International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 01 | Jan 2020 www.irjet.net p-ISSN: 2395-0072

Improved Model for Big Data Analytics using Dynamic Multi-Swarm Optimization and Unsupervised Learning Algorithms

C.P. Oleji[1], E.C. Nwokorie[1], G. Chukwudebe[1], E.O. Nwachukwu[2]

[1]Department of Computer Science, Federal University of Technology Owerri, Imo, State, Nigeria.
[2]Department of Computer Science, University of Port Harcourt, Rivers State, Nigeria.

Abstract - *Big Data poses enormous challenges to traditional machine learning algorithms. In Information Technology, Big Data is a collection of datasets so large and complex that they become difficult to compute and process using existing database management tools or traditional data processing applications. In this work we have proposed an improved model for Big Data analytics using dynamic multi-swarm optimization and unsupervised learning algorithms. It explored the concept of Swarm Intelligence, Reference Distance Weight (RDW) of dissimilarity measure, Euclidean distance measure, and Chi-square relative frequency of dissimilarity measure to develop improved clustering algorithm for Big Data analytics called DynamicK-reference Clustering Algorithm. Seven datasets (voting, Iris, wine, Australian Credit Approval, German Credit Data and Statlog Heart) obtained from University of California Irvine (UCI) Machine Learning Repository was used to demonstrate the clustering performance of the proposed hybridized algorithm. The results show that the proposed hybrid clustering algorithm could be more proficient for clustering mixed large datasets than Fuzzy Artificial Bee Colony (FABC) algorithm, k-nearest Neigbors (KNN) algorithm, Particle swarm optimization (PSO) algorithm, Artificial Bee Colony (ABC) Optimization algorithm and PSO based K-prototype algorithm. It is robust and very efficient at expelling outliers from its dissimilar clusters/classifications.*

Key Words: **Dynamic Multi-swarm, Big Data, Unsupervised Learning, Clustering Algorithm, Analytic, K-prototype**

1. INTRODUCTION

The massive amount of data being generated per second through various social media platforms, online marketing platforms, and business websites among others generally defines Big Data [1]. These data may be pre-processed and analyzed upon acquisition to make better event predictions and knowledge discoveries for the benefit of man. Hence, the fast growing of the application of new technology development in this modern time, has brought about vary large unstructured and structured information from various areas of its applications and processes. Very large unstructured and structured data has the following characteristics: large volume, velocity, variety, veracity, volatility, validity and variable. These characteristics are used to describe the concept and challenges of big data analytic. Big Data consists of complex or very large data that the traditional data mining applications cannot efficiently analyze.

Data mining techniques search for consistent patterns and/or systematic relationships between variables, and validate the findings by applying the detected pattern to form new subsets of data. It is an interdisciplinary subfield of computer science. It is the computational process of discovering patterns in large data set involving methods at the intersection of artificial intelligence, Unsupervised Machine learning, Optimization algorithms, business intelligence, statistics, high performance computing and database system. Unsupervised Learning is a branch of machine learning that learns from test data, in which the domain is unknown. In the Swarm Intelligence (IS) family, Dynamic Multi-swarm optimization is a variant of PSO centered on the use of multiple sub-swarms instead of one (standard) swarm. It has the potential to regroup each swarm for next task after completing its assign task. Particle swarm optimization is a population based stochastic optimization technique that can be used to fine the finest result to complex numerical and qualitative problems.

This paper is aimed at developing an improved model for big data analytics using SI and unsupervised machine learning algorithms. Traditional unsupervised machine learning algorithms has been found to be weak and inefficient to handle big data analytical challenges. Unsupervised Machine learning algorithms such as K-prototype clustering algorithm is an integration of k-means and k-mode clustering algorithms. K-means clustering algorithm is well known for clustering large numerical datasets. The use of k-means is often limited to numerical attributes. K-mode clustering algorithm was developed to extend K-means to cluster categorical attributes which is based on simply mismatches between objects in the given region. It is practically more useful for mixed-type object. "These partitional clustering algorithms appear not to be efficient in handle big data analysis, because of the rigorous nature of the datasets to be analyzed [15]". Swarm Intelligence (SI) systems respond well to the rapidly changing environment of the dynamics of big data manipulations, making use of their inherent auto-configuration and self-organization capability. This adaptability character of SI allows them to autonomously adapt their individual behavior to the external environment dynamically on run-time processes, with substantial flexibility.

The general approach in Dynamic Multi-swarm Optimization (DMSO) is that each sub-swarm concentrates on a specific region of the high dimensional dataset while a specific diversification method decides where and when to launch the sub-swarm [15]. "The system copies the social characters shown by swarms of animals. In this algorithm a point in the search space, which is a possible solution, is called a particle. The group of particles in a specific alteration is called 'swarm'. While looking out for food, the birds are either scattered or go collectively before they find out the place where they are able to locate the food. While the birds are on the search for food moving from one location to another, there is often a bird which is able to smell the food effectively, in other words, the bird is discernible of the location where the food is likely to be found, having superior food resource data. As they tend to convey the data, particularly the excellent data at any time while looking for the food from one location to another, attracted by the excellent data, at the end, the birds will throng at the location where there is strong possibility for locating food. The integration of this mechanism with traditional unsupervised machine learning algorithm will enhance its performance by obtaining the global best initial search for k value, which is the major drawback of partitional algorithm that makes it to converge prematurely. However, many researchers tend to solve these problems but end up creating more problems to be solved [15]."

