ISSN 2278-3091
Volume 9, No.4, July – August 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse268942020.pdf
https://doi.org/10.30534/ijatcse/2020/268942020

Enhanced Evolutionary Computing Assisted K-Means Clustering Algorithm for BigData Analytics

Sarada B[1], Dr Vinayaka Murthy M[2], and Dr Udaya Rani V[3]
[1]School of C & IT, Research Scholar, Bangalore, India, saradasaikonda@gmail.com
[2]School of CS&A, Professor, Reva University, Bangalore, India, dr.m.vinayakamurthy@gmail.com
[3]School of C & IT, Associate Professor, Reva University, Bangalore, India, udayamurthy@yahoo.com

ABSTRACT

The exponential rise in internet technologies and allied applications has given rise to the technology named Big Data that intend to process gigantically huge data to assist real-time analytics or decision purposes. However, high pace increasing data heterogeneity, non-linearity, multi-dimensional features and unannotated data characteristics forces classical approaches to undergo huge computational overheads and limited accuracy that confines its suitability for major Big Data analytics purposes. With this motivation, in this paper a robust Big Data analytics model has been developed by incorporating Min-Max normalization, Dual Phased Feature Selection (DPFS) and enhanced Adaptive Genetic Algorithm (AGA) assisted K-Means clustering. Here, the use of Min-Max normalization helps alleviating key issues like data heterogeneity, data imbalance and pre-mature convergence during computation. Unlike classical feature selection approaches, DPFS exploited the efficacy of both Pearson correlation assisted significant test as well as T-Test analysis that ensure optimal feature selection for further computation. In addition, the use of AGA assisted K-Means clustering algorithm has accomplished computationally efficient and reliable clustering for efficient Big Data analytics purposes. Noticeably, the use of adaptive fitness sensitive GA parameter selection has strengthened our proposed system to exhibit better performance without imposing computational overheads. The computational efficacy of AGA-K-Means can strengthen MapReduce to be used for real-time Big Data analytics applications.

Key words: Big Data Analytics, Adaptive Genetic Algorithm, K-Means Clustering, Dual Phase Feature Selection.

1. INTRODUCTION

The exponential rise in internet technologies and allied applications has given rise to the gigantic data that requires optimal computing techniques to assist reliable and time efficient decision purposes. Emergence of cloud technologies and allied Big Data too has been playing decisive role for decentralized computation and decision processes. However, heterogeneous data types and gigantic volume make major existing approached confined to assure optimal performance. Increasing heterogeneity and complexity of the data has been forcing academia-industries to develop certain more accurate and time efficient Big Data computing platform. To deal with such scenarios, Hadoop and/or Apache Spark platforms have been playing vital role; however up surging voluminous and complex data nature demands better techniques. In other words, the classical MapReduce techniques are becoming confined, especially in terms of computation time and accuracy over large scale heterogeneous datasets. A traditional MapReduce method with classical classifiers such as K-Means clustering often imposes a computational overhead that eventually makes it limited for Big Data computation. On the other hand, majority of the classical machine learning methods are confined to group un-labeled data often called unannotated data which is common in current day Big Data computation scenario. In such cases, developing certain efficient and robust Big Data computing technology with enhanced data processing technologies are of paramount significance. This fact has revitalized academia-industries to develop more efficient and robust computing paradigm to accomplish reliable and time efficient decision process in Big Data [10].

Realizing the fact that in current day scenario where most of the data used to be unannotated data or unlabeled data, clustering based approaches can be vital. [4] Dierckens et al. (2017) found K-Means clustering efficient for major Big Data computations. Numerous researches Singh and Balabantaray [15] [3] too found that the computational efficacy of K-means makes it a potential solution for document clustering purpose and numerous Big Data computations. However, authors have found classical K-Means algorithm confined, especially under Big Data computing environment where it undergoes local minima and convergence [28]. Random cluster selection nature of the classical K-Means algorithm limits its efficacy [12]. It requires certain more efficient and robust K-Means clustering algorithm to cope up with emerging Big Data computation [12].

Realizing the fact that K-Means algorithm predominantly relies on the position of initial centers. To alleviate such issues, Qi[13] developed an enhanced K-Means algorithm using a hierarchical optimization principle originated by k* cluster centers (k* > k) to minimize the risk of randomly seeds selection-based clustering. In addition, authors proposed a cluster pruning concept to augment K-Means algorithm to reduce clusters to reduce search space for further computation. A similar effort was made by Sinha and Jana [17] who focused on performing automated cluster formation to cope up with Big Data analytics problems. Considering significance of distance metric in K-Means clustering, Niu [9] applied block function which collects instances as blocks to cluster attributes. Unlike Euclidean distance-based clustering authors applied Manhattan distance. A similar effort was made by Amorim [2] who used Minkowski algorithm to reduce noise sensitivity on the classification accuracy. However, this approach can't achieve optimal performance merely by optimizing distance information. Enabling K-Means with feature-sensitive clustering can enhance performance [6]. Frahling [6] used core-sets to cluster objects with similar set of points for multi-cluster scenarios. Akhtar [1] used data dimensional density feature to perform initial centroid selection in K-Means clustering. Saini [14] found that the sequential nature of the initial set of centers forces K-means++ algorithm to undergo adversaries especially for large scale data. To achieve better computation speed, Kanungo [11] developed k-means with k-dimensional tree (kd-tree) in which a k-d tree data structure was considered to form points in a k-dimensional space. Though, this approach was found time-efficient, its efficacy over a large-scale dataset seems unexplored, which is must for Big Data computation. Gheid [7] too recommended multiparty additive scheme for privacy preserved Big Data computation; however, this approach could be limited with heterogeneous datasets, which is common in Big Data or cloud applications. Author [5] found that parallel clustering concept can be better alternative for K-Means clustering, especially for Big Data analytics[3],[8]. In these approaches' authors introduced the concept of merging where data parallelization was performed by distributing data objects to all parallel processes. It also performed iterative centroid update by merging the centroid sets from each parallel process to form a final centroid set. This approach can be suitable for Big Data computation; however, at the cost of increased computation and hence demands more efficient computation to cope up with time as well as accuracy expectations. Authors [10],[16] have found that K-means algorithm can be efficient for MapReduce assisted Hadoop for Big Data analysis; however, requires alleviating issues like convex hull and the heel point solution to achieve proper centroid estimation. Authors [30] who focused on predicting stock market price by using Genetic Algorithm and datamining techniques.

