Asian Journal of Research in Computer Science

11(3): 1-8, 2021; Article no.AJRCOS.71463
ISSN: 2581-8260

A Survey on Unsupervised K-Means Algorithm in Big Data Environment

Fatama Sharf Al-deen[1*] and Fadl Mutaher Ba-Alwi[2]

[1]Department of Computer Science, Sana'a University, Sana'a, Yemen.
[2]Department of Information System, Sana'a University, Sana'a, Yemen.

Authors' contributions

This work was carried out in collaboration between both authors. Both authors read and approved the final manuscript.

Article Information

DOI: 10.9734/AJRCOS/2021/v11i330262
Editor(s):
(1) Dr. Francisco Welington de Sousa Lima, Federal University of Piaui, Brazil.
Reviewers:
(1) Hilman F. Pardede, Indonesian Institute of Sciences, Indonesia.
(2) Siti Zulaiha Ahmad, UiTM Cawangan Perlis Kampus Arau, Malaysia.
Complete Peer review History: https://www.sdiarticle4.com/review-history/71463

Review Article

Received 01 June 2021
Accepted 04 August 2021
Published 24 August 2021

ABSTRACT

Due to the rapid development in information technology, Big Data has become one of its prominent feature that had a great impact on other technologies dealing with data such as machine learning technologies. K-mean is one of the most important machine learning algorithms. The algorithm was first developed as a clustering technology dealing with relational databases. However, the advent of Big Data has highly effected its performance. Therefore, many researchers have proposed several approaches to improve K-mean accuracy in Big Data environment. In this paper, we introduce a literature review about different technologies proposed for k-mean algorithm development in Big Data. We demonstrate a comparison between them according to several criteria, including the proposed algorithm, the database used, Big Data tools, and k-mean applications. This paper helps researchers to see the most important challenges and trends of the k-mean algorithm in the Big Data environment.

Keywords: K-mean; Big Data; unsupervised learning; clustering.

1. INTRODUCTION

With the rapid development of technologies, the use of machine learning has become one of the most important fields in the modern era, and it is involved in many applications. Machine learning is divided into supervised learning and unsupervised learning. Supervised learning (SL)

**Corresponding author: E-mail: am.mohomed1985@gmail.com;*

is the machine-learning task of learning a function that maps an input to an output based on example input-output pairs. It infers a function from labeled training data consisting of a set of training examples. Unsupervised Learning (UL) is taking out data patterns and sample to classify unlabeled data. [1]. The clustering function is one of the most important uses of unsupervised learning, clustering sections data into double clusters in which cluster differs from one to another [2].

There are many algorithms classifying data in which K-mean algorithm is the most popular one. The k-mean algorithm is unsupervised learning algorithm used to categorize unidentified data by dividing it into a number of groups (clusters) where each group shares a large share of its characteristics, in which, each cluster is represented by the center or means of the data points belonging to the cluster [3].

There are several developments and improvements that researchers have made to the k-mean algorithm, especially with the large volume of data and the emergence of the concept of Big Data, the development of the k-mean algorithm has become important.

Big Data is a new generation of technologies and architectures designed to economically extract value from very large volumes of a wide variety of data by enabling high velocity capture, discovery, and/or analysis [4]. A simple definition by Jason Bloomberg [5]: “Big Data: a massive volume of both structured and unstructured data that is so large that it's difficult to process using traditional database and software techniques [6].

The advent of Big Data has highly affected performance of k-mean algorithm. Therefore, many researchers have proposed several approaches to improve K-mean accuracy in Big Data environment. In this paper, we introduce a survey about different technologies proposed for k-mean algorithm development in Big Data. We demonstrate a comparison between them according to several criteria, including the proposed algorithm, the database used, Big Data tools, and k-mean applications.

This paper will be organized as follows: Section 2 presents the methodology of this research. Section 3 introduces the K-mean algorithm in a Big Data environment. Section 4 clarifies the unsupervised K-Means approaches in Big Data in terms of three criteria: speed, quality, and accuracy. In section 5, we presents results and discussion in a table. Section 6 represents the conclusion of the paper.

