Optimized K-Means Clustering Model based on Gap Statistic

Amira M. El-Mandouh[1], Hamdi A. Mahmoud[3]
Beni-Suef University
Cairo Egypt

Laila A. Abd-Elmegid[2], Mohamed H. Haggag[4]
Helwan University
Cairo Egypt

***Abstract*—Big data has become famous to process, store and manage massive volumes of data. Clustering is an essential phase in big data analysis for many real-life application areas uses clustering methodology for result analysis. The data clustered sets have become a challenging issue in the field of big data analytics. Among all clustering algorithm, the K-means algorithm is the most widely used unsupervised clustering approach as seen from past. The K-means algorithm is the best adapted for deciding similarities between objects based on distance measures with small datasets. Existing clustering algorithms require scalable solutions to manage large datasets. However, for a particular domain-specific problem the initial selection of K is still a significant concern. In this paper, an optimized clustering approach presented which is calculated the optimal number of clusters (k) for specific domain problems. The proposed approach is an optimal solution based on the cluster performance measure analysis based on gab statistic. By observation, the experimental results prove that the proposed model can efficiently enhance the speed of the clustering process and accuracy by reducing the computational complexity of the standard k-means algorithm which achieves 76.3%.**

***Keywords*—Big data; mapreduce; k-means; gap statistic**

I. Introduction

Cluster analysis is a vital exploratory mechanism widely applied in many fields such as biology, sociology, medicine, and business. Clustering aims to group a set of data items, known as data points, into similar clusters [1]. The process examines the similarity between various data points according to some distance measure. The main idea is to put in one cluster the points that have the least distance from one another. Accordingly, different points in different groups have a larger distance from each other [2]. There are three main types of clustering techniques; Distance-based, Density-based, and hierarchical.

K-means, proposed by MacQueen, is an unsupervised learning distance-based algorithm [3]. It is the famous used algorithm for cluster analysis. It considers a simple, easy, and recursive procedure to assign the data points into clusters according to the specified similarity measurement. The main feature of k-means is the linear complexity of both time and space. Additionally, it has many variants characterized as disk-based as they do not require the existence of all data points in memory [4].

In the K-Means clustering algorithm based on Euclidean distance which measures the similarity, the k data objects farthest from each other are more representative than the k data objects randomly selected [5][6]. It is a process to organize the specified objects into a group of classes called clusters. It had calculated similarities among objects for specific criteria. It solves the well-known clustering problem by considering certain attributes and performing an iterative alternating fitting process. In each iteration, the distance was calculated which causes the low algorithm efficiency and high consuming time. It introduced a simplified data structure to save some details in each iteration and utilized this information in the next iteration. The proposed method does not demand to calculate the distance of each data point from each cluster center in each iteration due to which running time of the algorithm is reduced.

Estimating the cluster's number is a critical difficulty in cluster analysis processing, which is taken as a beginning in almost clustering techniques. It would most possibly recover the underlying cluster structure given a reasonable guess of the correct number of cluster.

The distance metric plays a vital role in clustering techniques. A distance metric is a function which represents a distance within instances of a dataset. It gets a similarity of data objects by using distance metrics which lead to developing robust data mining algorithms. A set with a metric is known as metric space [7]. The various methods are available for clustering like Euclidean Distance, Manhattan distance, Chebychev Distance, Minkowski Distance.

The rest of the paper is structured as follows. In Section II, related research work is discussed. Whereas, the basic concepts of Map reduce and Gap statistic, utilized in the proposed approach, are presented in Section III. The proposed approach is presented in Section IV. The efficiency of the proposed approach is proved in the experimental study given in Section V. Finally, the conclusion of the proposed work is introduced in Section VI.

II. Related Work

One of the critical issues of cluster analysis is expecting the optimal number of clusters suitable to the processed data set [8].

