International Journal of Engineering & Technology, 7 (2.27) (2018) 50-55

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

(N,α)-means algorithm for clustering big data

Md Tabrez Nafis [1]*, Ranjit Biswas [1]

[1]*Department of CSE, Jamia Hamdard University, INDIA*
**Corresponding author E-mail: tabrez.nafis@gmail.com*

Abstract

The k-means algorithm is a popular algorithm for clustering data, but it is not appropriate for clustering big data. In this paper the authors modify the existing k-means algorithm to develop a new algorithm called by (N,α)-means algorithm. The proposed (N,α)-means algorithm is developed to cluster N number of big data into α number of clusters. In our approach by (N,α)-means algorithm the result is achieved in n number of sequential steps, in each step executing k-means algorithm twice.The method provides wide opportunity to many data points to stand as leaders and to justify their leadership with the progress of time. This new algorithm, if incorporated in the existing popular data mining tools (viz. Rapid Miner, Orange, Weka, Knime, Oracle Data Mining, etc.), is expected to play a better role in case of data mining of big data.

Keywords:*Big Data; (N,α)-Means; Multiset; Bag; Multiset Space; Leader-Set.*

1. Introduction

Clustering is a technique in Machine Learning which is an important action for grouping of data points. The data points which belong to a group [15], [31] should have similar properties and/or features, whereas the data points falling in different groups should have dissimilar properties and/or features [11], [15], [31].For a given multiset [3] of data points, we use to apply a clustering algorithm to classify each data point into a specific group. Some of the important and popular clustering algorithms [11], [15] are: k-means algorithm, k-medians algorithm, DBSCAN clustering,Agglomerative Hierarchical Clustering, Mean-shift clustering, EM-Clustering (using GMM), etc.However, one of the most useful existing algorithms is k-means algorithm for clustering structured data. The k-means algorithm aims to partition N data into k number of clusters,in which each data belongs to the cluster with the nearest mean, serving as a prototype of the cluster. This belongingness is unique, i.e.
One data will surely belong to some cluster out of these k clustersand will not belong to more than one cluster.
A large number of applications of the k-means algorithm are reported by various authors in diverse areas for clustering data or information or text for various analysis. Consider few recent applications, their merits and demerits. In [26] an under sampled k-means approach for handling imbalanced data has been proposed. However, in [20] the authors have proposed k-nearest neighbor distance based under sampling for improved opinion mining for positive or negative options only. The K-Means clustering method using ArtificialNeural NetworkBackPropagation (ANN-BP)has been applied in [8] to predict the minimum wage to workers in their own business environment. An enhanced of K-Means algorithm is proposed in [28] to achieve better and meaningful result of ranking-based clustering that eventually accelerates the clustering process. The phishing becomes considerably bigger issues in online networking, for example, Facebook, twitter and Google+, etc.A novel approach for phishing emails real time classification using k-means algorithm is proposed in [18]. In [9] the authors proposed an enhanced algorithm for the prediction of chronic, autoimmune disease called Systemic Lupus Erythematosus(SLE). TheHybridK-meansJ48DecisionTreealgorithm(HKMJDT)hasbeenproposedfor theeffectiveandearly prediction of the SLE. The use of social big data analysis for social media is increasing rapidly. The authors in [14] proposed a method to apply text clustering for analysis by related topics of texts extracted using text mining of social big data. The Privacy pre-serving data mining (PPDM) has emerged as a main research area [24] for data confidentiality and knowledge sharing in between the communicating parties. An Information Network is the network formed by the interconnectivity of the objects formed due to the interaction between them. In [16] the authors made a good survey of data mining techniques on Information Network, and network size is considered by the number of nodes and/or edges.Forest based clustering for gene classification reported in [22] is also quite interesting.
But none of the above techniques and methods can be applied for clustering big data. There are a number of challenges and issues in real life environment while dealing with big data [1-5], [7], [17], [21], [23],[32]. There are difficulties for applying classical clustering algorithms/techniques to big data due to its 4Vs or 5Vs challenge [4], in particular the security aspects [19], mainly because of the fact that the big data deals with terabytes and petabytes of data. There are few methods proposed by the researchers [10], [12], [13], [25], [33],[34] for clustering big data but there is still a lot to do to find better methods. We consider here the classical and probably the most popular clustering algorithm which is k-means clustering algorithm. The k-means algorithm is not suitable for clustering Big Data even if be structured. One of the significant reasons is that the value of k could be too small compared to the massive size N of big data. For example, consider the following two problems:-
Problem-1:
A problem for classical data set

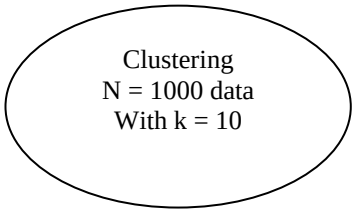

Fig. 1: K-Means Algorithm Is Applicable.