2. LITERATURE REVIEW

"Imran and Rao [2] proposed a novel technique on class imbalance Big Data using Analogous under Sampling Approach". Their proposed Under Sampled Imbalance Big Data (USIBD) knowledge discovery framework is robust and less sensitive to outliers where non-uniform distribution of data is applied. "Ali and Site [3] presented a simplified analytical model towards Big Data analysis using Ridge Regression Method". Their simulation result represented a mapping model of Gaussian data from big data in sufficient scale. Their proposed model presents the new gateway for big data statistical and mathematical analysis. Ridge regression cannot really handle uncertainties and variation in big data analytics and dynamic systems. "Aishwaray and Jyothi [4] presented handing big data analytic using swarm intelligence". Their intention was to prove that big data analytics problem can be solved using swarm intelligence and its application on Hadoop architecture. They used PSO algorithm to create clusters of a given dataset. Their experimental results show that PSO algorithm is efficient in solving the analytical problems faced in big data analysis. However, they were not able to develop effective hybridized algorithm with swarm intelligence to solve big data analytics problems. "Big Data Analytic with swarm intelligence was presented by [5]. They described the association between big data analytics and IS techniques. Their work compared the mean and standard deviations of three PSO variants (such as; Cooperative coevolving particle swarm Optimization (CCPSO), Competitive Swarm Optimizer (CSO) and Dynamic Multi-swarm Optimizer (DMSO) in solving problem with 1,000 decision variables. The results show that DMSO produces more accurate results than CCPO2 and CSO algorithms. They concluded that the solutions found by PSO variants are good enough for solving challenges of big data analytics with huge number of decisions". Cao and Jiao [6] proposed big data: A parallel Particle Swarm Optimization-Back Propagation (BP) Neural Network Algorithm Based on Map Reduction. They used the POS algorithm to optimize the BP neural network's initial weights and thresholds. And also, to improve the accuracy of the classification algorithm the MapReduce parallel programming model was utilized to achieve parallel processing of the BP algorithm. Their results showed higher significant classification accuracy and improved time complexity of the process. However, neural network is a supervised learning model that is applicable in a known domain problem. Their proposed system will not be efficient when solving problems in an unknown domain. Sathesh and Hemalatha [7] proposed an innovative potential on rule optimization using Fuzzy Artificial Bee Colony Algorithms. Their system was used to optimize rules to get the best classification accuracy of real-world datasets. They proposed system was compared with the traditional bee colony optimization and particle swarm optimization algorithms. Another interesting work of [8] presented a new method of Fuzzy Clustering by using the combination of the Firefly Algorithm (FA) and the Particle Swarm Optimization algorithm. They used their proposed algorithms to compute the maximum distance between the clusters centers and the cost of clustering the datasets to improve the objective function of fuzzy clustering algorithm. The approach they used improved the accuracy of fuzzy clustering algorithm. But there was no clear intelligent procedure for switching of the two swarm optimization algorithms to get the global initial value of k. It will be time taking for both of the optimization algorithm to evaluate the high dimensional dataset before switching to the next optimization algorithm.

"Prabha and Visalakshi [9] proposed a new variant of binary Particle Swarm Optimization and K-Prototype algorithms to reach global optimal solution for clustering optimization problems". Their proposed algorithm was implemented and evaluated on standard benchmark dataset taken from UCI machine learning repository. Their results produced accurate clustering, but it is time consuming for a swarm to search for X particles moving around D-dimensional search space. In the work of [10], they used meta-heuristic mechanism to produce solution to big data tasking scheduling and analytic for decision making, which the traditional algorithm has inefficient analytic strength to handle. They suggested that using other heuristic approach will solve big data analytical challenges. In addition, Palake and Vikas [10] commented that the size of big dataset posed challenges on traditional algorithms; they provide solution to the problem by using Genetic algorithm because of its self-sufficient to handle big dataset in global space. They concluded that in order to solve these problems, in future, meta-heuristic algorithm should be combined with any other techniques to help improve the efficiency of Big Data Analytics.

3. PROPOSED SYSTEM

This paper proposes an improved model for big data analytics using dynamic multi-swarm optimization and unsupervised learning algorithms. The concept of Reference Distance Weighted (RDW) of dissimilarity measure, Euclidean Distance (ED) measure and Chi-square Relative Frequency (CRF) of dissimilarity measure was used to develop an improved clustering algorithm to efficiently cluster mixed dataset. The improved clustering algorithm was hybridized with Dynamic Multi-Swarm Optimization (DMSO) Algorithm to form a robust and efficient clustering algorithm called "DynamicK-reference Clustering Algorithm" for Big Data analytics.