Optimal centroid selection with huge non-linear, heterogeneous, and unannotated data with multi-dimensional feature is a highly complicate problem for which heuristic-based K-Means algorithm can be vital [16]. Towards such objective, different Evolutionary Computing (EC) algorithms such as Genetic Algorithm (GA), Particle Swarm Optimization (PSI), Artificial Immune System (AIS), Ant Colony Optimization (ACO), Imperialist Competitive Algorithm (ICA) [19] etc. have been developed. However, most of these algorithms employ fixed stopping criteria that forces algorithm to undergo huge computational overheads and time consumption [17],[18]. Here the Author [31] said that Hadoop solves the main problem of processing and storage. It can adversely affect real-time Big Data analytics performance [17]. In such cases, maintaining optimal trade-off of computational cost and efficiency is must. With this motive, in this paper a highly robust EC algorithm named Adaptive Genetic Algorithm (AGA) has been developed for K-Means clustering centroid estimation for Big Data analytics purposes. Unlike classical GA algorithm, our proposed AGA heuristic model applies dynamic GA parameter update such as the probability of crossover and the probability of mutation that eventually augments overall computation by reducing number of iterations and search space while alleviating the issue of local minima and convergence.

The use of AGA based K-Means can help K-Means (assisted) MapReduce to perform computation without imposing computational overheads or time. In addition to the enhanced K-Means clustering algorithm to achieve robust Big Data analytics, we have incorporated numerous enhancement such as pre-processing, feature extraction and selection that enables proposed Big Data analytics model to exhibit better performance. Considering data heterogeneity and non-linearity of the data elements, we have applied Min-Max normalization that helps key issues of data imbalance and pre-mature saturation during computation. In addition, Pearson correlation assisted significant has been used over normalized data that strengthens T-Test based feature selection to avoid computational overheads and time consumption. The use of DB-Index as fitness estimation method enables proposed AGA-K-Means based Big Data analytics to assist feature sensitive optimal centroid estimation. It strengthens proposed model to achieve higher accuracy for MapReduce based Big Data analytics purposes, where AGA-K-Means can help achieving higher performance without imposing computational overheads. The overall proposed system has been developed using Python programming language which was tested over numerous benchmark data. The efficiency of the proposed system has been assessed in terms of classification accuracy, precision, F-Measure, Recall, Specificity, balanced accuracy, heterogeneity, etc.

The remaining sections of the presented manuscript are divided as follows: Section II discusses our contribution and the proposed model, followed by results and discussion in Section III. Section IV presents overall conclusion and the references used in this paper are presented at the end of the manuscript.

2. OUR CONTRIBUTION

This section primarily discusses the proposed enhanced evolutionary computing assisted K-Means clustering model for Big Data analytics. In addition, the other enhancements such as data normalization, DPFS followed by AGA-K-Means clustering are discussed in this section.

Recalling the fact that the modern computational system especially Big Data computation which functions over gigantic datasets to assist transient decision purposes, requires certain highly robust and efficient algorithm. Towards this objective the use of K-Means clustering can be of utmost significance. However, classical K-Means clustering algorithms suffer varied limitations due to random initial cluster centre, here onwards called centroid selection. Therefore, strengthening Big Data technology with an enhanced K-Means algorithm can be vital to achieve higher accuracy, scalability and robustness for data clustering over large scale datasets. To achieve it, enhancing K-Means clustering algorithm with better centroid selection can be a potential solution. With this motivation, in this paper we have focussed on developing a novel and robust EC assisted K-Means clustering algorithm to strengthen Big Data MapReduce component for better performance. This research emphasizes on cumulative performance optimization where multi-level optimization measures have been incorporated to achieve optimal performance. Considering key issues of Big Data computation a few such as multi-dimensional data or features, unstructured or imbalanced datasets, heterogeneous data elements, curse of dimensionality, confined classifiers etc. in this research paper we focus on developing a multi-level optimization measure that could resolve aforesaid issues effectively. Towards these objectives, in this paper we have emphasized on incorporating better pre-processing algorithm, feature extraction, feature selection, enhanced K-Means clustering etc to assist Hadoop MapReduce computation for better Big Data analysis. Noticeably, in major literatures authors have recommended enhancing K-Means clustering centroid selection to achieve better performance. With this motive, unlike classical efforts, in this paper an enhanced EC approach named Adaptive Genetic Algorithm has been developed to assist better K-Means centroid estimation. This approach can significantly enhance classical K-Means to achieve low computation time, high accuracy and better scalability which can be of paramount significance for Big Data analytics.

The overall proposed EC assisted K-Means based BigData analytics model comprises the following steps:

- Step-1 Data Acquisition
- Step-2 Data pre-processing
- Step-3 Dual Phase Feature Selection (DPFS)
- Step-4 Enhanced EC assisted K-Means clustering algorithm for data classification.

Table 1 given below presents a snippet of the proposed methodological paradigm.

Table 1: Snippet of the implementation paradigm

Variables	Specifications
Datasets	Large scale multi-dimensional data with heterogeneous features/elements
Pre-processing	Outlier analysis Min-Max normalization
Feature Extraction/Selection	Pearson correlation assisted Significant test T-Test analysis DB-Index feature selection
Classifier	Enhanced EC assisted K-Means clustering
Performance assessment	Confusion metrics (Accuracy, precision, recall and F-Measure)

A snippet of the proposed Big Data analytics model is given in Figure 1.