Problem-2:
A problem for Big Data

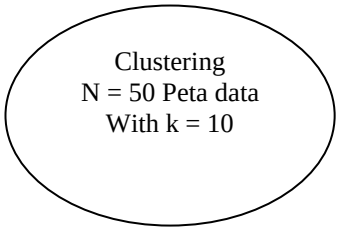

Fig. 2: K-Means Algorithm Is Not Suitable.

In such a case (as in Figure 2), k-means algorithm if directly be applied in big data in just one go for clustering, then clusters may have significant deviation from the basic properties of 'likely similardata', unlike the situation of Figure 1.

1.1. The major demerits of k-means algorithm if applied to problem-2

It is known to the researchers that the following are the demerits of the k-means algorithm if applied to cluster data, even if the number N be not large.But for the Problem-2,the demerits are much more significant and alarming because of the reason that k << N. The major demerits in case of Problem-2 are:

i) How to initially choose k number of leaders where k is too too small compared to the large value of N?
ii) Ifthe initial k leaders or most of them be too close (being almost close neighbors), then the situation will be worse as k << N.
iii) The number of tie cases of d(x,C) during execution of k-means algorithm could be many in count.

1.2. Problem statement

Consequently, in this research work we pose the following important problem of big data:
"How to have α number of clusters out of N (very large amount) structured data, where α << N?"
The solution for this problem is proposed in the next section by developing a new algorithm called by "(N,α)-means Algorithm" which is a slightly modified version of the k-means algorithm.In the notation "(N,α)-means Algorithm", N stands for the size of Big Data and αstands for the desired number of clusters.

2. (N, α)-means algorithm

In our theory, we regard the collection B of the N number of big data as a 'multiset' (or bag) because it is a collection of data points where repetition of data may sometimes exist;and we regard the object (B,d) as a multiset space [3] with respect to the metric d over the multiset B (see [3], [6], [27], [30]).Any cluster of the multiset space (B,d) is also a multiset space being a sub-multiset of the universe B. For details of multiset (and bag structure) and of multiset space theory of a population data, one could see [3].
We use the following notations, all-through in our work here:

i) B = the collection of N big data (which forms a multiset i.e. bag)
ii) CL(C_j)=the latest cluster in our steps constructed from the leader C_j.
iii) CL(C_{ij})=the newly formed cluster by union of two existing clusters C_i and C_j.Here C_{ij} denotes the mean of the data of the cluster CL(C_{ij}). Once the cluster CL(C_{ij}) is formed, the old clusters C_i and $C_j$are no more existing as clusters in this latest status.
iv) Any cluster CL(C_i)is shown by drawing a circle. However, a figure of circle with centre at C_i with dash border means the cluster CL(C_i) is under construction.

The proposed method of "(N,α)-means Algorithm" consists of several sequential steps. There are n = Floor($\log_\alpha N$)number of sequential steps as described below, where the notation 'Floor' denotes the mathematical Floor function.At each of the n steps in this method the classicalk-means algorithm will be applied twice, except in Step-1 where the classicalk-means algorithm will be applied once only.The value of k will be different for different steps, reducing at each step, and at the last step (i.e. at the nth step) the value of k is α producing finally α number of clusters of the big data multiset B.
Step-1:
Consider the N number of big data which is here the multiset B. We now apply k-means algorithm with$k = \alpha^n = N_1$to cluster the N number of data multiset B starting with the randomly choosen N_1 number of leaders, say, $C_1^1, C_2^1, C_3^1, \ldots C_{N_1}^1$.
Suppose that on completion of the execution of the k-means algorithm the result clusters are : CL(C_1^1), CL(C_2^1), CL(C_3^1), …… , CL($C_{N_1}^1$)and the new centers at the result are: $newC_1^1, newC_2^1, newC_3^1, \ldots\ldots, newC_{N_1}^1$.If there is no confusion, let us rename them again by the same notations$C_1^1, C_2^1, C_3^1, \ldots\ldots, C_{N_1}^1$ respectively, the new set of N_1 leaders after this second time execution of k-means algorithm in this step.
Thus at the end of step-1, the N number of big data B are clustered into N_1 number of clusters as shown in Figure 3.