First a simple robust numerical clustering algorithm called RDW-K-means clustering algorithm were developed with Reference Distance Weighted (RDW) and Euclidean distance measure. The RDW calculates distance between feature-vectors with real values (i.e obtaining classified future data-points of clusters) in a given region of the data points. This mechanism improves the traditional K-means clustering algorithm to obtain accurate similarities clusters of attributes in a given region. Secondly RDW-K-means clustering algorithm were integrated with Chi-square relative frequency dissimilarity measure used in K-representative clustering algorithm to develop improved clustering algorithm called K-reference Clustering algorithm for clustering unstructured (mixed) dataset. Thirdly, K-reference clustering algorithm was hybridized with Dynamic Multi-swarm Optimization algorithm to develop robust and efficient algorithm called DynamicK-reference Clustering algorithm that will overcome the problems that Big Data characteristics posed on traditional data mining techniques on creating clusters of mixed datasets. The intelligent concept of Dynamic Multi-Swarm will guide the proposed clustering algorithm to obtain accurate convergences of the objective function by providing best global initial value of K clusters. Dynamic multi-swarm optimization algorithm was chosen among other meta-heuristic algorithms because it has the capacity to achieve an effective balance in a multi-model problems, instead of trying to reach a compromise between exploration and exploitation which could weaken both mechanisms of the search process like in the case of Artificial Bee Colony and Firefly algorithms etc., Dynamic Multi-swarm system separates them into distinct phases. Each phase is more focused on either exploitation (individual sub-swarms) or exploration (diversification method)". The general approach in multi-swarm optimization is that each sub-swarm focuses on a specific region while a specific diversification method decides where and when to launch the sub-swarm. It partitions the entire space into many sub-spaces, each swarm optimizes the location of the best global particle in a local environment.

4. METHODOLOGY

Object Oriented Analysis and Design Methodology were used for the design and implementation of the hybrid clustering algorithm with java programming language and MATLAB. This methodology features: the algorithm, architectural design and flowchart Diagram.

4.1 Proposed RWD-K-means Clustering Algorithm

The Proposed RWD-K-means Clustering Algorithms are as follow:

Step one: obtain the global best value of clusters k from the result of Dynamic Multi-swarm algorithm.

Step two: Randomly selecting the centroids (v_1, v_2...v_k) in the data set.

Step three: Calculate the Reference Distance Weight RDW (1) of the corresponding centroids (v_1, v_2...v_k) of the data points in the dataset.

Step four: Calculate RDW

(s, t, α) =

$$\frac{\sum_{i=1}^{n}\alpha_i\left|\frac{s_i-t_i}{s_i}\right|}{n}=\frac{\sum\left|1-\frac{t_i}{s_i}\right|}{n}, i=1,2,\dots n.[12] \qquad (1)$$

Where: RDW is the corresponding reference weight vector to the s_i,

$s = \{s_1, s_2 \dots s_n\}$ is the features-vector of reference, associated to a class, vector from whom the distance of the data points are measured [12];

$t = \{t_1, t_2 \dots t_n\}$ is the feature-vector associated to the problem that must be solved or object that must be classified in the dataset, vector up to which measured the distance of the data point in the dataset;

$\alpha = \{\alpha_1, \alpha_2.., \alpha_n\}$, is a vector called relevance vector, whose components α_i are parameters specific for each feature in part and assigned to each feature, called relevance factor, proportional to the importance/weight of the respective feature under the conditions of problem to be solved or the objects (data points) of the datasets to be clusters [12].

Step five: Find the distance between the centroids using the integration of Reference distance weighted and Euclidean Distance equation.

$$d_{ij} = ||RDW_i(s_i - t_k)||^2 \qquad (2)$$

Step six: Update the centroids Stop the process when the new centroids are nearer to old one. Otherwise go to step-four.

4.2 Dissimilarity Measure of Categorical Domain

The categorical domain of mixed dataset analysis from the review of literatures is analyzed using the dissimilarity measures between categorical objects and the representative of a cluster, is defined based on simple matching approach of chi-square expression. The process is as follows:

Given C = {S_1, …, S_P) is a cluster of categorical objects, with

S_i = (s_1, …, $s_{i,m}$), 1 ≤ i ≤ p, and S = (s_1, … , s_m) a categorical object.

Note that in some cases S may or may not belong to C. Assume that Q = (q_1, …,q_m), with t_j ={(c_j, rf_{cj})|c ∈ D_j}, is a representative of cluster C.

Now, the dissimilarity between object S and representative Q is defined by

$$d(S,Q)=\sum_{j=1}^{m}\sum_{c_j \varepsilon D_J} rf_{cj}.\delta(s_j,c_j) \qquad (3)$$

Where c = cluster and rf = relative frequency between the clusters.

Under such a definition, the dissimilarity d(S, Q) is mainly dependent on the relative frequencies of categorical values within the cluster and simple matching between categorical values. It is also of interest to note that the simple matching dissimilarity measure between categorical objects can be considered as a categorical counterpart of the squared Euclidean distance measure.

It is easily seen that

$$d(S,Q)=\sum_{j=1}^{m}\sum_{c_j \varepsilon D_J} rf_{cj}.\delta(s_j,c_j) \qquad (4)$$

$$=\sum_{j=1}^{m}\sum_{c_j \varepsilon D_j, c_j \neq x_j} rf_c \qquad (5)$$

$$=\sum_{j=1}^{m}\left(1-rf_{x_j}\right), \qquad (6)$$

where rf_{xj} is the relative frequency of category x_j within C.

