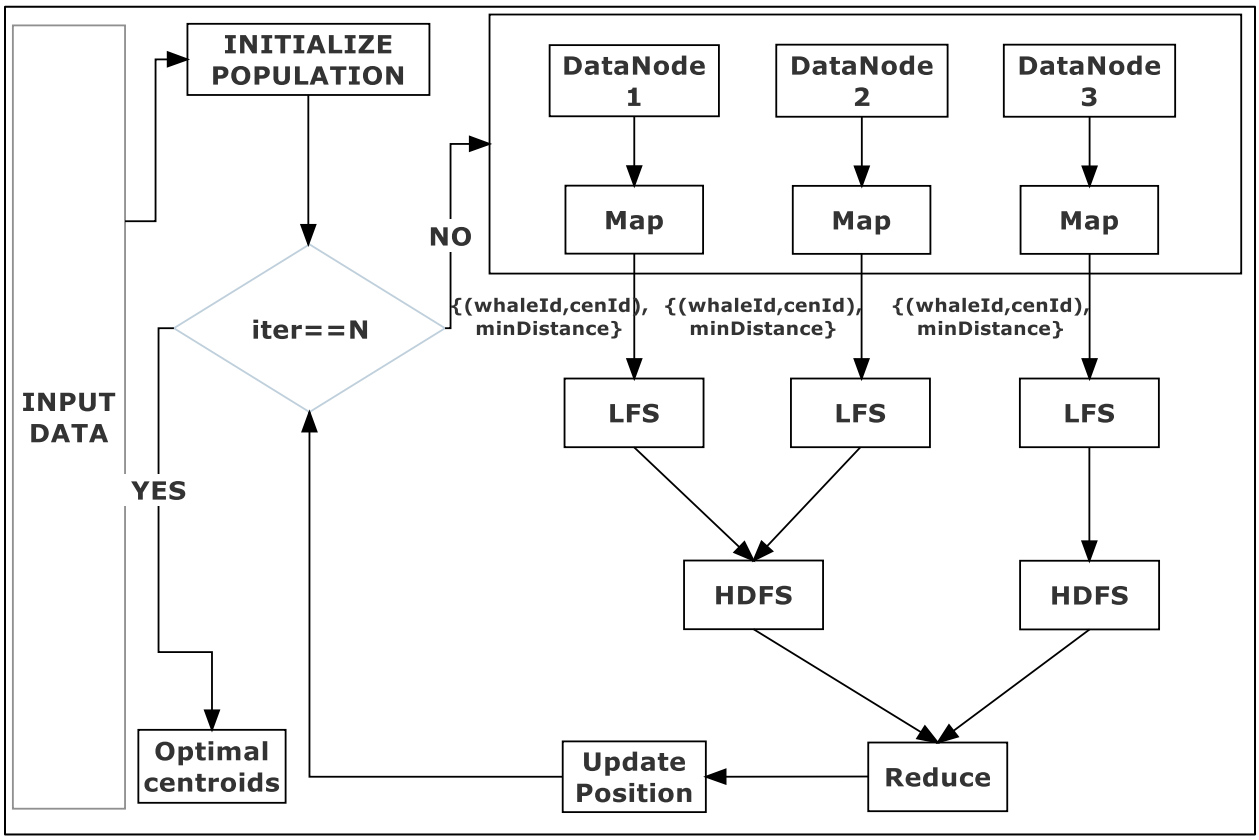

Figure 1: Architecture of MR-TSWOAK

4.1 Experimental setup

The experiments related to TSWOAK were done on a single machine, with specification as 2.5 GHz Intel i5 processor, 8GB RAM, and 1 TB hard-disk.

For experiments related to the MR-TSWOAK, Hadoop cluster of capacity of five nodes was made. Each node had specification as 2.5 GHz Intel i5 processor, 8 GB RAM, 1 TB hard-disk. Java version used was openjdk "1.8.0_191". The version of Hadoop was 2.6.0.