Lu Xin-guo et al. [9] presented a gene cluster approach due to most similarity tree. It's an adequate gene cluster method and can generate the preferred global clusters. It is responsible for the separation of equality combinations of equality association including similarity measure called λ. The research results confirmed that the CMST has a superior

performance on classical cluster approaches of K-means and SOM. According to their work, the Gap statistic is recommended to estimate the most optimal similarity measure λ and an optimal self-adaptive gene cluster method based on CMST (OS-CMST). The clustering algorithm of OS-CMST can obtain the relevant similarity measure threshold and then the number of clusters. The standard difficulty of SOM and K-means is the amount of groups is determined at the beginning. Keyan Cao et al. [10] concentrated on the clustering of multidimensional mass data based on density in MapReduce. The researcher emphasis that the classical clustering algorithm cannot be applied to the important modern data on the mass multidimensional data processing speed requirements and the standard clustering algorithm does not consider the multi-dimensional characteristics of the data itself. So, their paper proposed proposes a large-scale multidimensional data clustering algorithm based on density and information entropy. The algorithm uses the idea of DBSCAN clustering algorithm.

Jianlou Lou [11] proposed an optimized gap statistics algorithm based on area density statistics method. Their algorithm applied bad data. By observation, it decreases the computational complexity of iterative computation processing. Also, it improves the computing speed and computing time decreased.

Sithara et al. [12] presented a hybrid clustering algorithm KHM-ABC that is a combination of K-harmonic means & ABC algorithm to achieve a perfect clustering. The results indicated that the performance is better than the other algorithms concerning the quality of clusters. KHM-ABC used artificial bee colony algorithm to optimize K-harmonic means clustering algorithm, and ABC algorithm provides global optimum solutions. The datasets used are iris, wine, yeast, and spam. Cluster quality was checked using silhouette index scores. Silhouette index scores calculated for KHM-ABC, ABC, K-means K-harmonic means and PAM. The performance of KHM-ABC was high compared to the other algorithms. The value of k is not self-learned. In the pre-processing stage, the k value was fixed using gap statistics method and silhouette width method.

Ruqi Zhang et al. [13] preferred a two-step optimization approach for large-scale sparse clustering: the first, k-means clustering over the large-scale data to generate the primary clustering results; the second, clustering learning over the initial findings by developing a spare coding algorithm. The model ensures the scalability of the second round for large-scale data. Also, researchers apply non-linear approximation and dimension reduction algorithms to speed up the sparse coding methods. By using synthetic and real-world datasets, the experimental results demonstrate the promising performance of the LSSC algorithm.

Archana Singh et al. [7] implemented the k-means approach using three different metrics; Euclidean, Manhattan, and Minkowski distance metrics. The research concluded in its comparative study that K-means gives the best performance when using Euclidean distance metric.

A detailed discussion of k-means and its main features is presented in [14]. Also, the study focused on the limitations and how they can be reduced. The study highlighted the criticality of the issue of estimating the suitable number of clusters.

Due to our prior work in clustering on big data, parallel K-Means algorithm showed that it is very efficient and takes less time to build the clusters. It is also very easy to implement. The drawbacks of this algorithm the number of clusters formed by this algorithm is fixed. In the classic k-means, the cluster centers are chosen depend on data chunk in mappers thus different clusters are formed during different runs for same input dataset. The main contribution of this work that the number clusters formed by this clustering algorithm is automated based on gap statistics evaluation criterion. It is hard to apply data mining clustering techniques in Big Data because of the great mass of data and the complexity of clustering algorithms which have very high treatment costs [15].

III. Preliminaries

The proposed model considers: firstly, the MapReduce programming model which trade with big datasets. Secondly, Gap statistic measure to optimize the number of clusters in the k-means technique. The following section explains in details the tw*o concepts.*

A. MapReduce Model

MapReduce is considered as an important programming paradigm for processing and generating big datasets with a parallel, distributed algorithm [15]. It assumes that the Maps are independent and executes them in a parallel manner. MapReduce consists of two main functions known as Map function and Reduce function. In the Map stage, the big dataset is splitted into a set of mappers. Each mapper contains sub-dataset which called data chunk. The Map function has a pair <key, value> that associates the input data. In the Reduce stage, the lowest nodes reach their results back to the parent node which had asked them It computes a partial result using the Reduce function including all the corresponding values for the identical key to a unique pair <key, value> that shown in Fig. 1.