The enhanced k-means algorithm RDW_K-means clustering algorithm was integrated with k-representative clustering algorithm to produced RDW-K-reference clustering algorithm for clustering mixed dataset. The K-reference clustering algorithm consists of RDW_K-means algorithm and k-representative algorithm. The dissimilarity between two mixed-type objects S and T, which are described by attributes A^r_1, A^r_2,.., A^r_p, $A^c_{p+1,\ldots}$ A^c_m, can be measured by:

$$d_2(S,T) = RDW_i\sum_{j=1}^{p}(s_j-t_j)^2+\gamma\sum_{j=p+1}^{m} rf\delta(s_j,t_j) \qquad (7)$$

where the numerical attribute is the first term in equation (7) and the second term is the simple matching dissimilarity measure on the categorical attributes. To avoid the computing process for favoring either the categorical or numerical attributes the parameter weight γ was added to the process. The influence of γ in the clustering process was discussed in [13]. Using equation (7) for mixed-type objects, it is more convenient to modify the cost function of equation (7) as follows:

$$P(W,Q) = RDW_i\sum_{l=1}^{k}\left(\sum_{l}^{n} w_{i,l}\sum_{j=1}^{p}(x_j-y_j)^2+\gamma\sum_{l=1}^{n} w_{i,l}\sum_{j=p+1}^{m} rf\delta(x_j,y_j)\right) \qquad (8)$$

Let

$$P_l^r = RDW_i\sum_{j=1}^{p} w_{ij}\sum_{j=1}^{p}(s_{i,j}-t_{i,j})^2 \qquad (9)$$

And

$$P_1^c = RDW_i\gamma\sum_{i=1}^{n} w_{ij}\sum_{j=p+1}^{m} rf\delta(s_{i,j}t_{i,j})^2 \qquad (10)$$

Equation (10) can be re-written as:

$$P(W,Q) = RDW_i\sum_{l=l}^{k}(P_l^r + P_l^c) \qquad (11)$$

Since both P^r_l and P^c_l are nonnegative, minimizing P(W; Q) is equivalent to minimizing P^r_l and P^c_l for 1≤ l ≤ k; and rf is the relative frequency for the mixed data. The input of the proposed K-reference clustering algorithm is the global best value of k obtained from the swarm intelligent mechanism. It is used by the clustering algorithm to produce efficient clustering outputs (see Fig. 1).

The distance d ($\vec{x}$; $\vec{s}$) between two points and in the n-dimension space is defined as the Euclidean distance as shown in equation (1).

$$d(\bar{x},\bar{s}) = \sqrt{\sum_{i=1}^{n}(x_i - s_i)^2} \quad (12)$$

According to our experimental experience, the more peaks in the landscape, the more child swarms are relatively needed. r is relative to the range of the landscape and the width of peaks. In general, we set r according to the following equation:

$$r = \sqrt{\sum_{i=1}^{n}\left(x_i^u - x_i^l\right)}/\left(W_{\min} + c\left(W_{\max} - W_{\min}\right)\right) \quad (13)$$

where x_i^u and x_i^l are the lower and upper bound on the i-th dimension of the variable vector of n dimensions [14]. W_{min} and W_{max} are the minimum and maximum

4.3 Clustering Accuracy

A cluster is called a pure cluster if all the objects belong to a single class.

The clustering accuracy 'r' is defined as

$$r = \frac{1}{n}\sum_{l=1}^{k} a_l, \quad (14)$$

Where a_i is the number of data objects that occur in both cluster C_l and its corresponding labeled class. a_i has the maximal value and n is the number of objects in the data set. The clustering error e is defined as:

$$e = 1 - r. \quad (15)$$

E. Hybridization of the improved Clustering Algorithm with Dynamic Multi-Swarm Optimization Algorithms

The hybridization of the improved Clustering Algorithm with Dynamic Multi-Swarm Optimization Algorithm is as follows:

Step one: Initialized the swarm (intelligent search agents). Randomize position and velocities of each particle in search space, set all attractors to randomized particle (data points or objects) position, set swarm attractor to particle attractor 1 and set all stored function values to function floor.

Step Two: Evaluate function at swarm attractor of swarm n. (n is the sub-swarm deployed at that region).

IF new value is different from last iteration THEN

- Re-evaluate function values at each particle attractor.
- Update swarm attractor and store function values.

Step three: The sub-swarms examines each particle position

Step Four: Current Particle Moving to Different Swarm

Step Five: New Velocity Calculated for Current Particle

Step Six: After a new velocity is calculated, that velocity is used to move the current particle

Step Seven: Next, the error associated with the new particle position is determined and checked to see if it's a new best error for the current particle

Step Seven: Checking New Particle Position

Step Eight: Method Solve finished by returning the best position found

Step Nine: Obtain the global variable form best variable K of each sub-swarm.

Step Ten: Calculate the Reference Distance Weight (RDW) of the corresponding object representative.