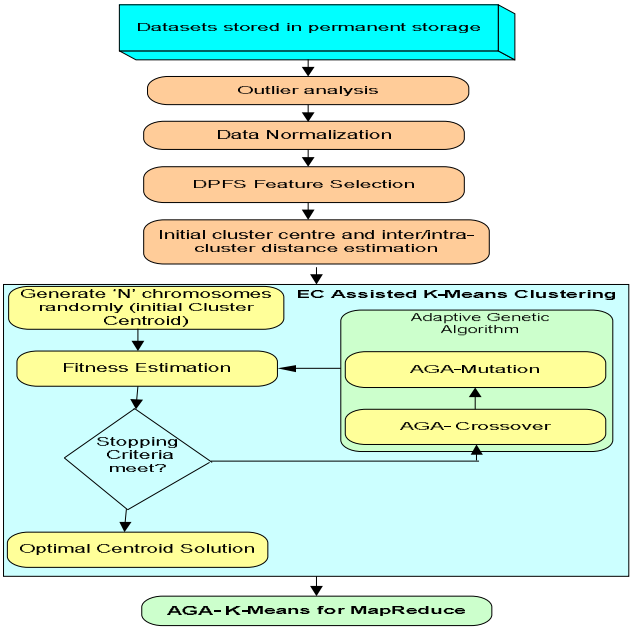

Figure 1: Enhanced AGA assisted K-means clustering for BigData analytics

A snippet of the pre-processing is given in the sub-sequent section.

A. Data Pre-processing

As already stated in above section to assess efficacy of the proposed Big Data analytics model we have considered data with heterogeneous features and multi-dimensional constructs.

Typically, in real time scenario, data with the above characteristics can be non-uniform and imbalanced that can adversely affect overall performance. Non-linearity of the data elements can force classifiers to undergo local minima and convergence. To alleviate such issues, data normalization can be vital and hence in this paper the obtained data after outlier analysis and tokenization has been processed for data normalization. In this paper Min-Max normalization algorithm has been applied that maps or normalizes input data elements in the range of 0 to 1. It performs linear transform over the input data and adjusts it in the defined [0-1] range that helps prohibiting any possible saturation during training and classification. Here, each data element x_i of the class X is mapped to the allied normalized value x'_i in the range of [0, 1]. Thus, the normalized value x'_i is obtained as (1).

$$Normaltzed(x_i) = x'_i = \frac{x_i - min(X)}{max(X) - min(X)} \quad (1)$$

In (1), the variable $\min(X)$ and $\max(X)$ signify the minimum and maximum values of X correspondingly.

In addition to the normalization, normalized data is processed for outlier analysis where "OLA-Analysis" has been applied to perform outlier analysis.

B. Dual Phase Feature Selection (DPFS)

Undeniably, Big Data datasets often encompass heterogeneous data elements distributed as multidimensional features or entities. Such scenario often exhibits the problem of "curse of dimensionality," signifying that the total features are much higher as compared to the total samples. On the other hand, there can be the possibility that a major fraction of data can have similar significance and allied features can be low influence on the classification result. In practice, these issues often influence the performance of Big Data analytic models.

Considering these issues, in this paper a Dual Phase Feature Selection (DPFS) technique has been developed. The proposed DPFS model applies Pearson Correlation based significant test and Independent Sample T-Test in sequence to retrieve most significant feature to perform further classification or analytics processes. In DPFS, at first the considered features have been tested Pearson correlation based significant test with correlation value of 0.5. In other words, in our proposed analytic model the retrieved data elements have been processed for significant test using Pearson correlation where the data elements with significant value of more than 0.5 has been taken into consideration for further analysis. This process enables reducing unwanted computation and helps achieving timely and accurate computation which is vital for Big Data analysis. Once performing significant test and obtaining the suitable data elements, we have applied a robust feature selection method named Independent Sample T-Test, often called T-Test analysis.

A snippet of the T-Test analysis-based feature selection is given as follows:

In our proposed Independent Sample T-Test Analysis (ISTA) method, we implement a commonly used filtering paradigm for data that employs univariate (regression) criteria on each data and/or features, if there is no significant relation between features. In our proposed method, we considered two-class problem with the Null Hypothesis (H_0) and the Alternate Hypothesis (H_1), where the first signifies that the two populations are equal, i.e.; there is no significant difference in between their mean values. In other words, it signifies that both features are same or homogeneous. Noticeably, similar features or homogeneity signifies that the features can have no significant effect on the overall classification or clustering and therefore can be reduced to make computation cost as well as time efficient. On contrary, the features possessing certain unique characteristics or differences can influence classification and therefore can be selected for further computation. With this motive, in this paper H_0 is rejected, while H_1 is accepted. Thus, the only features with significant differences or heterogeneity are considered for further classification. In our proposed method, we have applied T-Test analysis on each feature and the results have been compared with the P value, also called the absolute values of t-statistics for each feature as a measure of how significant it is for clustering the data elements or grouping the data.

Once performing the proposed DPFS feature selection, the selected data or feature set has been fed as input to the proposed enhanced EC assisted K-Means clustering algorithm to perform classification.

C. Enhanced EC assisted K-Means Clustering for Hadoop MapReduce

Recalling the fact that the classical K-Means clustering algorithms often undergo limited performance due to random initial cluster center selection or centroid selection, in this paper we have developed a novel EC assisted K-Means algorithm. Before discussing our proposed EC assisted K-Means clustering, a snippet of the classical K-Means algorithm is given as follows:

2.1 K-Means Clustering Algorithm

Classically, K-means clustering algorithm is an unsupervised learning method that performs grouping of the unlabeled data. K-Means algorithm emphasizes on finding groups in the data, where the total number of groups is indicated by a term K. Functionally, it operates iteratively to assign each data element to one of the K groups based on the features provided.