4.2 Performance analysis of TSWOAK based clustering

The proposed method TSWOA was tested on seven UCI datasets namely, Iris, Seeds, Glass, Cancer, Balance, Haberman, and Wine, and the results were compared on the basis of intra-cluster distance with five popular meta-heuristic algorithms namely, PSO, K-Means, GSA, BA, and GWO. Description of seven datasets used is given in Table 1. The setting parameters of each meta-heuristic algorithm used for experimentation are described in Table 2. The intra-cluster distance found by each meta-heuristic algorithm on each dataset is given in Table 3.

Table 1: Dataset description

Dataset	Number of Datapoints	Number of Attributes	Number of Clusters
Balance	625	4	3
Cancer	638	9	2
Iris	150	4	3
Haberman	306	3	2
Wine	178	13	3
Glass	214	9	6
Seeds	210	7	3

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

Parameter name	K-Means	BA	PSO	GWO	GSA	TSWOAK
Population size	-	40	40	40	40	40
Cognitive Constant (c1)	-	-	1	-	-	-
r0	-	9	-	-	-	-
Gamma ()	-	9	-	-	-	-
Alpha (α)	20	2	-	2	9	-
fmin	-	0	-	-	-	-
fmax	-	2	-	-	-	-
Inertial Constant (w)	-	-	0.5	-	-	-
Social Constant (c2)	-	-	1	-	-	-
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	GWO	PSO	GSA	TSWOAK
Iris	97.34084	96.65552	96.65826	96.78998	96.65548	**96.65548**
Seeds	587.31957	311.79816	311.88200	312.68370	311.79804	**311.79804**
Haberman	30507.0207	2566.98889	2567.02562	2566.99548	2566.98989	**2566.98889**
Glass	292.75724	243.70331	265.81420	238.51144	286.11855	**214.44399**
Wine	2370689.68700	16371.05448	16307.09242	16298.98906	17038.59226	**16292.18465**
Cancer	19323.1738	2964.38718	2964.390179	2969.23958	2970.17834	**2964.38697**
Balance	3472.32142	1424.04307	1423.82106	1423.96787	1423.82042	**1423.82040**

Table 3 shows that the proposed algorithm TSWOAK gives the minimum intra-cluster distance with respect to other algorithms used in experimentation.

The convergence speed of the proposed algorithm TSWOAK is more than the convergence speed of other algorithms with which the comparison has been done. Figure 2 and figure 3 shows the convergence graph of the algorithms on Wine and Glass datasets respectively. In both the diagram, the Y-axis depicts the best

intra-cluster distance for an algorithm, the X-axis depicts the number of iterations for which the algorithm has been run.