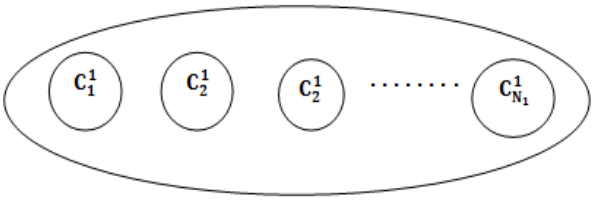

Fig. 3: Clustering of B IntoN_1Number of Clusters In Step-1.

Step-2:
In each step from Step-2 onward the k-means algorithm will be applied twice to two different data sets.In this step, first we apply k-means algorithm with$k = \alpha^{n-1} = N_2$to cluster theN_1 centers $C_1^1, C_2^1, C_3^1, \ldots C_{N_1}^1$as the data points.Suppose that on completion of the execution of the k-means algorithm the new centers at the result are:$C_1^2, C_2^2, C_3^2, \ldots C_{N_2}^2$.
Next we apply k-means algorithm to cluster B into N_2 number of clusters starting with the new leaders :$C_1^2, C_2^2, C_3^2, \ldots C_{N_2}^2$.Suppose that on completion of the execution of the k-means algorithm the result clusters are : CL(C_1^2), CL(C_2^2), CL(C_3^2), …… , CL($C_{N_2}^2$)and the new centers at the result are:$newC_1^2, newC_2^2, newC_3^2, \ldots\ldots, newC_{N_2}^2$.If there is no confusion, let us rename them again by the same notations$C_1^2, C_2^2, C_3^2, \ldots C_{N_2}^2$ respectively, the new set of N_2 leaders after this second time execution of k-means algorithm in this step.
Thus at the end of step-2, the N number of big data B are clustered into N_2 number of clusters as shown in Figure 4.

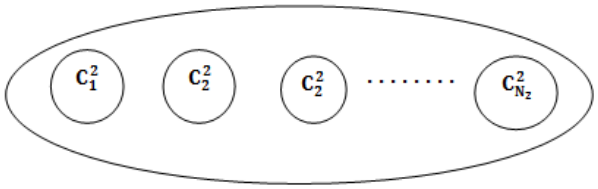

Fig. 4: Clustering of B IntoN_2Number of Clusters in Step-2.

Step-3:

In this step, first we apply k-means algorithm with$k = \alpha^{n-2} = N_3$to cluster the N_2 centers $C_1^2, C_2^2, C_3^2, \ldots C_{N_2}^2$ as the data points. Suppose that on completion of the execution of the k-means algorithm the new centers at the result are:$C_1^3, C_2^3, C_3^3, \ldots \ldots, C_{N_3}^3$.

Next we apply k-means algorithm to cluster B into N_3 number of clusters starting with the new leaders :$C_1^3, C_2^3, C_3^3, \ldots \ldots, C_{N_3}^3$.Suppose that on completion of the execution of the k-means algorithm the result clusters are : CL(C_1^3), CL(C_2^3), CL(C_3^3), …… , CL($C_{N_3}^3$),and the new centers at the result are:new$C_1^3, newC_2^3, newC_3^3, \ldots \ldots, newC_{N_3}^3$.If there is no confusion, let us rename them again by the same notations$C_1^3, C_2^3, C_3^3, \ldots \ldots, C_{N_3}^3$respectively, the new set of N_3 leaders after this second time execution of k-means algorithm in this step.

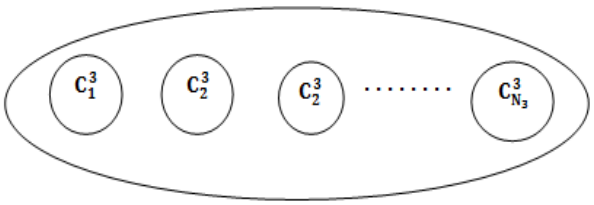

Fig. 5: Clustering of B IntoN_3Number of Clusters.

Thus at the end of step-3, the N number of big data B are clustered into N_3 number of clusters as shown in Figure 5.And so on,till Step-n (n number of steps will be continued).

Step-(n-1):

This is the last but one step.In this step, first we apply k-means algorithm with $k = \alpha^2 = N_{n-1}$ to cluster the N_{n-2}centers $C_1^{n-2}, C_2^{n-2}, C_3^{n-2}, \ldots C_{N_{n-2}}^{n-2}$ as the data points. Suppose that on completion of the execution of the k-means algorithm the new centers at the result are:$C_1^{n-1}, C_2^{n-1}, C_3^{n-1}, \ldots C_{N_{n-1}}^{n-1}$.