Thus, the overall data elements are clustered based on the respective feature similarity. This method results centroid for each cluster which signifies the collection of feature values defining resulting groups and therefore the selection of centroid does have decisive role for achieving accurate groups. Functionally, being an iterative refinement-based approach K-Means takes the number of clusters K and the data set as the input and exploits features of each data element that helps grouping of the data elements based on respective similarity. It

starts with initial estimates for the K centroids, which is typically assigned randomly from the data set. The overall process of K-Means clustering can be visualized in two sequential phases, data assignment and centroid update. A snippet of these phases is given as follows:

Step-1 Data Assignment

In this phase, each centroid represents one of the clusters formed. In this method, each element is assigned to its nearest centroid based on the squared Euclidean distance. Consider, c_j be the centroids in set C, then each data element x is allocated to a cluster based on the following conditions

$$\underset{c_j \in C}{arg\,min}\, dist(c_j, x)^2 \quad (2)$$

In Eq. 2 the component $dist(.)$ refers the standard (L_2) Euclidean distance. Now, consider the set of data points assigned to each $i-th$ cluster centroid be s_i, the centroid is estimated using following approach.

Step-2 Centroid Update

In this phase, the centroids are updated iteratively by exploiting the mean value of all data elements assigned to that centroid's cluster. Mathematically,

$$c_i = \frac{1}{|s_i|}\sum_{x_j \in s_i} x_i \quad (3)$$

This process continues until certain predefined stopping criterion is met.

Noticeably, the stopping criterion can be the number of data elements change clusters, maximum iterations, the sum of the distances is minimized etc. Considering the need of an optimized K-Means clustering algorithm for Big Data computation, in this paper an enhanced EC algorithm named Adaptive Genetic Algorithm (AGA) has been developed. The proposed AGA-K-Means model achieves optimal centroid for each cluster, and thus enables most accurate grouping of the (un annotated) data elements.

2.2 AGA-K-Means Clustering Algorithm

In AGA-K-Means clustering can be stated as a stochastic model maintaining the data points as population called the chromosomes signifying the population $P = \{Ch_1, Ch_2, \dots, Ch_p\}$. Here, each individual states a solution or sub-solution to the at hand clustering problem. Each solution Ch_i is obtained to estimate its "fitness value" that helps identifying optimal or superior solution. Once identifying optimal solution with suitable fitness value is used as centroid. In case of inappropriate solution, AGA-K-Means introduces new population to achieve solution with better fitness value and efficiency. Unlike classical GA algorithm where static parameters (crossover and mutation parameters) are used to perform solution retrieval, in this paper Adaptive GA parameter update model is developed.

Some of the key functions of AGA-K-Means clustering are discussed as follows:

Binary Data Presentation

In the proposed method, each data elements of the considered dataset are considered as the population for centroid's candidates. In our proposed method the length of chromosomes is equal to the size of the dataset, where i^{th} gene of the chromosome stated i^{th} data elements in the dataset. In proposed AGA-K-Means algorithm for a data element i to be the centroid, the i^{th} gene is labeled as "1"; otherwise "0". In AGA-K-Means the total number of clusters, K is chosen in the range of K_{min} to K_{max}, where K_{min} equals 2, while K_{max} can be $l/2$ or $\sqrt{l}$, where l signifies the length of chromosome. A snippet of the population initialization is given as follows:

Population Initialization

Let, S_p be the population size and Ch_p be the population having p chromosomes, (i.e., $p = 1, 2, \dots, S_p$). In our proposed model a positive, integer value K_p is arbitrarily selected from the range of $[K_{min}, K_{max}]$ and the gene connected to the index of the selected data elements is assigned "1" and rest as "0". Once deploying chromosomes (signifying sub-solution for cluster's centroid), the fitness value for each chromosome is obtained. A snippet of the applied fitness value estimation of objective function is given as follows:

Fitness Function Estimation

Undeniably, being an iterative calculation assisted stochastic method; AGA applies Objective Function which is required to be achieved to yield optimal centroid as solution. Towards these objectives we have estimated fitness value of each chromosome. In this paper we have estimated DB-Index value to derive Fitness Value of each chromosome.

DB Index based Fitness Estimation

In the proposed method to introduce parallel K-Means algorithm the considered dataset has been split into different small sub-sets based on the two key features; homogeneity within the clusters and the heterogeneity within different clusters. To corroborate the clusters, we have applied DB index that provides effective solution to validate the cluster constituted. The dispersion measure of a cluster $C_i, i = 1, \dots, K_p$, existing in the chromosome set Ch_p is retrieved using equation (4).

$$S_{i,q} = \left(\frac{1}{|C_i|}\sum_{x \in C_i} \|x - z_i\|_2^q\right)^{1/q} \quad (4)$$

In Eq. 4, the variable $S_{i,q}$ states the dispersion measure between data element i and probable centroid candidate q. The other variable z_i signifies cluster centroid of i-th cluster C_i.

Now, we estimate the maximum intra-cluster similarity index (ICSI) that states the similarity in between i-th cluster C_i with other cluster is obtained using equation (5). Mathematically,

$$R_{i,qt} = \max_{j,j \neq i} \left\{ \frac{S_{i,q} + S_{j,q}}{d_{ij,t}} \right\} \quad (5)$$

Unlike classical approaches in this paper we have estimated distance vector $d_{ij,t} = d(C_i, C_j)$ named Minkowski distance between C_i and C_j. In the denominator of Eq. 4, the variable t signifies the order. Thus, the DB index, DB_p of a chromosome Ch_p can be obtained using (6).

$$DB_p = \frac{1}{k_r} \sum_{i=1}^{k_r} R_{i,qt} \quad (6)$$

In our proposed AGA-K-Means clustering algorithm, the fitness value of each chromosome has been obtained using (7).

$$Fitness(Ch_p) = \frac{1}{DB_p} \quad (7)$$

To assist computationally efficient environment, we have assigned q and t as 1 and 2, correspondingly. Once, estimating fitness value of each chromosome the chromosomes with higher value is selected for the next generation and assessed for its suitability as centroid. In case of insufficient fitness value, we initiate Crossover process. A snippet of the crossover process is given as follows:

Typically, crossover process also called reproduction intends to assure that the at hand population or the chromosome possesses better solution or fitness as compared to the previous one. In our proposed method, we do hypothesize that the chromosome with higher fitness value can provide better solution. Unlike classical crossover approaches we have applied "Winner Substitution" followed by the roulette wheel selection. This approach enables each chromosome to have higher probability to be considered as potential solution.