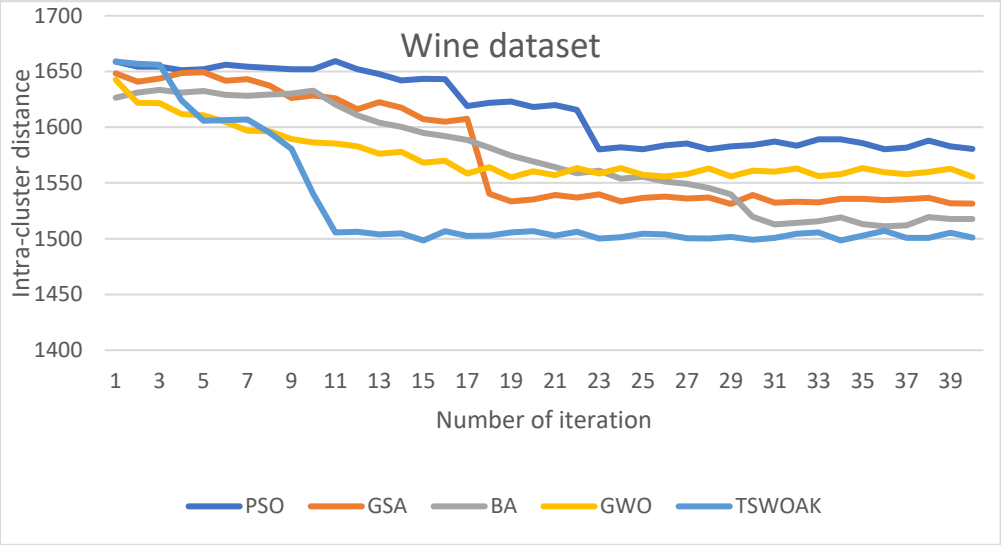

Figure 2: Convergence graph of Wine

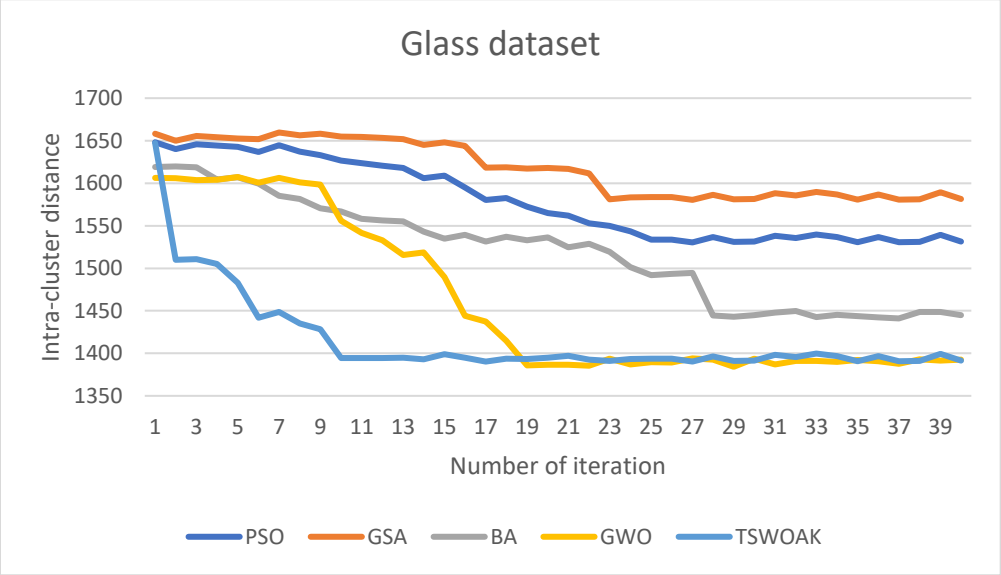

Figure 3: Convergence graph of Glass

4.3 Performance analysis of MapReduce-based TSWOAK

Section 4.2 have shown that the proposed algorithm TSWOAK has given better result in comparison to other algorithms. Hence, the authors have proposed a MapReduce-based version of TSWOAK called as MR-TSWOAK to handle extremely large datasets. The performance of MR-TSWOAK is compared with other state of art MapReduce-based meta-heuristic algorithms on the basis of F-Measure and computation time. Table 4 describes the large datasets to be used. Table 5 gives the results of different algorithms on different dataset on the basis of F-measure and computation time.

Table 5 shows that the performance of MR-TSWOAK is better than other algorithms, but the parallel K-Means is computationally fast.

Speedup-performance is also a major criteria to evaluate a parallel-computing algorithm. The calculation of speedup-performance is done according to equation 12.