B. Gap Statistics

The gap statistic was developed by Tibshirani et al. [16]. It is a kind of data mining algorithm aims to improve the clustering process by efficient estimation of the best number of clusters. This method is designed to apply to any cluster technique and distance measure. K-means algorithm is executed to determine the number of clusters in a given dataset. It calculates sum of the distance of all objects from cluster mean which known as the dispersion. It creates some amount of sample datasets of original and gets the mean dispersion of these sample datasets. Every gap is described as a logarithmic difference between the mean dispersion of reference datasets and dispersion of the original dataset [12]. The gap is maximized when applying the minimum value of k. The idea behind their approach was to find a way to standardize the comparison of logWk with a null reference distribution of the data [17]. So, the optimal number of clusters K is the value for which logWk comes the farthest below this reference curve in Fig. 2.

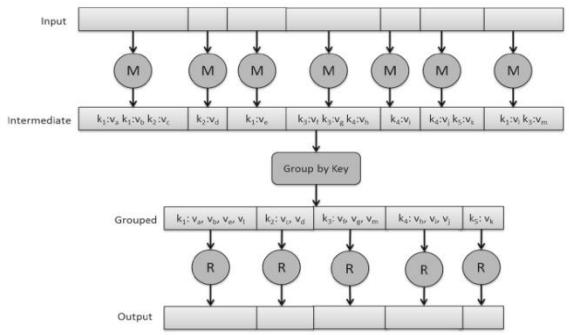

Fig. 1. The MapReducer Programming Model [5].

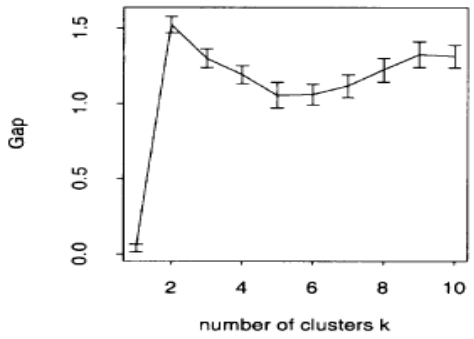

Fig. 2. Gap Curve [16].

IV. Optimized Clustering Approach

The proposed model consists of three main phases shown in Fig. 3. The portioning phase is the primary phase which deals directly with big data. In this phase, the data is spitted into a set of data chunks according to the available hardware environment. At the end of this phase, the big data is converted to a set of small datasets to be moved to the mapper phase. The mapper phase; it is the second phase. It receives a set of data chunks which is stored in a group of mappers. The main task is done in this phase which is executing the k-means algorithm on each mapper. So, the data chunk is locally clustered using the optimal number of clusters determined by the proposed optimized k-means algorithm. In the third phase; the reducer collects the local key-value pairs produced by each mapper. Then the results are merged to generate a global cluster center. The next sections explain in more details each phase.

A. Partitioning Data

The big input dataset is spited into mappers. Input data chunk is fed to each map function in form of data points.

B. Optimized K-Means Clustering Approach

K-means algorithm is evaluated on every data chunk using different numbers of clusters which ranges from 2 to maximum numbers of clusters. In order to determine the optimal number of clusters on every data, the Gap Statistics clustering evaluation is calculated. First, the distance D_K is computed by the sum of all Euclidean distance between all data points' pairs in cluster k

$$D_K = \sum_{x_i \in C_k} \sum_{x_j \in C_k} ||x_i - x_j||^2 = 2n_k \sum_{x_i \in C_k} ||x_i - \mu_k||^2 \quad (1)$$

Second, Within-cluster is computed by a sum of all squares around the cluster mean.

$$W_k = \sum_{k=1}^{K} \frac{1}{2n_k} D_k \quad (2)$$

The third step the "estimated gap" statistic is calculated using eq. 3.

$$Gap_n(k) = E_n^*\{log\ W_k\} - log\ W_k \quad (3)$$

Where the expected value $E_n^*\{log\ W_k\}$ is determined by Monte Carlo [16] sampling from a reference distribution

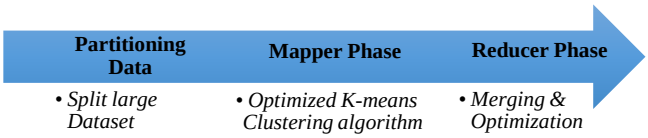

Fig. 3. The Flow Chart of Proposed Approach.