Our proposed method generates new chromosomes Ch_{new} by augmenting the chromosome having the maximum fitness value in such approach that each centroid candidate is replaced by the data point closest to the allied mean center. If Ch_{new} has the higher fitness than Ch_j, then Ch_j is replaced by Ch_{New}. This approach enables time efficient as well as cost effective computation. Unlike classical GA in this paper an enhanced GA called AGA has been developed that performs dynamic GA parameter update that eventually makes overall computation highly efficient and robust for Big Data analysis purposes.

Adaptive GA Parameter update

The proposed EC-K-Means algorithm applies dynamic or adaptive genetic parameters (i.e., P_c and P_m) that avoids the issue of over-fitting, under-fitting, local minima and convergence. One of the most efficient and noticeable contribution of our proposed model is that unlike traditional GA algorithm where the number of iterations is used as stopping criteria, where EC-K-Means considers adaptive update approach. It is supported by the fact that, during training there can be multiple chromosomes having similar fitness value that could force K-Means to undergo local minima and convergence where it can show updated centroid as any random outcome. Considering this fact, in this paper AGA-K-means estimates the number of chromosomes having similar fitness value which is further used to update P_c and P_m dynamically (8).

$$(P_c)_{k+1} = (P_c)_k - \frac{C_1 * N_{NCF}}{7} \quad (8)$$

$$(P_m)_{k+1} = (P_m)_k - \frac{C_2 * N_{NCF}}{7}$$

In Eq. 8, the variables $(P_c)_{k+1}$ and $(P_m)_{k+1}$ signify the updated crossover and mutation probability, while its current probability is given by $(P_c)_k$ and $(P_m)_k$. Furthermore, we have considered static values for the coefficient parameters C_1 and C_2 as 0.1 and 0.01, respectively. The parameter N_{NCF} signifies the total chromosome having similar fitness value. In such manner, developing AGA based K-Means we performed centroid estimation. Thus, the developed model accomplishes not only optimal efficiency but also computational efficacy as well as reliability. Noticeably, in classical GA algorithms the "Number of Iterations" or "Generations" is given as the stopping criteria that often creates voluminous search space which leads huge computational overhead [28] [29]. Unlike such approaches, in our proposed AGA model, the stopping criterion (Figure 1 and Figure 2) has been derived as a dynamic variable characterizing a condition when 95% of the chromosomes have the same fitness value. If the number of chromosomes with the same fitness value exceeds 95%, it may cause over-saturation and hence can affect performance. Therefore, reaching this condition AGA is terminated and with the optimally retrieved centroid value, AGA-K-Means performs classification. Figure 2 depicts the implementation schematic of AGA-K-Means assisted Hadoop-MapReduce Big Data analytics model.

The implementation of the proposed AGA –K-Means clustering algorithm for Hadoop MapReduce is discussed in the sub-sequent sections.

2.3 MapReduce with AGA-K-Means Clustering

In AGA-K-Means based MapReduce implementation, the dataset after pre-processing used to be in numeric value signifying vector values of each data input. Further, the DPFS method is applied over the processed data inputs (i.e., numeric vector) it results into reduced vector values representing optimal dataset with significant information.

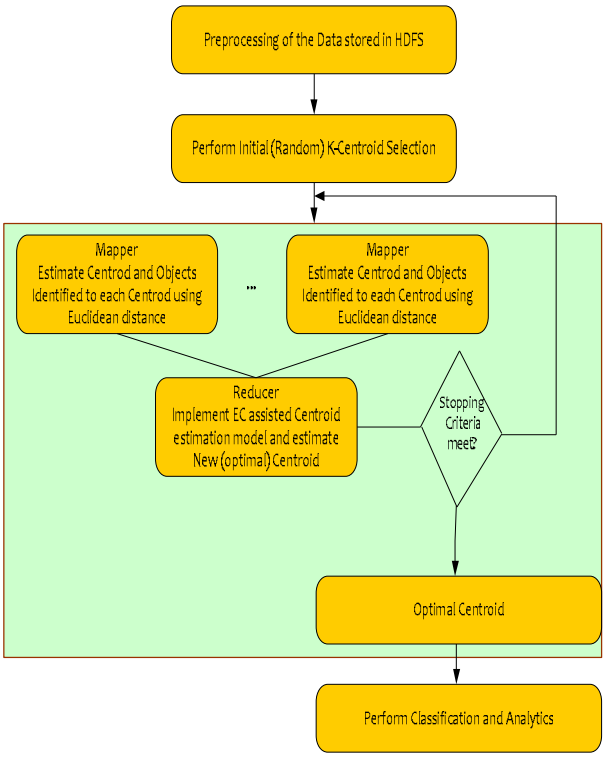

Figure 2: AGA-K-Means assisted MapReduce for Big Data Analytics

The retrieved value is then given as input to the AGA-K-Means algorithm with initial value of cluster k. Here, k number of vector values, centroid (selected randomly from the dataset), can be given as input a table named centroid table or centroid data and the input data can be divided across the slave nodes, where centroid data can be assigned or copied to each slave node by the HDFS. Thus, the proposed AGA-K-Means can be executed in parallel over certain defined (Hadoop) cluster size or nodes. For implementing AGA-K-Means algorithm with MapReduce, overall process can be split into two phase, first centroid estimation (and allied data element assignment) and second centroid update. In first step it executes iterative distance estimation (between centroids and dataset objects or data elements) to assign each data element to the closest centroid. In second step, centroids can be updated once data elements are assigned after (each) iteration.

Thus, the proposed AGA-K-Means algorithm was applied in such manner that the map achieves the job of assigning each object to the nearest center while the reduce job accomplishes the procedure of updating the new cluster centers, until it remains unchanged. Overall implementation schematic is given in Figure 3.