2. METHODOLOGY

In this systematic literature review, we have collected scientific papers that were relevant to our research which is unsupervised K-means in big data. We read them carefully and found that the researchers were interested in studying this algorithm because of its importance in the field of classifying data. After that, we classified these approaches based on three categories: approaches improving the algorithm speed, approaches improving the clustering quality, and approaches improving the accuracy. Lastly, we collected the studies in each paper and made a comparison between them in terms of the algorithm used, the database, the tools used, the applications used, the advantages in each paper and the disadvantages of it .

3. UNSUPERVISED K-MEANS IN BIG DATA

K-Means is one of the algorithms that solve the well-known clustering problem, the algorithm classifies objects to a predefined number of clusters, which is given by the user (assume k clusters). The idea is to choose random cluster centers, one for each cluster. These centers are preferred to be as far as possible from each other. Starting points affect the clustering process and results. Here the Centroid initialization plays an important role in determining the cluster assignment in effective way. In addition, the convergence behavior of clustering is based on the initial centroid values assigned. Although, the K-means algorithm has a strong qualifications in classifying data ,it is very slow with dealing with Big data and the optimal number of clusters is not defined by algorithm programmatically. Also, Some results are not accurate as the algorithm starts with picking center points randomly.

4. Unsupervised K-Means Approaches in Big Data

Many researchers have made a wide range of improvements to the K-mean algorithm in order to improve the performance of the algorithm and the ability to cluster data at a high speed. The concept of big data and its techniques have influenced these developments and research. In this section, we focus on research approaches

that focused on algorithm development in Big Data. We will classify these approaches into three categories based on the improvement they made.

4.1 Speed Improvement of K-Means Algorithm

This section explains some approaches that have focused on improving the speed of K-means algorithm in clustering big data.

The authors in [7] have demonstrated a parallel and distributed framework to increase data scalability using k-mean algorithm in big data environment. The authors used a protocol named AMQP to apply asynchronous communication mechanism. They used MRI data to test their framework and deploy these data on the cloud. The evaluation showed an improvement in data scalability. Moreover, the authors in [8] have proposed a new algorithm. The objective of the recommended algorithm was to unite the search process of GA to generate new data clusters. The suggested approach was instrument using the common MapReduce programming model on Hadoop framework. It also applied parallel K-Means to increase velocity of the quality of the search procedure during clusters formation. Experiments were conducted with many synthetic datasets to appraise the performance of the proposed algorithm. In addition, the main goal in [9] was to improve the efficiency of the algorithm and reduce the time complexity of the algorithm, the distributed database was used to simulate the shared memory space and parallelize the algorithm on the Hadoop platform of cloud computing. The fundamental cluster was joined according to the area between the two cluster centers. The points that were not divided into any cluster were divided into the clusters nearest to them. Furthermore, the density of data essence was calculated, and each basic cluster was composed of the center points whose density was not less than the given threshold and the points within the density range. Furthermore, the authors in [10] have used two types of firefly algorithms to manage the problem of initialization and local minima in K-mean algorithm. The evaluation results showed an improvement comparing to the traditional K-mean. Additionally, the authors in [11] have presented that optimized clustering approach was calculated the better number of clusters (k) for precise domain problems. The proposed approach was a better solution based on the cluster deed measure analysis based on gab statistic. Moreover, the authors in [12] have offered Grid-K-means algorithm by planning datasets as firstly proposed to overcome drawbacks of the traditional K-means algorithm. Then, a new clustering validity index (BCVI) was formulated to better evaluate by utilizing grid points as the weighted representative points to process datasets . The quality of clustering results generated by the Grid-K-means algorithm . Established upon the monotonous feature of BCVI and the linear combination of intra-cluster compactness and inter-cluster separation of clusters, BCVI consumes were much lower time cost in finding the optimal clustering number (K opt) than the commonly used method that utilized the empirical rule K max to calculate the(K opt). In addtion, the authors in [13] have presented a new optimization methodology which was proposed. On the basis of an unsupervised learning framework for the purpose of designing a new clustering algorithm was replaced the distance criteria into intensity level majority for group comparable image pixels in a cluster. The proposed method for finger vein image localization based on better clustering algorithm had many stages such as using some important pre-processing steps.