Step Eleven: Calculate similarity and dissimilarity of the particles

Step Twelve: Compute dissimilarity measure for the mixed dataset (categorical and numerical)

Step Thirteen: Compute the Relative Frequency (rf) of the dissimilarity matrix of the mixed dataset

Step fourteen: Proceed to compute the clusters using the rf of the dissimilarity matrix as a parameter

Step Fifteen: If no object has changed clusters after a full cycle test of the whole data set

Step Sixteen: Stop

Step Seventeen: else go to Step Ten

4.4 Flowchart diagram of the hybridized system

The Flowchart diagram of the hybridized system is shown in Fig. 1.

The components of the Flowchart diagram of the proposed system includes: "data mining model", "deployment of sub-swarms to different regions of the dataset" and "improved k-reference clustering algorithm". The dataset extracted from UC Irvine machine learning repository was processed using data mining model such as data processing, data integration, data transformation, data cleansing and data reduction.

The result of the data mining model is used as an input for the Dynamic Multi-swarm intelligent search optimization algorithm to determine the landscape of the datasets. The mechanism of the particle attractor and swarm attractor were used to deployed sufficient sub-swarm to search for the initial global best value of k in the whole region of the dataset. The activities of these sub-swarms are controlled by

specific diversification method via the design decision to choose the best global value of k (initial value of k) from the results of each of the sub-swarms.

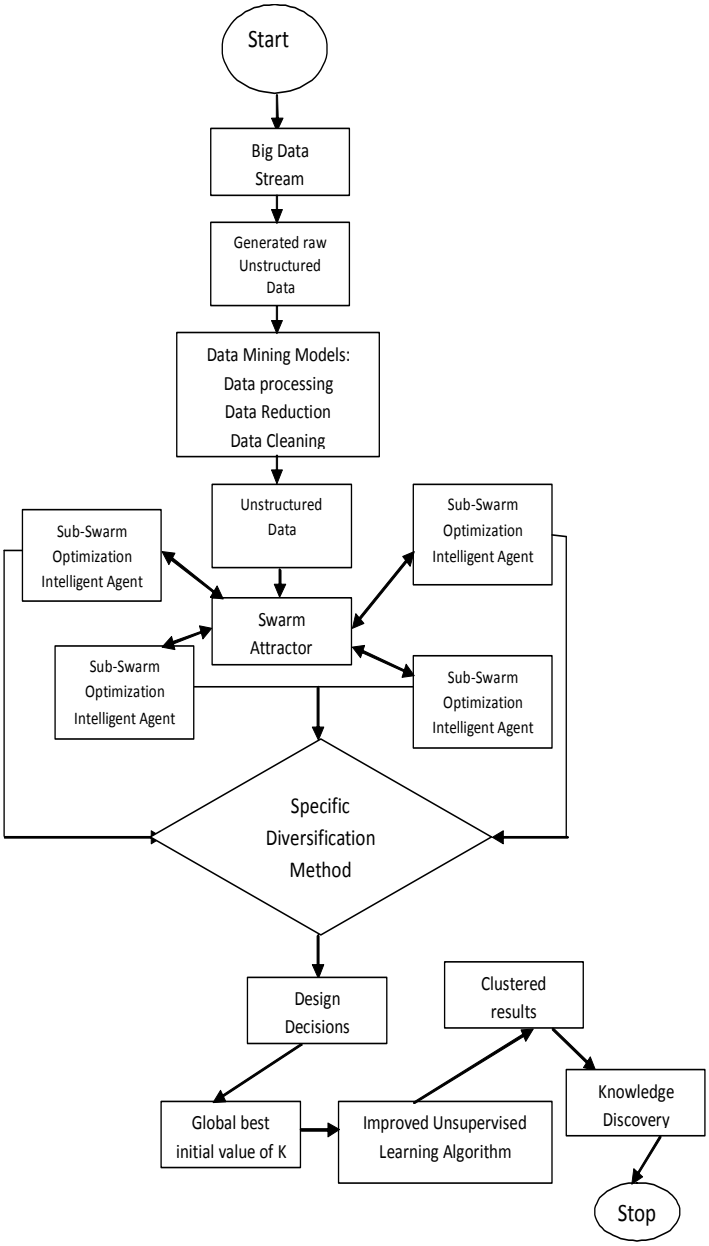

Fig-1: Flow Chart Diagram

The result of the data mining model is used as an input for the Dynamic Multi-swarm intelligent search optimization algorithm to determine the landscape of the datasets. The mechanism of the particle attractor and swarm attractor were used to deployed sufficient sub-swarm to search for the initial global best value of k in the whole region of the dataset. The activities of these sub-swarms are controlled by specific diversification method via the design decision to choose the best global value of k (initial value of k) from the results of each of the sub-swarms.

The global best value of k obtained by the swarm intelligence mechanism is used by the clustering algorithm to produce efficient clustering outputs for knowledge discovery.

5. RESULTS

The system that has been developed is shown in Fig. 2.