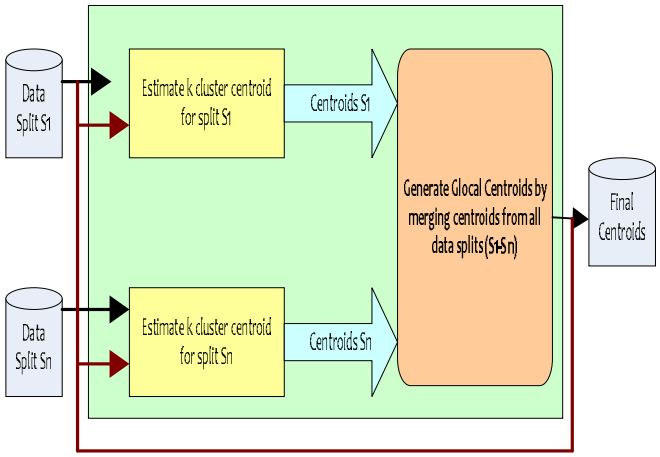

Figure 3: AGA-K-Means iterations in Hadoop-MapReduce cluster.

3. RESULTS AND DISCUSSIONS

The results obtained, and their inferences are also discussed in this section. The overall algorithmic innovation and optimization made in this paper emphasized on developing a robust Big Data computation environment that could deal with multi-dimensional features, large scale data with heterogeneity and non-linearity, and computational efficacy. To achieve it, enhancements were made at each preliminary phase of Big Data computation, such as data pre-processing, feature selection and classification. Towards these goals, outlier detection approach followed by Min-Max normalization was applied to achieve normalized dataset. Min-Max normalization mapped overall datasets in the range of [0-1] and thus avoids pre-mature convergence and saturation. Similarly, unlike classical feature selection approaches this research work applied DPFS where the normalized data elements were at first processed for significant test using Pearson coefficient values of the data elements. Here, we assigned threshold Pearson coefficient (p) or significant level as 0.5, i.e., the data elements with $p < 0.5$, were removed from the datasets and thus helped achieving suitable features for further computation. In other words, a data element with $p \geq 0.5$ is selected for further analytics computation.

This process can be stated as dimensional reduction measure that exploits significant level of each element towards targeted analytics purposes. Once performing Pearson coefficient-based feature selection, the data elements (with $p \geq 0.5$) were processed by means of a statistical feature selection approach called T-Test analysis that exploits homogeneity and heterogeneity information about the data elements to perform feature selection. Thus, the selected features were fed as input to the proposed AGA-K-Means clustering algorithm. Noticeably, in Hadoop-MapReduce computing environment the selected feature can be stored in HDFS, which can then be fetched to the MapReduce to perform clustering based data classification (Figure 3).

Unlike classical K-Means clustering algorithm, we applied an enhanced EC concept named AGA for K-Means clustering centroid estimation, which is highly related to the classification accuracy and hence the efficacy of Big Data analytics model. As depicted in Figure 3, the proposed AGA-K-Means clustering algorithm can function in conjunction with MapReduce component to perform data grouping or classification. Unlike conventional GA algorithm where GA parameters (crossover and mutation probabilities) are assigned as fixed value, in this paper it was updated dynamically. However, the initial crossover and mutation probabilities were assigned as $P_c = 0.8$ and $P_m = 0.2$, respectively. AGA-K-Means clustering was initiated with K=3. Undeniably, examining efficacy of any Big Data analytics model with merely one set of features or data elements seems confined. Therefore, assessing an algorithm with diverse feature sets, different attribute sizes, different classes etc can be vital to characterize a Big Data analytics model. With this motive, in this research the proposed model was tested with different datasets comprising large data elements or attributes, features and dimensions etc. A snippet of the data considered in this research is given in the sub-sequent sections.

3.1 Descriptive Assessment

Before examining efficacy of the proposed Big Data analytics model, a descriptive assessment of the data under study is given in this section.

Table 2 presents a brief statistical representation of the considered datasets in this study. Recalling the research intend to design a robust Big Data analytics model that could deal with big size datasets with large scale instances, different features and multiple classes, the data under study and respective performance assessment can be justifiable for (proposed) system generalization. In this research different datasets comprising Iris [20], Pendigits [21], BNG Spambase[22], BNG Mushroom [23],BNG Satellite Image[24] , Musk[25], Bank Marketing[26] and BNG Anneal[27] have been considered. Table 2 presents the number of instances, the number of features and allied number of classes for each dataset.

Table 2: Data description

Datasets	No. of Instances	No. of Features	No. of Classes
Iris	150	5	3
Pendigits	10992	17	2
BNG Spambase	4601	58	2
BNG Mushroom	1000000	23	2
BNG Satellite	1000000	37	6
Musk	6598	166	2
Bank Marketing	4521	17	2
BNG Anneal	1000000	39	6

Observing the datasets considered for this study, it can be found that the different datasets possess diverse statistical features such as number of instances, features and classes that can have significant impact on the efficacy of any classification or grouping algorithm. Therefore, assessing the performance of the proposed algorithm or computing model under such data-diversity can be of great significance that can help generalizing the proposed approach for different Big Data analytics purposes. The detailed discussion of the simulation results obtained is given as follows.

3.2 Performance Assessment

To examine performance efficiency of the proposed Big Data analytics model confusion matrix has been obtained in terms of true positive (TP), true negative (TN), false positive (FP) and false negative (FN). Thus, employing these matrix values, performance has been assessed in terms of accuracy, precision, f-measure, recall, specificity, heterogeneity, correctness, completeness, balanced accuracy etc.

Table 3 presents the definitions of these performance variables.

Table 3: Performance parameters

Parameter	Mathematical Expression	Definition
Accuracy	$\frac{(TN+TP)}{(TN+FN+FP+TP)}$	Signifies the proportion of predicted fault prone modules that are inspected out of all modules.
Precision	$\frac{TP}{(TP+FP)}$	States the degree to which the repeated measurements under unchanged conditions show the same results.
F-measure	$2.\frac{Recall.Precision}{Recall+Precision}$	It combines the precision and recall numeric value to give a single score, which is defined as the harmonic mean of the recall and precision.
Recall or Sensitivity	$\frac{TP}{(TP+FN)}$	It indicates how many of the relevant items are to be identified. The portion of the data set that tested positive out of all the positive tested The probability that the test is positive
Specificity	$\frac{TN}{(TN+FP)}$	Focus on how effectively a classifier identifies the negative labels.
Balanced Accuracy	$\left(\frac{\text{Recall} + \text{Specificity}}{2}\right)$	Accuracy under imbalanced computing environment such as different training and testing sets or cross-validations

Table 4 presents the performance parameters and their respective values.