4.2 Quality Improvement of K-means Algorithm

This section displays approaches that have improved the algorithm in terms of the quality of clustering such as creating high quality clusters and developing the initialization process of center points.

The authors in [14] have presented a robust K-means clustering algorithm, particularly, flexible subspace clustering. The aim of enhancing the robustness, the l2.p-norm was inserted into the objective function. The aimed method incorporated feature selection and K-means clustering into a unified framework, which chose the refined features and made good use of the clustering performance. When we chose an appropriate p according to the various data and thus obtained additional robust performance. Furthermore, the authors in [15] have introduced a novel k-means variant of the primary algorithm was suggested. In this novel setting, cluster middle contend against another to attract the largest number of similar objectives or entities to their cluster. The approach leveraged the power of bargaining game modeling in the k-means algorithm for clustering data. The middle kept changing their positions so that they had smaller distances with the maximum possible data than

other cluster middle. To exhibit the superiority and efficiency of GBK-means over conventional clustering algorithms, particularly, k-means and fuzzy k-means, the authors treated two different syntactic and real-world data sets, a series of two-dimensional syntactic data sets; and ten benchmark data sets that are widely used in different clustering studies. The authors called this new algorithm the game-based k-means (GBK-means) algorithm. Moreover, the authors in [16] have constructed an unsupervised learning schema for the k-means algorithm. It was free of initialization without parameter choice and it simultaneously found an optimal number of clusters. The authors suggested a novel unsupervised k-means (U-k-means) clustering algorithm with automatically finding better number of clusters without showing any initialization and parameter selection. In addition, the authors in [17] have improved an existing clustering algorithm. A novel algorithm was introduced by joining Spectral clustering with k-means with NFPH. The proposed system substituted the initialization method for cluster centroids in classical k-means algorithms. They should solve some of the limitations of the k-means algorithm. The authors intended to select the most appropriate first centroid rather than selecting randomly. Test data sets from the medical domain which were available for research purposes would be used to train the model and an open source data mining application called WEKA was utilized for testing. From tests carried out on 10 different UCI data sets, using the proposed solution found that the clustering error was reduced up to 2 percent while the processing time enlarged from 4~5 seconds. The growth in processing time was caused by the replacement of the initialization method of k-means. The proposed method reduced the clustering error of the spectral clustering algorithm.

4.3 Accuracy Improvement of K-Means Algorithm

This section explains approaches that focused on improving the accuracy of clustering such as increasing the diagnosis precision using K-means algorithm.

the authors in [18] have introduced Spark-based three-layer wavelet packet decomposition approach, it was stored in Hadoop Distributed File System (HDFS), and worked as the input of ACO-K-Means clustering algorithm. It also efficiently preprocessed the running-state monitoring data to obtain eigenvectors. ACO-K-Means clustering algorithm was appropriate for rolling bearing fault diagnosis that improved the diagnosis accuracy. ACO algorithm was adopted to acquire the global optimal initial clustering centers of K-Means from all eigenvectors. The K-Means clustering algorithm established upon weighted Euclidean distance was employed to perform clustering examination on all eigenvectors to obtain a rolling bearing fault diagnosis model. The ACO-K-Means clustering algorithm was executed on a Spark platform. Moreover, the authors in [19] have demonstrated three major techniques methods of k-means_AMV. First, to put the practical use of the contextual data in an adaptive manner and region around. A central pixel was built by detecting the spectral similarity between the main pixel and its eight neighboring pixels. Second, the extension for the adaptive region was stopped, the k-means clustering method was implemented to determine the label of each pixel within the adaptive region. Third, an existing AMV technique was utilized to refine the label of the main pixel of the adaptive region. Changing magnitude image (CMI) was scanned and processed in this manner. The label of each pixel in the CMI could be refined and the binary change detection map could be generated. Three images scenes related to many land cover change events were adapted to examine the effectiveness and performance of the proposed k-means_ AMV approach. Additionally,16The authors in [20] have put forward a recent clustering algorithm named hybrid clustering so that overcoming was the defects of existing clustering algorithms. The recent hybrid algorithm and existing algorithms were compared on the bases of precision, recall, F-measure, execution time, and accuracy of results. Furthermore,the authors in [21] have presented the enhanced K-means algorithm based on the versatile clustering number for the segmentation of tomato leaf images. The whole attempt images were acquired from the tomato we grew. The white paper background images were used for planning algorithm. The inborn background images were the algorithm validated data. Through a sequence of pretreatment experiments, the value of the clustering number in this algorithm was automatically determined by calculating the DaviesBouldin index. The initial clustering hlfway was given to prevent the clustering calculation from falling into a local optimum. Finally, authors verified the accuracy of segmentation by two types of objective assessment measures the clustering F1 measure and Entropy.