Next we apply k-means algorithm to cluster B into N_{n-1} number of clusters starting with the new leaderss:$C_1^{n-1}, C_2^{n-1}, C_3^{n-1}, \ldots C_{N_{n-1}}^{n-1}$.Suppose that on completion of the execution of the k-means algorithm the result clusters are : CL(C_1^{n-1}), CL(C_2^{n-1}), CL(C_3^{n-1}), …… , CL($C_{N_{n-1}}^{n-1}$),and the new centers at the result are:$newC_1^{n-1}, newC_2^{n-1}, newC_3^{n-1}, \ldots \ldots, newC_{N_{n-1}}^{n-1}$.If there is no confusion, let us rename them again by the same notations $C_1^{n-1}, C_2^{n-1}, C_3^{n-1}, \ldots \ldots, C_{N_{n-1}}^{n-1}$ respectively, the new set of N_{n-1} leaders after this second time execution of k-means algorithm in this step.

At the end of Step-(n-1), the N numbers of big data B are clustered intoN_{n-1} number of clusters (see Figure 6).

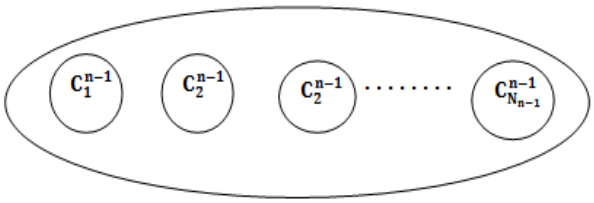

Fig. 6: Clustering of B IntoN_{n-1} Number of Clusters.

Step-n:

This is the last step.In this step, first we apply k-means algorithm with $k = \alpha = N_n$ to cluster the N_{n-1}centers $C_1^{n-1}, C_2^{n-1}, C_3^{n-1}, \ldots \ldots, C_{N_{n-1}}^{n-1}$ as the data points.Suppose that on completion of the execution of the k-means algorithm the new centers at the result are:$C_1^n, C_2^n, C_3^n, \ldots C_{N_n}^n$.

Next we apply k-means algorithm to cluster B into N_n number of clusters starting with the new leaders:$C_1^n, C_2^n, C_3^n, \ldots \ldots, C_{N_n}^n$.Suppose that on completion of the execution of the k-means algorithm the result clusters are : CL(C_2^n), CL(C_3^n), …… , CL(C_k^n),and the new centers at the result are:newC_1^n, newC_2^n, newC_3^n, … new$C_{N_n}^n$.If there is no confusion, let us rename them again by the same notations $C_1^n, C_2^n, C_3^n, \ldots \ldots, C_{N_n}^n$ respectively, the new set of N_n leaders after this second time execution of k-means algorithm in this step.

At the end of Step-n, the N numbers of big data B are clustered into α number of clusters (see Figure 7).

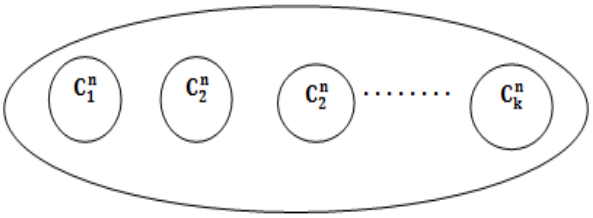

Fig. 7: Clustering of B Into A (= K) Number of Clusters.

It may be noted that in Step-n here, apparently it seems that we apply k-means algorithm directly to cluster B. but a careful observation will reveal that it is not so in reality. A lot of demerits are reduced in prior steps before arriving at Step-n in our proposed (N,α)-means Algorithm.

Summary Diagram of"(N,α)-means Algorithm" for big data:

The complete process (all the steps) is shown in Figure 8 below.

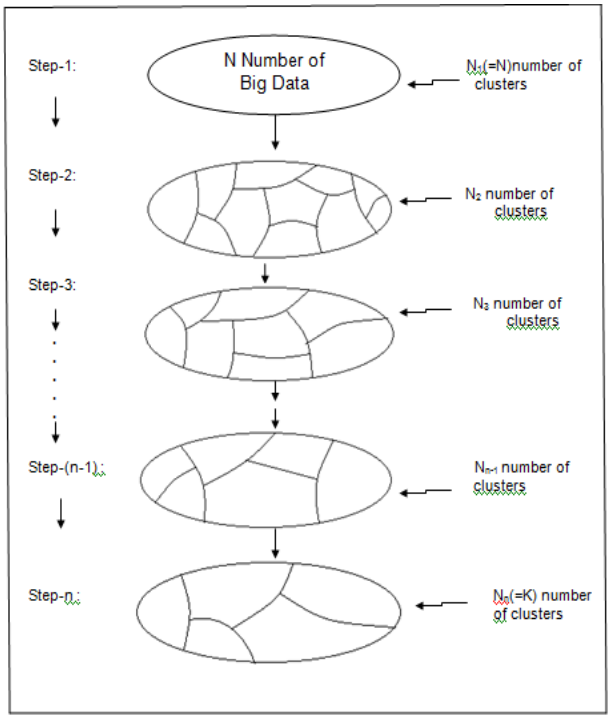

Fig. 8: All the N Steps of (N, A)-Means Algorithm for Clustering Big Data B.