Table 4: Description of large datasets

Dataset	Number of Datapoints	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	Parallel K-Means	DFBPKBA	MR-ABC	MR-TSWOAK
Reproduced CMC	Computation time	10.33E + 04	**8.24E + 04**	10.34E + 04	10.33E + 04	10.32E + 04
	F-Measure	0.324	0.298	0.378	0.387	**0.391**
Reproduced Vowel	Computation time	13.22E + 04	**10.50E + 04**	13.23E + 04	12.21E + 04	13.21E + 04
	F-Measure	0.627	0.586	0.622	0.634	**0.635**
Reproduced Iris	Computation time	9.23E + 04	**8.05E + 04**	9.24E + 04	9.26E + 04	9.22E + 02
	F-Measure	0.785	0.667	0.790	0.842	**0.846**
Reproduced Wine	Computation time	13.22E + 04	**11.20E + 04**	13.23E + 04	12.21E + 04	13.21E + 04
	F-Measure	0.627	0.586	0.622	0.634	**0.635**

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

Where T_{base} is time taken by algorithm to run on one machine. T_N is time taken by the algorithm to run on N machines. Speedup performance of MR-TSWOAK is studied on Replicated CMC and Replicated Iris datasets. Figure 4 and 5 are the speedup graph of MR-TSWOAK running on Replicated CMC and Replicated Iris datasets. The speedup graph has Y-axis for computation time, and X-axis for the number of nodes in the cluster. The speedup performance of MR-TSWOAK running on CMC dataset is 4.7548, when there are five nodes in the cluster. The speedup performance of MR-TSWOAK running on Replicated Iris dataset is 4.4561, when there are five nodes in the cluster. This shows that MR-TSWOAK can be used on large datasets.

4.4 MR-TSWOAK application in Recommender systems

The recommendation system developed over MR-TSWOA is tested on MovieLens dataset, which contains 100,000 datapoints and has ratings given by users to different movies. There are 1000 users in the dataset and 1700 movies in the dataset. Each user has given rating to at least 20 movies. The results of the test are compared with the results of PCA-GAKM, PCA-SOM, SOM, UPCC, K-Means, PCA-K-Means, GAKM, ABC-KM on the basis of Mean Absolute Error (MAE). The results are given in Table 6.

Table 6 shows that MR-TSWOAK perform better than the already present collaborative-filtering clustering algorithm. Figure 6 shows the comparison of MAE with different methods.

Table 6: MAE for different approaches for different number of clusters

Cluster	5	10	15	20	25	30	35	40
K-WOA	0.74169	0.69	0.68813	0.69628	0.68675	0.68037	0.68635	0.68704
PCA-GAKM	0.79	0.77	0.77	0.77	0.781	0.785	0.786	0.788
PCA-SOM	0.82	0.79	0.79	0.79	0.8	0.805	0.806	0.807
SOM-CLUSTER	0.819	0.81	0.81	0.81	0.81	0.805	0.81	0.81
UPCC	0.825	0.825	0.824	0.828	0.824	0.824	0.825	0.825
K-Means Cluster	0.825	0.82	0.818	0.815	0.815	0.814	0.813	0.812
PCA-K-Means	0.85	0.845	0.841	0.84	0.84	0.84	0.84	0.84
GAKM-Cluster	0.815	0.805	0.804	0.804	0.804	0.804	0.803	0.803
ABC-KM	0.773	0.764	0.764	0.771	0.78	0.784	0.782	0.787