Algorithm: Mapper Phase

Input: D dataset is having n data points.
Output: Optimal k clusters Centers and Data Pointes nearest to them
Step 1: In each Mapper
Prepare Input Data chunk in the form of n data points
Initialize Max-K-Cluster
Step 2: The mapper function finds the optimal K center among k centers for the input point.
For each k=2 to Max-K-Cluster
Clustered-Data=K-means(K)
Distance=Compute-Distance (Clustered-Data)
Within-cluster= Compute-Within-Cluster-Distance (Clustered-Data)
Step 3: Evaluate the optimal number of clusters
For each number of clusters k,
Gap=Compares log(W(k)) with E[log(W(k))]*
Optimal-K=Generate-Optimal (K, Gap)
Step 4: Data Clustering using Optimal-K
Cluster Data into K clusters
Clustered-Data = each k center and all data point which is nearest to it.

Fig. 4. Mapper Phase Algorithm.

Algorithm: Reduce Phase

Input: each k center and all data point which is nearest to it.
Output: *The reducer phase generates global center using and data points.*
Step 1: Collects Data from all mappers
Key-Center=Collect(k center, Data-Points)
Step 2: Merging Clusters
Clustered-Data=Merge (Key-center, data points)
For each K =2 to no cluster Centers
Sum= Calculate-Sum(data points)
Count= Calculate-Count(data points)
Global-K-Center =Calculate-Mean (Sum, Count)
Step 3: Clustering Data Generation
Clustered-Data= generate (Global-K-Center ,all data points)

Fig. 5. Reduce Phase Algorithm.

Finally, the optimal number of clusters is chosen as the smallest k such that Gap (k) ≥ Gap (k+1). The map function finds the nearest center among an optimal k centers which considered as key for the input point. The mapper phase produced <key, value > pairs. The clustering using an optimal number of cluster occurred in the mapper phase which shown in Fig. 4.

C. Merging and Optimization

The output of mappers <key, value > where key is local cluster center and value is set of all data point that nearest to this centered is received from mappers. The data points is grouped by key, the center of all clustered data is calculated for each cluster that returned as the global cluster center. Set of clusters are optimized with clusters global center and data points located in it as value. Reduce phase will show in Fig. 5 in detail.

V. Experimental Results

A. Dataset

In this experiment, four large-scale datasets conducted, available in the UCI repository whose statistics are summarized in the following:

1) Covtype dataset: It is consists of 581012 data points for predicting forest cover type about cartographic that received from a known survey called US Geological Survey (USGS) and US Forest Service (USFS) data. Each sample belongs to one of seven classes.

2) Covtype-2 dataset: it is similar to Covtype dataset except for a number of classes. Each sample belongs to one of two classes.

3) Poker dataset: it contains 1, 025, 010 data points. There are 10 classes in the dataset, each depicting a type of poker hand.

4) Poker-2 dataset: it is similar to Poker Dataset except for the number of the class which is 2 class.

B. Experiments Evaluation Metrics

The Optimized model evaluates the clustering quality of the proposed model using accuracy and time has been taken in the processing. The speed up measurement is presented to evaluate the time performance.

1) Accuracy (Chen and Cai2011): Accuracy is the first reasonable evaluation measurement. The accuracy of an analysis is how close a result comes to the actual value. Accuracy used to estimate the performance of the proposed approach. A larger Accuracy value indicates better clustering performance. The accuracy is defined as:

$$Accuracy = \frac{\text{true positives value} + \text{true negatives value}}{\text{true positives value} + \text{true negatives value} + \text{false positives value} + \text{false negatives value}}$$

2) Time taken: The second metric is the time that consumed in the execution. It is recorded in seconds. Due the various conuration, the execution time would be different from machine to another.