The algorithm is presented below, for which the inputs are B, N and αas introduced above, and the output is the collection CL of α number of clusters of B.

(N-k)-Algorithm (B, N, α, CL)

1) n=Floor($\log_\alpha N$)
2) $N_1=\alpha^n$
3) C=any sub-multiset of Bof cardinality N_1(to be treated as initial leaders)

4) Apply k-means algorithm to cluster the leader-set C into k number of clusters
5) CL=collection of all the k clusters
6) C=collection of revised leaders
7) Fori=2 to n,Dofrom Step-8 to Step-13
8) $N_i=\alpha^{(n+1)-i}$
9) $k=N_i$
10) Apply k-means algorithm to cluster the leader-set set C into k number of clusters
11) C=collection of revised leaders of Step-10
12) Apply k-means algorithm to cluster Binto k number of clusters starting with the leaders of C
13) C=collection of revised leaders of Step-12
14) CL=Collection of all the k clusters of B
15) Stop

3. Inappropriateness of k-means algorithm for clustering big data: a justification

In subsection 1.1 earlier, it is mentioned about the demerits of k-means algorithm. Although these are well known demerits, but for our problem (subsection 1.2) here the consequences of these demerits are much more significant. In this section we consider one issue only out of many issues. For this we present a simple justification to show that the k-means Algorithm is not always appropriate for big data, but the philosophy of our proposed (N,α) algorithm can well address the situation in many cases.It is known that k-means algorithm [11]updates the placement of a data point x during its execution, placing x from one cluster to another newly formed cluster according to its revised closeness. In this section our issue is not only about updation of the placement ofx, but about more appropriateness too.

For the sake of presentation here, we consider a very simple problem:"to partition N big data into two clusters (here 2 << N)".

It is obvious that the possibility of tie cases of the value d(x,C) to occur in big data is more. For clarification we present here two cases, Case-1 and Case-2 independently.In Case-1 we apply k-means algorithm directly and form two clusters. In Case-2 we apply the philosophy of our proposed (N,α)-means algorithm to justify that clustering of big data can be done in a better way. However, in this simple problem we will show the merit of our philosophy by applying two steps only in Case-2 here;i.e.in step-1 of Case-2 we initially make three clusters and then in step-2 of it we finally make two clusters, which is our desired goal here.

One major advantage in (N,α) algorithm is that a large number of data points are initially given opportunity to become leaders, and because of variable but reducing nature of the parameter k (whose value become equal to α at the last step only) the leaders loose their appropriateness to lead. In this sense this method can be regarded to be more democratic.

Case-1:Clustering of big data by direct application of k-means algorithm.

In this case, k-means algorithm is applied directly so that in one go, all the N data are clustered into two clusters CL(C_1) and CL(C_2).But there is a possibility that during the execution of the k-means algorithm tie-up cases have occurred corresponding to few data points. For example (see Figure 9), for the data x a tie-up case has occurred during the construction of the clusters CL(C_1) and CL(C_2), because of the reason that

d (x, C_1) = 7 = d (x, C_2)

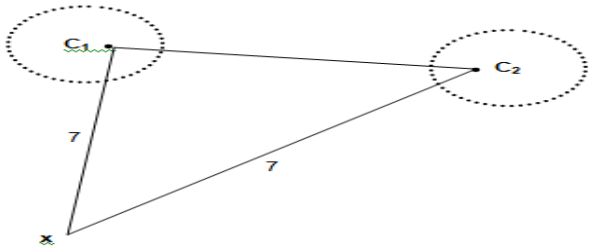

Fig. 9: A Tie Case Where CL (C1) and CL (C2) are Under Construction.

Hence x is put in any one (not in both) of the clusters CL(C_1) and CL(C_2).With no loss of generality, let us consider that x is in CL(C_1).Consequently, the final result of k-means algorithm applied directly on N big data yields in two clusters CL(C_1) and CL(C_2)as shown in Figure 10 below.