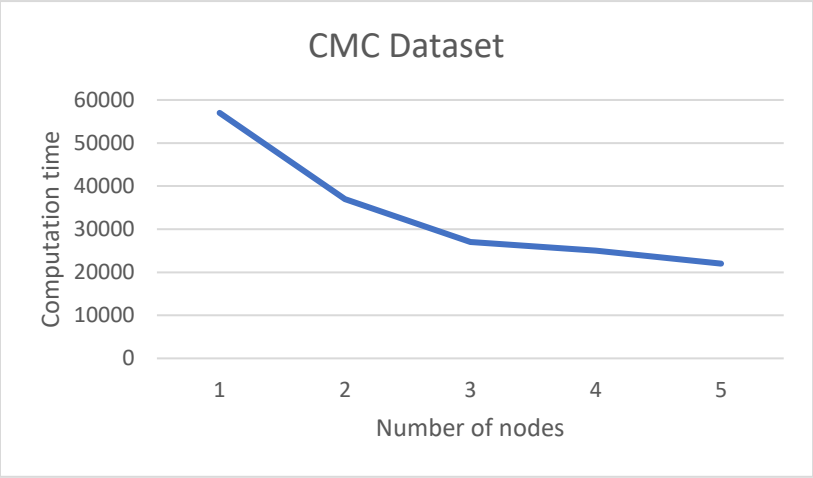

Figure 4: Speedup graph of Replicated-CMC dataset

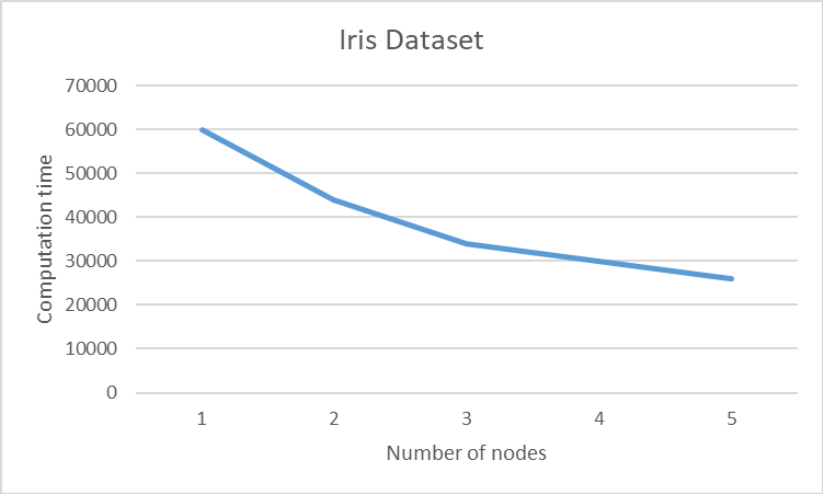

Figure 5: Speedup graph of Replicated-Iris dataset

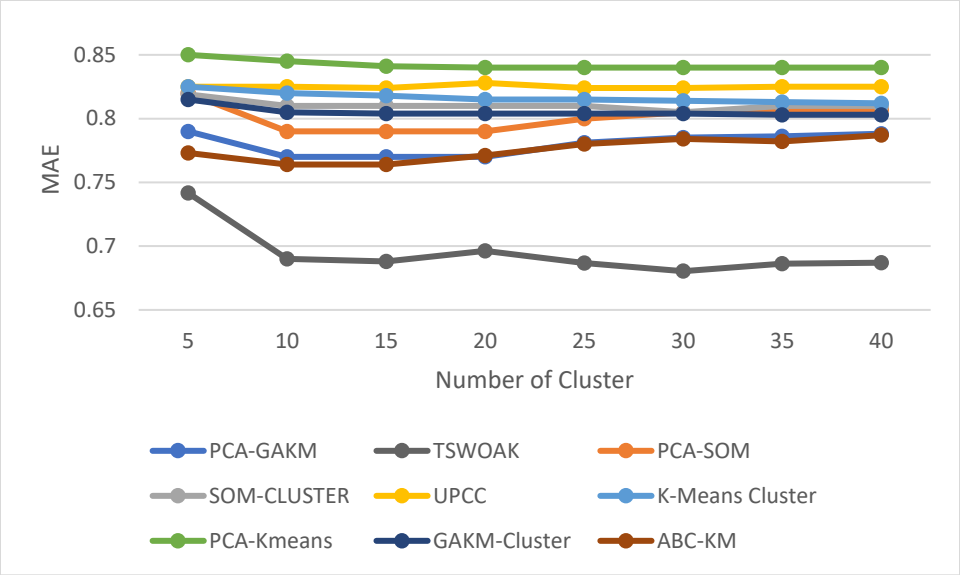

Figure 6: Comparison of MAE with different methods

5. CONCLUSION

In the 21st century, it is very important for e-commerce business to assist their customers into buying better product for themselves. For this kind of assistance, recommendation systems have come into picture and quite popular these days. Collaborative-filtering is a popular approach for recommendation. K-means is a popular algorithm used for collaborative-filtering. Though recommendation systems have been successful in assisting customers, but still they suffer from the problem of getting into local-optima, and have scalability problems. To overcome these problems, the authors have proposed a method that uses the hunting behaviors of Humpback whales and the concept of tournament selection. This method is called as Tournament Selection empowered Whale Optimization Algorithm optimized K-Means (TSWOAK). This method is tested on seven benchmark UCI datasets namely, Haberman, Cancer, Iris, Glass, Wine, Balance, Seeds, and the results are compared with the results of start of art algorithms namely, K-Means, BA, PSO, GWO, GSA, on the basis of intra-cluster distances. The comparison has shown that the proposed method works better than already present clustering algorithms. To solve the problem of scalability, the authors have adopted TSWOAK on MapReduce programming model, called MR-TSWOAK. The MapReduce-based model MR-TSWOAK was tested on four large datasets namely, Reproduced CMC, Reproduced Vowel, Reproduced Iris, Reproduced Wine, and the results are compared with the results of Parallel K-PSO, DFBPKBA, Parallel K-Means, MR-ABC, on the basis of F-Measure and Computation time. The comparison has shown that the F-measure found by MR-TSWOAK is better than those by other algorithms. The comparison has implied that the MapReduce-based method can be used for clustering purpose. The authors have further tested MR-TSWOAK on MovieLens dataset to check for Recommendation System applicability. The results were compared to the result found by state of art collaborative-filtering algorithms namely, PCA-GAKM, PCA-SOM, SOM, UPCC, K-Means, PCA-K-Means, GAKM, ABC-KM on the basis of Mean Absolute Error (MAE). The comparison has shown that the MAE found by MR-TSWOAK is lower than MAE found by other algorithms. The comparison has implied that the proposed method MR-TSWOAK can be used for Recommendation Systems application.