Table 4: Performance analysis

Dataset	Accuracy	Precision	F-Measure	Recall/ Sensitivity	Specificity	Completeness	Correctness	Heterogeneity	Balanced Accuracy
Iris	0.886	0.895	0.923	**0.954**	0.707	0.594	0.122	1.905	**0.830**
Pendigits	0.860	0.861	0.848	0.899	0.751	0.396	0.651	4.390	0.825
BNG Spambase	0.906	**0.928**	0.891	0.887	0.772	0.714	0.714	4.820	0.829
BNG Mushroom	**0.929**	0.899	0.921	0.912	0.858	0.810	0.799	4.210	**0.885**
BNG Satellite	0.899	0.837	**0.965**	0.901	0.834	**0.859**	**0.883**	**6.143**	0.867
Musk	0.901	0.882	0.863	0.899	0.852	0.837	0.852	4.681	0.875
Bank Marketing	0.882	0.841	0.799	0.883	0.809	0.430	0.683	4.010	0.846
BNG Anneal	0.909	0.873	0.929	0.899	**0.864**	0.772	0.871	4.19	0.882

Once performing the classification, confusion matrix has been obtained for the specific data that eventually estimates the performance values as presented in Table 3 and Table 4.

An illustration of the results obtained, and allied calculation is given as follows:

Iris data which comprises a total of 150 instances and can be classified into three distinct feature sets, Setosa, Versicolor, and Virginica, each containing 50 instances. Simulating the developed algorithm and performing classification, we get the values of the different confusion matrix components, like TP=104, TN=29, FP=12 and FN=5. Now putting these obtained values in the equations as mentioned in Table 3, we get the following:

$$Accuracy = \frac{(29+104)}{(29+5+12+104)} = 0.8866$$

$$Precision = \frac{104}{(104+12)} = 0.8956$$

$$Recall = \frac{104}{(104+5)} = 0.9541$$

$$F-Measure = 2.\frac{0.9541 \times 0.8956}{0.9541+0.8956} = 0.9239$$

$$Specificity = \frac{29}{(29+12)} = 0.7071$$

$$\text{Balanced Accuracy} = \left(\frac{0.9541+0.7071}{2}\right) = 0.8307$$

Noticeably, converting above outcomes into percentile, we get accuracy of 88.6%, precision 89.5%, recall 95.41%, specificity 70.7% and balanced accuracy 83%.

Big Data analytics is hypothesized to be performing over a large-scale dataset (i.e., VOLUME) with more features (VERIETY) while assuring swift computation (VELOCITY). However, in real time computation it might require exhibiting consistent or reliable performance over small as well as large scale datasets. With the motive to assess performance of AGA-K-Means clustering algorithm we have examined its performance (efficacy) for Iris (with 150 instances and three class problem) as well as BNG Mushroom/Satellite/Anneal which contains 1,000,000 instances each with 2, 6 and 6 class classification problem, respectively. Observing overall performance, it can be found that the proposed algorithm can be of utmost significance for Big Data analytics purposes irrespective of size or dimensionality. It shows robustness of the proposed model deal with real-time computation problems.

To assess efficiency of the proposed AGA-K-Means based MapReduce model over classical K-Means algorithm, we have compared its performance in terms of accuracy. Noticeably, considering space constraints in this manuscript, we have compared aforesaid algorithms in terms of only "accuracy". The comparative performance for the different datasets is given in Table 5.

Table 5: Comparative accuracy performance

Dataset	Accuracy	
	K-Means	AGA-K-Means
Iris	0.794	0.886
Pendigits	0.777	0.860
BNG Spambase	0.668	0.906
BNG Mushroom	0.715	0.929
BNG Satellite	0.718	0.899
Musk	0.763	0.901
Bank Marketing	0.663	0.882
BNG Anneal	0.692	0.909

Observing the results obtained in Table 5, it can easily be found that the proposed AGA-K-Means algorithm outperforms classical K-Means clustering algorithm for classification. Noticeably, the performance by K-Means algorithm, especially for BNG Mushroom/Spambase/Anneal it can be found that it performs inferior as compared to our proposed AGA-K-Means clustering algorithm. It justifies superiority and efficiency of our proposed model over classical K-Means clustering algorithm.

4. CONCLUSION

Considering the significance of a robust and efficient Big Data analytics model for modern day computation, this research work primarily focused on making effort by incorporating enhancement at each step of the processing comprising pre-processing, feature selection and classification. Noticeably, this research made efforts to alleviate major issues like data heterogeneity and imbalance, multi-dimensional features, computational cost and complexity; and different limitations of the classical clustering approaches such as K-Means algorithm. Considering heterogeneous data elements, which is common in modern internet-based technologies at first outlier detection and Min-Max algorithm-based normalization was performed that significantly alleviates the issue of data imbalance and premature convergence.

The use of DPFS model that sequentially combined Pearson Correlation based significant test followed by T-Test analysis method for feature selection strengthened the proposed method to achieve computational efficiency. Realizing the fact that K-Means clustering method is one of the most efficient approach for grouping unlabeled data or un annotated data, in this research it was considered as a classifier for data analytics. However, considering random initial centroid based existing approaches and their limitations, in this paper an enhanced evolutionary computing method called Adaptive Genetic Algorithm (AGA) has been developed to enhance K-Means clustering algorithm for efficient Big Data classification and allied analytics. The use of AGA has enabled optimal centroid selection that not only enhances classification accuracy but also ensures minimum computational overheads and allied costs. Thus, the overall developed model can play vital role in Big Data analytics purposes to achieve optimal performance. Predominantly, the use of AGA- K-Means clustering algorithm can enhance MapReduce to achieve higher computational efficiency even with gigantic datasets with multidimensional and heterogeneous features. The simulation-based performance assessment too has revealed that the proposed Big Data analytics method can be vital for real-time computation for reliable, scalable and highly accurate decision purposes.