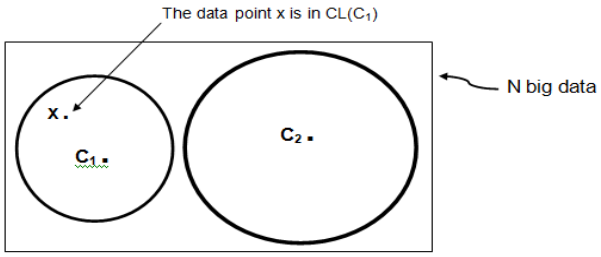

Fig. 10: The K-Means Algorithm Yields the Above Two Clusters CL(C1) and CL(C2) If Applied to N Big Data Directly.

Case-2:Clustering of big data using the philosophy of (N,α)-means algorithm.

We consider the same problem as above, a very simple problem, whichis"to partition N big data into two clusters (here 2 << N)".The case-2 consists of two steps as stated below.

Step-1:

In this step, we apply the k-means algorithm to cluster N big data initially into three (3) clusters CL (C_1),CL(C2) andCL(C_3).The Figure 11 shows the three clusters under-construction, and suppose that at this very moment the belongingness of the data point x is under consideration.

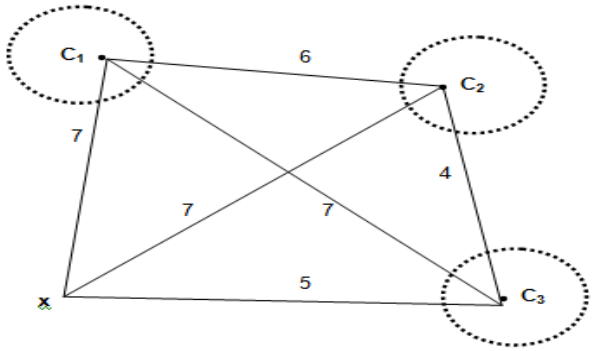

Fig. 11: Three Clusters CL (C_1), CL (C_2) and CL (C_3) are Under-Construction.

See Figure 11 above.During execution, x being closest to C_3, the element x thus going to belong to CL(C_3) in step-1. At the end in step-1 the three clusters are formed which are CL(C_1), CL(C_2) and CL(C_3)as shown below in Figure 12.

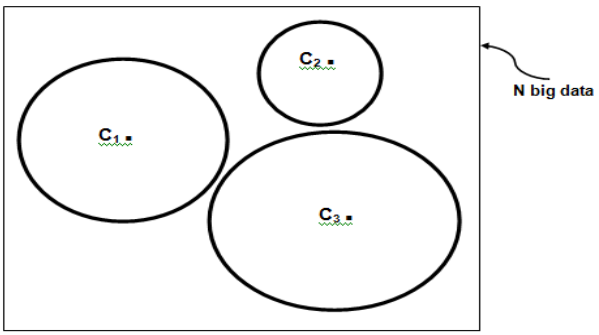

Fig. 12: Result of Step-1 Yielding Three Clusters CL (C_1), CL (C_2) and CL (C_3).

However, consider now an alternative (see Figure 13) for another situation.During execution, x being closest to C_1 and C_2 both, the element x thus may belong to CL(C_1) or CL(C_2) but not in both in step-1. With no loss of generality, let us put x in CL(C_1).

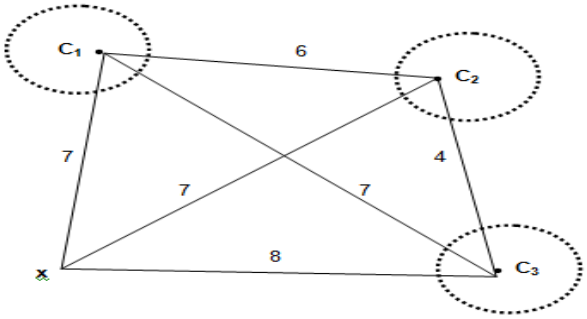

Fig. 13: Initially Three Clusters CL (C_1), CL (C_2) and CL (C_3) are Under-Construction.

At the end in step-1 for this situation too, the three clusters are formed which are CL(C_1), CL(C_2) and CL(C_3)as shown below in Figure 14 below. Obviously, the three clusters CL(C_1), CL(C_2) and CL(C_3) of Figure 12andof Figure 14need not be same.

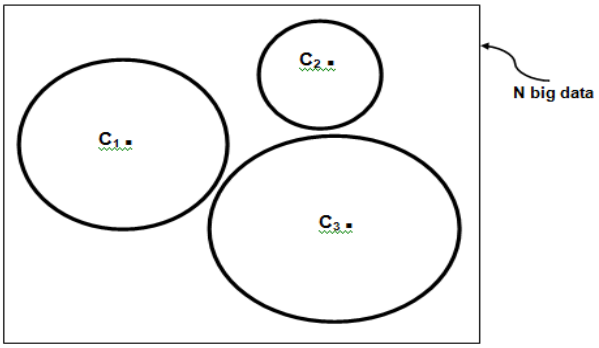

Fig. 14: Result of Step-1 Yielding Three Clusters CL (C_1), CL (C_2) and CL (C_3), (Different from Figure 12).