In future, Tournament Selection empowered Whale Optimization Algorithm shall be used to optimize the weight of Convolution Neural Network, which would be implemented over MapReduce. This shall be done to check if unsupervised learning is better or supervised learning is better for the product recommendation.

References

[1] Liu, Haifeng, Xiangjie Kong, Xiaomei Bai, Wei Wang, Teshome Megersa Bekele, and Feng Xia. "Context-based collaborative filtering for citation recommendation." IEEE Access 3 (2015): 1695-1703.

[2] Bobadilla, Jesús, Fernando Ortega, Antonio Hernando, and Abraham Gutiérrez. "Recommender systems survey." Knowledge-based systems 46 (2013): 109-132.

[3] Lam, Xuan Nhat, Thuc Vu, Trong Duc Le, and Anh Duc Duong. "Addressing cold-start problem in recommendation systems." In Proceedings of the 2nd international conference on Ubiquitous information management and communication, pp. 208-211. ACM, 2008.

[4] Ungar, Lyle H., and Dean P. Foster. "Clustering methods for collaborative filtering." In AAAI workshop on recommendation systems, vol. 1, pp. 114-129. 1998.

[5] Tripathi, Ashish Kumar, Kapil Sharma, and Manju Bala. "A novel clustering method using enhanced grey wolf optimizer and mapreduce." Big data research 14 (2018): 93-100.

[6] 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.

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

[8] 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.

[9] 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.

[10] 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.

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

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

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

[14] Tripathi, Ashish Kumar, Kapil Sharma, and Manju Bala. "Dynamic frequency based parallel k-bat algorithm for massive data clustering (DFBPKBA)." International Journal of System Assurance Engineering and Management 9, no. 4 (2018): 866-874.

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

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

[17] Miller, Brad L., and David E. Goldberg. "Genetic algorithms, tournament selection, and the effects of noise." Complex systems 9, no. 3 (1995): 193-212.