Fig- 2: Clustered output of Hepatitis dataset

It consists of "Random generated value of cluster", "Clustering sum of square error" and "Clustering Accuracy". The user clicks on "select dataset" to select the required dataset for analysis and then clicks on "Click to cluster" to analyze the results. The system outputs the "number of generated clusters", "sum of square error of created clusters", "clustering accuracy" and the "clustered datasets". The Clustering accuracy and Sum of square error used for the evaluation of the performance of the proposed algorithm were calculated using equations (14) and (15).

5.1 Discussion of Results

The proposed hybrid method was simulated with MATLAB R2018a version and Java programming language for data extraction respectively. Most of the reviewed literatures in this work did not implement their proposed systems; some used generated datasets for their evaluations. However, there are still some that implemented their works with real world datasets.

To evaluate the accuracy of the proposed hybrid clustering algorithm, seven (7) datasets were used. These included Voting, Iris, Wine, Hepatitis, Australian Credit Approval, German Credit Data, and Statlog Heart dataset all obtained from UCI Machine Learning Repository. Three datasets: Iris,

Wine, and Voting, were used to compare the performance of the proposed DynamicK-reference Clustering Algorithm with Fuzzy Artificial Bee Colony Optimization [7], KNN, PSO and ABC Optimization algorithms [7]. The results are shown in Fig. 3.

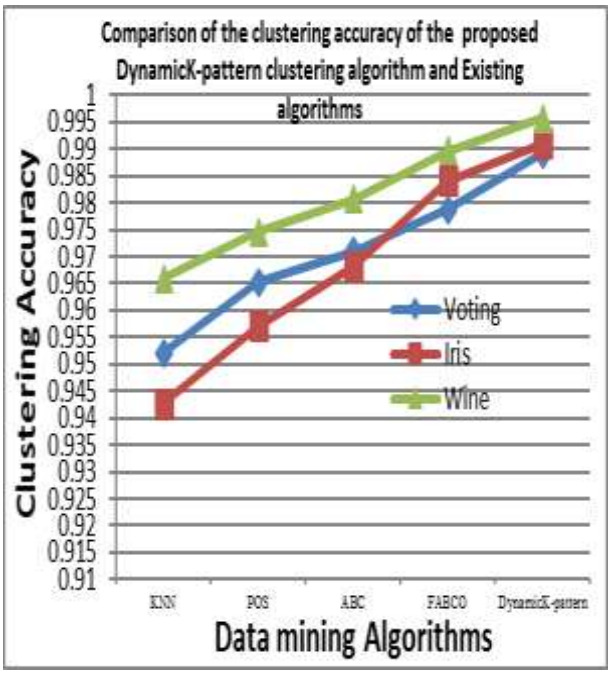

Fig-3: Graphical representation of comparison of the accuracy values of DynamicK-pattern and KNN. PSO. ABC AND FABCO algorithms.

The experimental results of DynamicK-reference clustering algorithm gave accuracy of (1.0000, 0.9927 and 0.9910) while for the existing algorithms, FABC Optimization algorithm gave accuracy of (0.9841, 0.9896 and 0.97865). ABC algorithm gave accuracy of (0.9682, 0.9896 and 0.97865). POS algorithm gave accuracy of (0.9571, 0.9746 and 0.96532). And KNN algorithm gave accuracy of (0.9428, 0.9659 and 0.95209). The results indicate that DynamicK-reference clustering algorithm has better clustering accuracy more than the Fuzzy Artificial Bee Colony Optimization, KNN, PSO and ABC Optimization algorithms.

Furthermore, the Clustering sum of square error of Fuzzy Artificial Bee Colony Optimization, KNN, PSO, ABC Optimization algorithms and the proposed DynamicK-reference clustering algorithm were compared. The result is shown in Fig. 4.

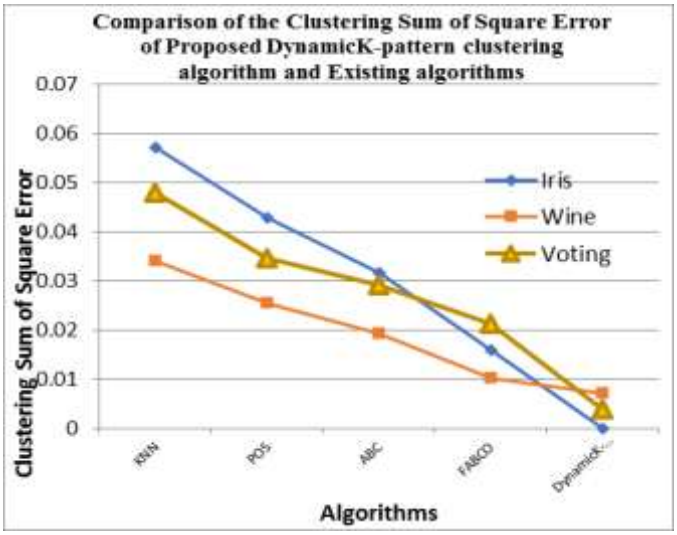

Fig-4: Graphical representation of comparison of the clustering sum of square error values of DynamicK-reference and KNN, PSO, ABC AND FABCO algorithms.