3) Speed up: It is a number that holds the corresponding performance of two methods processing the same issue. Also, it is the increase in speed of execution of a task performed on two similar structures with various sources. The speed up measure had used to assess the performance of the proposed approach, where Tc is the execution time on current method, and T_p is the execution time on classical k-means which is calculated as follows:

$$Speedup = \frac{T_c}{Tp}$$

TABLE I. Accuracy Results for Four Datasets

Methods \ Datasets	Covtype	Covtype-2	Poker	Poker-2
Basic K-means	56.72%	62.10%	62.67%	63.20%
K-means & Fuzzy Gaussian	62.10%	75.59%	63.39%	73.39%
Optimized K-means	67.59%	76.30%	72.10%	75.30%

TABLE II. Time Taken /Speed up for Four Dataset

Methods \ Data sets	Time / Speed up	Covtype	Covtype-2	Poker	Poker-2
k-means	Time	840.39	784.862	701.66	725.34
	Speed up	1×	1×	1×	1×
K-means & Fuzzy Gaussian	Time	935.0529	948.2365	831.4306	831.4306
	Speed up	0.89x	0.82x	0.84x	0.87x
Optimized K-means	Time	490.17	709.06	325.12	680.21
	Speed up	1.71x	1.1x	2.15x	1.06x

C. Results

In this section, the experiment's results display the evaluation of the proposed approach. The tests have been designed to contrast the results of the successive version about the big data versions of the algorithm. The experiments applied three methods, and compared them to examine the optimized K-means. The features of these methods are provided below

1) Basic k-means: The k-means clustering algorithm utilizes the Euclidean distance to calculate the similarities among instances. It can be seen as a baseline method. Both adaptive algorithm and iterative algorithm exist for the traditional k-means clustering. It needs to assume that the number of clusters is determined a priori.

2) K-means & fuzzy Gaussian: It is a parallel large-scale clustering approach based on Fuzzy Gaussian membership. It is based on the MapReduce programming model. All object relates to each cluster according to its degree. The degree is based on the probability of the instance which generated from each cluster's (multivariate) normal distribution.

3) Optimized k-means: It is the proposed approach, optimizing method to determine the optimal K according to a dataset. It is based on the gap statistics algorithm.

a) Accuracy: The methods which applied the four datasets are recorded the accuracy results in Table 1 which showed a comparison among Basic K-means, K-means & Fuzzy Gaussian, and Optimized K-means. By observation, several interesting points as follows:

- The proposed approach outperforms the classical k-means by 10.9%, 14.2% when applied on Covtype, Covtype2 by respectively. While the Poker, Poker2 achieve 9.4%, 12.1%.
- K-means is applied to four datasets. By observation, Optimized K-means outperformed of the other method. It achieves the best result in Covtype-2 due to reducing the number of classes.
- By comparing between K-means & Fuzzy Gaussian and Optimized K-means, the accuracy of Covtype-2 and Poker-2 is a very low enhancement, because of the number of the cluster label is only two classes.

b) Time taken: According to big data size, the time taken is a critical metric. Table 2 shows the running time of all the methods on four datasets. Among the comparisons, there are some useful points as follows:

- K-means & Fuzzy Gaussian is the highest time taken, but it records a good accuracy compared by Basic K-means.
- Optimized K-means outperformed on the other methods, it takes less time in execution,
- By observation, Covtype & Poker datasets take the lowest time when applying Optimized K-means, against Covtype-2 & Poker-2. The main reason due to the number of cluster label of them.

VI. Conclusions

Clustering techniques are the process of grouping objects that belong to the corresponding class. Related objects are grouped into one cluster, and different objects are arranged in another cluster. Many applications used clustering analysis in like data analysis, pattern recognition, and market research. K-means clustering is extremely fast, robust & easily understandable and manageable to implement. It gets many clusters (K) as input from the user. The user can indicate the suitable number of clusters by running a lot of experiments. Each instance is allocated to its nearest centroid, then the set of centroids is updated as the centers of mass of the instances attached to the same centroid in the previous step. So, the main problem in the K-means algorithm is fixing the number of clusters in advance. Specifically, when trade with big data it causes a critical challenge according to the data size and execution time Then compare several different clustering of the data and focus the optimal one which improves the accuracy and consume the time. Therefore, the optimized k-means proposed a model which can calculate the optimal number of clusters. It consumes time and can record the best accuracy.