Step-2:
We begin our action in Step-2 from Figure 12 (or from Figure 14).This step consists of sub-steps, where too we deviate from the existing k-means algorithm.
Substep-2.1:
We begin with three data which are the three centers C_1, C_2 and C_3; and then apply k-means algorithm to yield two clusters of them. Automatically the data points C_2 and C_3 jointly form a cluster whereas the other cluster is the data point C_1 only.
Substep-2.2:
Next we initialize with two new centres C_1 and C_{23}, where C_{23} is the calculated mean of the data points of the multiset {CL(C_2) ∪ CL(C_3)}.
Substep-2.3:
While considering the revised belongingness of the element x, it is now observed that (see Figure 15) the data point x is not going to belong to the new cluster CL(C_1). This shows very precisely and without any tie-up case that the element x has been the non-resident of the cluster CL(C_1)unlike in Case-1 above. and it is thus obvious that the (N,α)-means method can provide more appropriate placement of the data element xcompared to the Case-1.

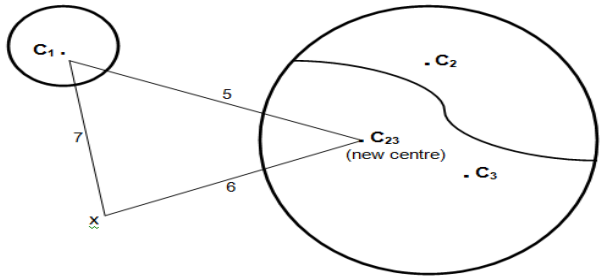

Fig. 15: The Data Point X Does Not Happen to Belong to the Cluster CL (C_1); The Two Clusters Under-Construction are CL (C_1) and CL (C_{23}).

Thus, the Case-2 is more appropriate for clustering N number of big data, instead of applying the k-means algorithm directly (as done in Case-1 above).Final result is shown in Figure 16 where the data point x is not in the cluster CL(C_1).

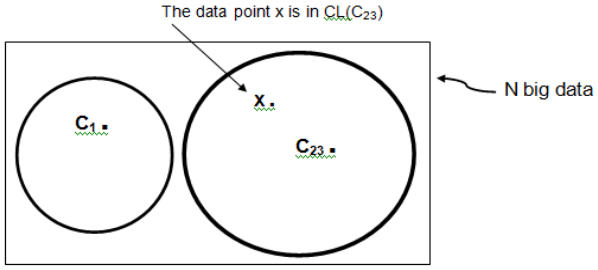

Fig. 16: Final Result of (N, A) Algorithm Yielding Two Clusters of N Big Data.

It is obvious that in this case the final result of Case-2 is not same as the final result of Case-1, although in both the cases we have clustered N big data into two clusters.

4. Conclusion

The existing k-means algorithm, which is an excellent algorithm to cluster data, is not always suitable to cluster big data if k is too small compared to the size N of big data B.The demerits are stated in subsection 1.1.One justification is given by a simple example by discussing one of the many issues. Consequently, a slightly modified version of it is developed by developing a new algorithm called by '(N,α)-means Algorithm' which can play a better role to cluster big data. Even for classical data sets (if not big data), wherever k-means algorithm is applicable the'(N,α)-means Algorithm' is also applicable and more appropriately, but the converse is not true. All the merits of k-means algorithm are in-built in '(N,α)-means Algorithm' and the demerits are reduced by its way of construction. A sufficient number (>α) of data points are initially provided scopes to become leaders, and in this sense the '(N,α)-means Algorithm'is much more democratic. There are a number of popular data mining tools viz. Rapid Miner, Orange, Weka, Knime, Oracle Data Mining, etc. existing in the market which are being fluently used by the researchers, scientists, analysts around the world. It is claimed that an improved version of these can be well developed by incorporating the '(N,α)-means Algorithm' to support big data deals. However in our future research work we will go for implementation of this algorithm and for its complexity analysis using Hadoop, Cassandra and Quantcast File Systems [1], but under the ADS distributive system [4] using the exclusive data structures [5] for big data. Optimizing the performance of Hadoop clusters through efficient cluster management techniques has been proposed in [29]. We will extend the notion Hadoop clustering using our proposed '(N,α)-means Algorithm'in our next work.