Table 1. Comparison of improved k-means algorithms in Big Data

Reference	Algorithm	Dataset	Tools	Application	improvement
[7]	Distributed k-means	medical image	AMQP protocol	Cloud micro-services	reduce the communication cost
[8]	GAPKCA	multiple synthetic datasets	Hadoop	in marketing and finance	speed up document clustering process by 0.54 s on average
[9]	Improved K-Means	KDD99	Hadoop	Big Data Mining	reduce the time complexity
[10]	Improving K-means	Seven datasets,Sonar, Ozone, Wisconsin			demonstratestatistically significant superiority in both distance and performance measures for clustering tasks
[11]	Optimized K-Means	Four dataset Covtype, Covtype-2, Poker	MapReduce	Gap Statistic	enhance the speed of the clustering process and accuracy by reducing the computational complexity
[12]	Grid-mapping K-means	12dataset Iris, HTRU2, 4K2			s faster and more accurate than the traditional ones
[13]	K-Means	SDUMLA-HMT	unsupervised learning framework	Finger Vein Image Localization	faster than the other clustering algorithms
[14]	K-Means	seven datasets, Cars, Wine, Ionosphere, and other	unified		more robust and better performance
[15]	GBK-means	series of two-dimensional, ten benchmark		bargaining game	able to cluster data more accurately
[16]	U-k-means	Eight datasets, Iris, Seeds, Australian, and other	MATLAB		solving the initialization problem
[17]	k-Means	10 dataset Liver, Diabetes, Audiology	NFPH	medical domain	able to produce a higher quality of clusters
[18]	ACO-K-Means	rolling bearing datasets	Spark	GPU	improve the diagnosis accuracy

[19]	k-Means	image datasets	Novel Land Cover Change Detection	aerial photography	better detection accuracies and visual performance
[20]	hybrid clustering	National Climatic Data Center	Hadoop		more accurate, and has better precision, recall, and F-measure values.
[21]	K-means	Image	image-processing	Segmentation of tomato leaf images	accuracy of segmentation by two kinds of objective assessment measures, the clustering F1 measure and Entropy

5. RESULTS AND DISCUSSION

In this section, we will summarize all approaches in the literature in Table 1. The summary table outlines these approaches in terms of their algorithms, dataset, tools, their applications, and their improvement.