The simulation results for DynamicK-reference clustering algorithm gave clustering sum of square error of (0.0000, 0.0073 and 0.0039) while for the existing algorithms: FABC Optimization algorithm gave clustering sum of square error of (0.0159, 0.0104 and 0.02135); ABC algorithm gave clustering sum of square error of (0.0318, 0.0193 and 0.02909); POS algorithm gave clustering sum of square error of (0.0429, 0.0254 and 0.03468); And KNN algorithm gave clustering sum of square error of (0.0572, 0.0341 and 0.04791).

The results show that DynamicK-reference clustering algorithm has a minimal Clustering sum of square error when compare with Fuzzy Artificial Bee Colony Optimization, KNN, PSO and ABC Optimization algorithms. This implies that DynamicK-reference clustering algorithm has the potentials to choose the best initial value of K clusters from the high dimensional dataset at a minimal error rate, with more accurate mechanism than the existing algorithms. This can be attributed to its robust and dynamic mechanism. It achieves an effective balance in multi-model problems, by separating the exploitation and exploration of the searching mechanism in the high dimensional space regions into distinct phases. Each phase is more focused on either exploitation (individual sub-swarms) or exploration (diversification method)" to obtain the best global particle in the high dimensional space of the huge dataset. The robust mechanisms of Dynamic Multi-swarm Optimization algorithm distinguished it from other meta-heuristic algorithms.

Moreover, the rest of the dataset (Hepatitis, Post-operative patient, Australian Credit Approval, German Credit Data, and Statlog Heart) was used to compare the Clustering accuracy and sum of squared error of the proposed Hybrid Clustering Algorithm and PSO based K-prototype algorithm [9].

The experimental results of DynamicK-reference clustering algorithm gave accuracy of (0.9669, 0.9889, 0.9610, and 0.9509) while PSO based K-prototype algorithm gave accuracy of (0.7521, 0.8229, 0.6261, and 0.8387). "The same Lambda values for the four benchmark datasets used for PSO based K-prototype algorithm experimental analysis [9] was also used to analyze the performance of the proposed system, to ensure equal comparison of the performance of both hybridization of Dynamic Multi-swarm and Binary particle Swarm Optimization Algorithms with k-prototype Algorithm [15]." The lambda values are shown in Table 1.

Table-1: Details of Datasets and Lamda Values

S/N	Hepatitis	Australian Credit Approval	German Credit Data	Statlog Heart
Lambda value	0.0533	0.0680	0.0995	0.0845

The graphical representation of the comparison of the clustering accuracy of the proposed DynamicK-reference and Binary swarm-based k-prototype clustering algorithm is shown in Fig. 5.

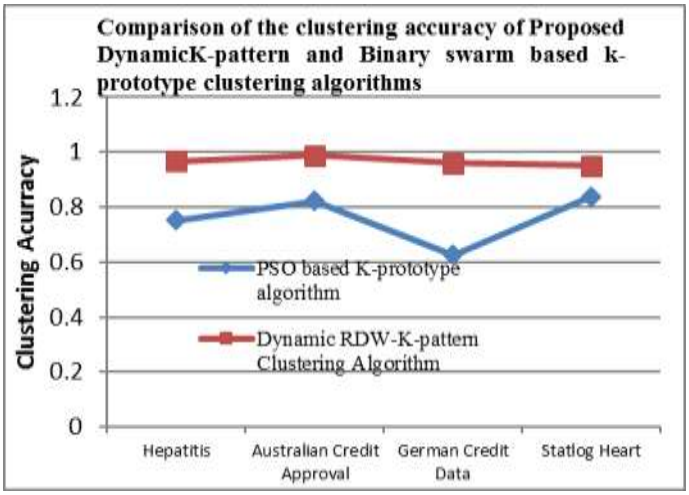

Fig- 5: Graphical representation of comparison of the accuracy values of DynamicK-reference and Binary swarm-based k-prototype clustering algorithms.

The result demonstrated that the proposed hybrid clustering algorithm is highly proficient for clustering very large mixed datasets. It also confirmed that DynamicK-reference clustering algorithm has better clustering accuracy more than the Binary Swarm based K-prototype clustering algorithm for the four datasets used. Furthermore, the Clustering sum of square error of Binary Swarm based k-prototype and the proposed DynamicK-reference clustering algorithm was compared. The result is shown in Fig. 6.

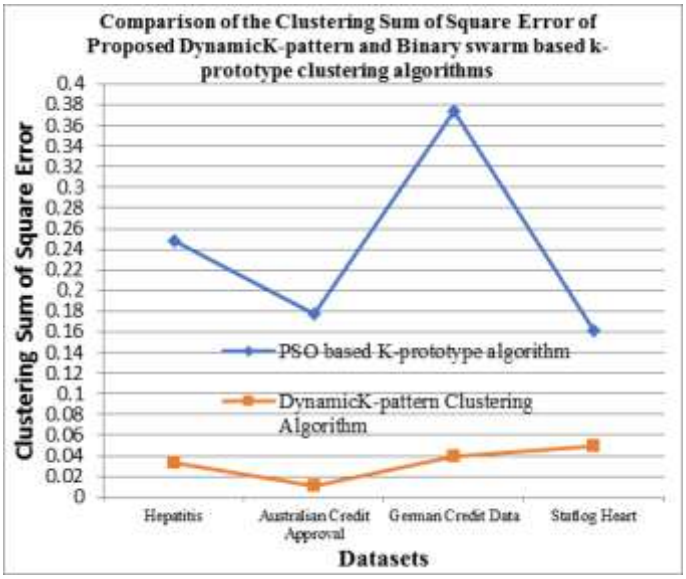

Fig-6: Graphical representation of comparison of the clustering sum of square error values of DynamicK-reference and Binary swarm-based k-prototype clustering algorithms.