References

[1] Feng Chen, Pan Deng, Jiafu Wan, Daqiang Zhang, Athanasios V. Vasilakos, Xiaohui Rong, "Data Mining for the Internet of Things: Literature Review and Challenges", International Journal of Distributed Sensor Networks, Vol. 11, No.8, 2015.

[2] Divya Pandove,Dr. Shivani Goel, "A Comprehensive Study on Clustering Approaches for Big Data Mining", IEEE SPONSORED 2ND INTERNATIONAL CONFERENCE ON ELECTRONICS AND COMMUNICATION SYSTEM (ICECS), 2015.

[3] Barkha Narang, Poonam Verma, Priya Kochar, "Application based, advantageous K-means Clustering Algorithm in Data Mining – A Review", International Journal of Latest Trends in Engineering and Technology (IJLTET), Vol 7, No. 2, 2016.

[4] M. Emre Celebi, Hassan A. Kingravi, Patricio A. Vela, "A comparative study of efficient initialization methods for the k-means clustering algorithm", Vol. 40, No. 1, pp. 200-210, 2013.

[5] Sergio Ramírez-Gallego, Alberto Fernández, Salvador García, Min Chen, Francisco Herrera, "Big Data: Tutorial and guidelines on information and process fusion for analytics algorithms with MapReduce", Information Fusion, Vol. 42, pp.51-61, 2018.

[6] Shruti Aggarwa, Parminder Singh, "Comparative Study of Various Enhanced K-Means Clustering Algorithms", International Journal of Computer Science and technology (IJCST), Vol. 5, No. 1, 2014.

[7] Archana Singh, Avantika Yadav, Ajay Rana, "K-means with Three different Distance Metrics", International Journal of Computer Applications, Vol. 67, No.10, 2013.

[8] Anil K. Jain, "Data clustering: 50 years beyond K-means", pattern recognition letters, Vol. 31, No.8, pp. 651-666, Elsevier Publisher, 2010.

[9] Lu Xin-guo Lin Ya-ping Li Xiao-long Yi Ye-qing Cai li-jun Wang Hai-jun, "Gene Cluster Algorithm Based on Most Similarity Tree", Proceedings of the Eighth International Conference on High-Performance Computing, 2005.

[10] Keyan Cao, Ibrahiem Musa, Jiadi Liu, "An Adaptive Density Clustering Algorithm for Massive Data", 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), pp. 1700-1707, IEEE, 2017.

[11] Lou Jianlou, Jizhe Xiao, Hongjian Zheng, and Zhaoyang Qu., "Application of Optimized GSA Algorithm on Bad-data Detection of Electric Power Dispatching System", Advanced Science and Technology Letters (AST 2017), Vol.143, pp.169-175, 2017.

[12] Sithara E.P and K.A Abdul Nazeer, "A HYBRID K-HARMONIC MEANS WITH ABCCLUSTERING ALGORITHM USING AN OPTIMAL K VALUE FOR HIGH PERFORMANCE CLUSTERING", International Journal on Cybernetics & Informatics (IJCI) Vol. 5, No. 2, April 2016.

[13] Ruqi Zhang, Zhiwu Lu, " Large Scale Sparse Clustering", Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI), pp. 2336-2342, 2016.

[14] Kehar Singh, Dimple Malik, Naveen Sharma, "Evolving limitations in K-means algorithm in data mining and their removal", IJCEM International Journal of Computational Engineering & Management, Vol. 12, pp. 105-109, 2011.

[15] Btissam Zerhari, Ayoub Ait Lahcen, Salma Mouline, "Big Data Clustering: Algorithms and Challenges", Proceedings of International Conference on Big Data, Cloud and Applications (BDCA'15), 2015.

[16] Robert Tibshirani, Guenther Walther, Tervor Hastie, "Estimating the number of clusters in a dataset via the gap statistic", Royal Statistical Society, Vol. 63, No.2, pp. 411-423, 2001.

[17] Shuo Xu, Xiaodong Qiao, Lijun Zhu, Yunliang Zhang, Chunxiang Xue, Lin Li, "Reviews on Determining the Number of Clusters", Applied Mathematics & Information Sciences, Vol. 10, No. 4, pp. 1493-1512, 2016.