REFERENCES

1. Akthar N., Ahamad M. V., S. Khan. 2015**. Clustering on Big Data Using Hadoop MapReduce.** *International Conference on Computational Intelligence and Communication Networks (CICN)*, Jabalpur, pp. 789-795.
2. Amorim R. and Mirkin B. 2012. Minkowski Metric, **Feature Weighting and Anomalous Cluster Initializing in K-means Clustering**. Pattern Recogn; 45(3):1061.
3. Balabantaray, Chandra R., Sarma C, Jha M. 2015. **Document Clustering using K-means and K-medoids**. arXiv preprint arXiv:1502.07938.
4. Dierckens K. E., Harrison A. B., Leung C. K., Pind A. V. 2017. **A Data Science and Engineering Solution for Fast K-Means Clustering of Big Data**. *IEEE Trustcom/BigDataSE/ICESS*, Sydney, NSW, 2017, pp. 925-932.
5. Fahad A., Alshatri N., Tari Z., Alamri A., Khalil I., Zomaya Albert Y. 2014. **A Survey of Clustering Algorithms for Big Data: Taxonomy and Empirical Analysis**. IEEE Trans Emerg Top Comput 2014;2(3):267.
6. Frahling G. and Sohler C. 2006. **A Fast K-means Implementation using Coresets**. In: Proceedings of the twenty-second annual symposium on computational geometry; p. 135-43.
7. Gheid Z. and Challal Y. 2016. **Efficient and Privacy-Preserving k-Means Clustering for BigData Mining**. *IEEE Trustcom/BigDataSE/ISPA*, Tianjin, 2016, pp. 791-798.
8. Garg, Dweepna, Gohil P., and Trivedi K. 2014. **Modified Fuzzy K-mean clustering using MapReduce in Hadoop and cloud**. IEEE International Conference on Electrical Computer and Communication Technologies (ICECCT).
9. K. Niu., Z. Gao., H. Jiao., N. Deng. 2016. K-means+: **A Developed Clustering Algorithm for BigData**. *4th International Conference on Cloud Computing and Intelligence Systems (CCIS)*, Beijing, pp. 141-144.
10. Jing Z., Wu G., Hu X., Li S., Hao S. 2013. **A Parallel Clustering Algorithm with Mpi-m K-means**. J Comput;8(1):10.
11. Kanungo T., Mount D., Netanyahu N., Piatko C., Silverman R., Wu A. 2002. **An Efficient K-means Clustering Algorithm: Analysis and Implementation**. IEEE Trans Pattern Anal Mach Intell;24(27):881.
12. Kang, Ji S., Yeon Lee S., Lee Keon M. 2015. **Performance Comparison of Open MPI, and MapReduce in Practical Problems**. AdvMultimed.
13. Qi J., Yu Y., Wang L., Liu J. 2016. **K*-Means: An Effective and Efficient K-Means Clustering Algorithm.** IEEE International Conferences on Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom) (BDCloud-SocialCom-SustainCom), Atlanta, GA, pp. 242-249.
14. Saini A., Minocha J., Ubriani J., Sharma D. 2016. **New approach for clustering of big data: DisK-means**. *International Conference on Computing, Communication and Automation (ICCCA)*, Noida, pp. 122-126.
15. Singh H. 2016. **Clustering of Text Documents by Implementation of K-means Algorithms**. Streamed Info Ocean January-June;1(1).
16. Singh, Kumar V., Tiwari N., Garg S. 2011. Document **Clustering using K-means, Heuristic K-means and Fuzzy C-means**. In: Computational intelligence and communication networks (CICN), international conference on. IEEE.
17. Sinha A. and Jana P. K.. 2016. **A novel K-means Based Clustering Algorithm for BigData**. International Conference on Advances in Computing, Communications and Informatics (ICACCI), Jaipur, pp. 1875-1879.
18. Wu K., W. Zeng., T. Wu., Y. An. 2015. **Research and Improve on K-means Algorithm based on Hadoop.** 6th IEEE International Conference on Software Engineering and Service Science (ICSESS), Beijing, pp. 334-337.
19. Zadeh M. R. D., Fathian M., Gholamian M. R. 2014. **A New Method for Clustering Based on Development of Imperialist Competitive Algorithm**. In *China* Communications, vol. 11, no. 12, pp. 54-61, Dec. 2014.
20. https://www.openml.org/d/451
21. https://www.openml.org/d/1019
22. https://www.openml.org/d/135
23. https://www.openml.org/d/120
24. https://www.openml.org/d/1183
25. https://archive.ics.uci.edu/ml/datasets/Musk+(Version+2)

26. https://www.openml.org/d/1558
27. https://www.openml.org/d/244
28. Hongcheng T., 2012 **An Improved Adaptive Genetic Algorithm.** In: Tan H. (eds) Knowledge Discovery and Data Mining. Advances in Intelligent and Soft Computing, vol 135. Springer, Berlin, Heidelberg.
29. Derigs U., Kabath M., Zils M. 1999. **Adaptive Genetic Algorithms: A Methodology for Dynamic Autoconfiguration of Genetic Search Algorithms**. In: Voß S., Martello S., Osman I.H., Roucairol C. (eds) Meta-Heuristics. Springer, Boston, MA
30. M.Tawarish and Dr. K. Satyanarayana. **A Review on Pricing Prediction on Stock Market by Different Techniques in the Field of Data Mining and Genetic Algorithm.** *International Journal of Advanced Trends in Computer Science and Engineering*, Volume 8, No.1, pp. 23-26,2019.https://doi.org/10.30534/ijatcse/2019/01812019
31. Danish Ahamad, MD Mobin Akhtar and Shabi Alam Hameed. **A Review and Analysis of Big Data and MapReduce**. *International Journal of Advanced Trends in Computer Science and Engineering*, Volume 8, No.1, pp. 1-3, 2019. https://doi.org/10.30534/ijatcse/2019/01812019