The simulation results for DynamicK-reference clustering algorithm gave clustering sum of square error of (0.0331, 0.0111, 0.0390 and 0.0491) while for the existing algorithms, Binary particle Swarm Optimization Algorithm gave accuracy of (0.2479, 0.1771, 0.3739 and 0.1613). The results show that DynamicK-reference clustering algorithm has minimal clustering sum of square error than the Binary swarm based k-prototype clustering algorithm. Therefore, DynamicK-reference clustering algorithm has high analytic accuracy than the existing algorithms.

5. CONCLUSION

The results from this work shown that the integration of Swarm Intelligence and Unsupervised Machine Learning Algorithms provides efficient and robust mechanism to handle the uncertainties and analytical challenges posed by complex characteristics of big data on traditional Algorithms. In addition, this work had proved that the intelligent characteristic of Dynamic Multi-swarm agents and robust analytic behavior of Unsupervised Machine Learning Algorithms are capable of unveiling the hidden patterns of unstructured dataset of big data stream. The proposed hybridized clustering algorithm clustered the unstructured datasets of Big Data analysis efficiently. It possesses the potential to create clusters of similar internal structure of unstructured (mixed) datasets and produce accurate clustering interpretation for decision support system.

ACKNOWLEDGEMENT

We would like to thank and appreciate the Departments of Computer Science, Federal University of Technology, Owerri, Imo State, Nigeria for providing the enabling environment that facilitated this work.

REFERENCES

[1] Wu, X., Zhu, G. Q. and Ding, W: Data mining with Bigdata. IEEE Transactions on Knowledge and Data Engineering, vol 3, pp. 97-107, 2014.

[2] M. Imran, and V. S. Rao: A Novel Technique on Class Imbalance Big using Analogous under Sampling Approach. International Journal of Computer Application (2018), vol. 17, pp. 18-21, 2018

[3] A. Ali, and S. Site: A Simplified Analytical Model towards Big Data Analysis using Ridge Regression Method. International Journal of Computer Application, vol. 38, pp. 41-47, 2018

[4] M. S, Aishwaray, H. Bhargavi, and N. Jyothi: Handing Big Data Analytic Using Swarm Intelligent. International Journal of Scientific Development and Research (IJDR), Vol. 2, pp. 271-275, 2017.

[5] C. Shi, Z. Qingyu, and Q. Quande: Big data analytic with swarm intelligent, Industrial Management & Data Systems. 116(4), 646-666, 2016.

[6] Cao, H. Cui, and L. Jiao: Big Data: A parallel Particle Swarm Optimization-Back-Propagation Neural Network Algorithm Based on MapReduction. 11(6), 1-17, 2016.

[7] S. K. Kumar and M. Hemalatha: An Innovative Potential on Rule Optimization using Fuzzy Artificial Bee Colony. Research Journal of Applied Science, Engineering and Technology. 7(13): 2627-2633, 2014.

[8] H. Marzieh, and H. A. Ali, and J. M. A. Seyed: A new method of Fuzzy Clustering by using the combination of the firefly algorithm and the particle swarm optimization algorithm. WALIA Journal. Vol 31, No 3, pp. 245-252, 2015.

[9] A. K. Prabha and K. K. Visalakshi: Particle swarm optimization-based k-prototype clustering algorithm. IOSR Journal of Computer Engineering (IOSR-JCE). (2017), vol. 2, pp. 56-62, 2015.

[10] L. Diallo, A. Aisha-Hassan and F.R. Olanrewaju: Two Objectives Big Data Task Scheduling using Swarm Intelligence in Cloud Computing. Indian Journal of Science and Technology, 9(26), 1-10. 2016.

[11] S. Palak, and K. Vikas: Social Media Generated Big Data Clustering. International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India. 2-7, 2017.

[12] L.O. Mafteiu-Scai, Anew Dissimilarity Measure between Feature-Vectors. International Journal of Computer Applications, Vol.17. pp. 0975-8887, 2013.

[13] Z. Huang: Extensions to the k-means algorithm for clustering large cybernetics 28C, (1998), pp. 219-230, 1998.

[14] T. Backwell, and J. Branke: Multi-swarm Optimization in Dynamic Environments, Department of computing, Goldsmiths College, University of London New Cross, London SE14 6NW, U.K., 1-12, 2004.

[15] P. C. Oleji, N. E. Nwokorie and D. O. Onuodu, Clustering Mixed Dataset with Multi-Swarm Optimization and K-prototype Clustering Algorithm. Nigeria Computer Society (NCS) 26th National Conference & Exhibition; Information Technology for National Safety & Security, Abuja Nigeria, 27, 205-221., 2016.