6. CONCLUSION

Because of the great development in the field of machine learning, the use of the k-mean algorithm in many applications and the emergence of the term Big Data, many researchers have used and developed the k-mean algorithm. In this paper we presented a survey on the works that were developed the k-mean algorithm in order to improve the performance of the algorithm, and compared them with dependence on some criteria. In future work, we will propose a development on the k-mean algorithm.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. Kameshwaran K, Malarvizhi K. Survey on clustering techniques in data mining. International Journal of Computer Science and Information Technologies. 2014;5(2): 2272-2276.
2. Han J, Kamber M. Data mining concepts and techniques, Morgan Kaufmann Publishers," San Francisco, CA. 2001;335-391.
3. Yadav R, Sharma A. Advanced methods to improve performance of k-means algorithm: A Review. Global Journal of Computer Science and Technology. 2012; 12(9):47-52.
4. Gantz J, Reinsel D. Extracting value from chaos," IDC iview. 2011;1142(2011):1-12.
5. Demchenko Y, De Laat C, Membrey P. Defining architecture components of the Big Data Ecosystem. In 2014 International Conference on Collaboration Technologies and Systems (CTS), IEEE. 2014;104-112 .
6. Hey T, Tansley S, Tolle K. The Fourth Paradigm: Data-Intensive Scientific Discovery, edited by T," Hey, S. Tansley, K. Tolle (Microsoft Research, Redmond, Washington; 2009.
7. Benchara FZ, Youssfi M. A new scalable distributed k-means algorithm based on Cloud micro-services for High-performance computing. Parallel Computing. 2021; 101:102736.
8. Alshammari S, Zolkepli MB, Abdullah RB. Genetic algorithm based parallel k-means data clustering algorithm using MapReduce programming paradigm on hadoop environment (GAPKCA). In International Conference on Soft Computing and Data Mining, Springer. 2020;98-108.
9. Lu W. Improved K-means clustering algorithm for big data mining under Hadoop parallel framework. Journal of Grid Computing. 2019;1-12.
10. Xie H, et al. Improving K-means clustering with enhanced firefly algorithms. Applied Soft Computing. 2019;84:105763,
11. El-Mandouh AM, Mahmoud HA, Abd-Elmegid LA, Haggag MH. Optimized K-means clustering model based on gap statistic. Int J Adv Comput Sc. 2019;10(1):183-188.
12. Zhu E, Zhang Y, Wen P, Liu F. Fast and stable clustering analysis based on Grid-mapping K-means algorithm and new clustering validity index. Neurocomputing. 2019;363:149-170.
13. Sulaiman DM, Abdulazeez AM, Haron H, Sadiq SS.Unsupervised Learning Approach-Based New Optimization K-Means Clustering for Finger Vein Image Localization. In 2019 International Conference on Advanced Science and Engineering (ICOASE), IEEE. 2019;82-87.
14. Long Z-Z, Xu G, Du J, Zhu H, Yan T, Yu Y-F. Flexible Subspace Clustering: A Joint Feature Selection and K-Means Clustering Framework," Big Data Research. 2021 23: 100170
15. Rezaee MJ, Eshkevari M, Saberi M, Hussain O. GBK- means clustering algorithm: An improvement to the K-means algorithm based on the bargaining game," Knowledge-Based Systems. 2021;213: 106672.
16. Sinaga KP, Yang M-S. Unsupervised K-means clustering algorithm," IEEE Access. 2020;8:80716-80727
17. Sapkota N, Alsadoon A, Prasad P, Elchouemi A, Singh AK. Data summarization using clustering and classification: Spectral clustering combined with k-means using nfph. In 2019 International Conference on Machine

Learning, Big Data, Cloud and Parallel Computing (COMITCon), IEEE. 2019;146-151 .

18. Wan L, Zhang G, Li H, Li C. A novel bearing fault diagnosis method using spark-based parallel ACO-K-Means clustering algorithm. IEEE Access. 2021;9:28753-28768,
19. Lv Z, Liu T, Shi C, Benediktsson JA, Du H. Novel land cover change detection method based on K-means clustering and adaptive majority voting using bitemporal remote sensing images. IEEE Access. 2019;7:34425-34437, 2019.
20. Kumar S, Singh M. A novel clustering technique for efficient clustering of big data in Hadoop Ecosystem," Big Data Mining and Analytics. 2019;2(4):240-247, 2019.
21. Tian K, Li J, Zeng J, Evans A, Zhang L. Segmentation of tomato leaf images based on adaptive clustering number of K-means algorithm. Computers and Electronics in Agriculture. 2019; 165:104962.

Peer-review history:
The peer review history for this paper can be accessed here:
https://www.sdiarticle4.com/review-history/71463