References

[1] Ahad, Mohd Abdul and Biswas, Ranjit. (2017).Comparing and Analyzing the Characteristics of Hadoop, Cassandra and Quantcast File Systems for Handling Big Data, Indian Journal of Science and Technology, Vol.10(8),pp 1-6. https://doi.org/10.17485/ijst/2017/v10i8/105400.

[2] Anandan, R, S. Phani Kumar, S.,Kalaivani, K. and Swaminathan, P. (2018). A survey on big data analytics for enhanced security on cloud. International Journal of Engineering & Technology,Vol. 7, No. 2.21, pp. 331-334. ISSN 2227-524X.https://doi.org/10.14419/ijet.v7i2.21.12397.

[3] Biswas, Ranjit. (2016).Introducing 'NR-Statistics': A New Direction in Statistics,in G*eneralized and Hybrid Set Structures and Applications for Soft Computing:* edited by Sunil John, IGI Global, USA. https://doi.org/10.4018/978-1-4666-9798-0.ch023.

[4] Biswas, Ranjit. (2015). "Atrain Distributed System" (ADS): An Infinitely Scalable Architecture for Processing Big Data of Any 4Vsin *Computational Intelligence for Big Data Analysis Frontier Advances and Applications*: edited by D. P. Acharjya, SatchidanandaDehuri and SugataSanyal, Springer International Publishing, Switzerland 2015, Part-1, pp 1-53.

[5] Biswas, Ranjit. (2016). Introducing data structures for big data, Chapter-2 in *Effective Big Data Management and Opportunities for Implementation*, edited by Manoj Kumar Singh and Dileep Kumar, IGI Global (USA). https://doi.org/10.4018/978-1-5225-0182-4.ch002.

[6] Copson, E.T. (1968). *Metric Spaces.*Cambridge University Press. https://doi.org/10.1017/CBO9780511566141.

[7] Elgendy N. and Elragal A. (2014). Big data analytics: A literature review paper. Advances in data mining. Applications and theoretical aspects. Lecture Notes in Computer Science.pp 214-227. https://doi.org/10.1007/978-3-319-08976-8_16.

[8] EndahHiswati, Marselina. Achmad Fanany OnnilitaGaffar,Rihartantoand Haviluddin. (2018). Minimum wage prediction based on K-Means clustering using neural based optimized Minkowski Distance Weighting. International Journal of Engineering & Technology,Vol. 7, No. 2.2, pp. 90-93. ISSN 2227-524X.

[9] Gomathi, S.; Narayani, V.(2017). Early prediction of systemic lupus erythematosus using hybrid K-Means J48 decision tree algorithm. International Journal of Engineering & Technology, Vol. 7, No. 1.3, pp. 28-32. ISSN 2227-524X.

[10] Guha, S., Rastogi, R. (2001): An efficient clustering algorithm for large database. Inf. Syst. 26(1), 35–58. https://doi.org/10.1016/S0306-4379(01)00008-4.

[11] Han, J., Kamber, M., Pei, J. (2006). Data mining: concepts and techniques. Morgan Kaufmann.

[12] Hashmi A, S. and Ahmad T. (2016).Big data mining techniques. Indian Journal of Science and Technology, Vol.9 (37),pp 1-5.

[13] Havens, T.C., Bezdek, J.C., Palaniswami, M. (2013).Scalable single linkage hierarchical clustering for big data.in: 2013 IEEE Eighth International Conference on Intelligent Sensors, Sensor Networks and Information Processing, pp. 396–401. IEEE.https://doi.org/10.1109/ISSNIP.2013.6529823.

[14] Heeku, Jin; Su Jeong, Yoon. (2018). A study on social big data analysis using text clustering. International Journal of Engineering & Technology, Vol.7, No.2.12, pp. 1-4. ISSN 2227-524X.

[15] Kaufman, L., Rousseeuw, P.J. (1990).Finding Groups in Data: An Introduction on Cluster Analysis. John Wiley and Sons. https://doi.org/10.1002/9780470316801.

[16] Kodali, Sadhana; Dabbiru, Madhavi; Rao, B Thirumala. (2018). A Survey of Data Mining Techniques on Information Networks. International Journal of Engineering & Technology, Vol. 7, No. 2.6, pp. 293-300. ISSN 2227-524X.https://doi.org/10.14419/ijet.v7i2.6.11267.

[17] Kusuma, S; KasiViswanath, D. (2018). IOT and Big Data Analytics in E-Learning: A Technological Perspective and Review. International Journal of Engineering & Technology, Vol.7, No.1.8, pp. 164-167. ISSN 2227-524X.

[18] Mhaske-Dhamdhere, Vidya; Vanjale, Sandeep. (2017). A novel approach for phishing emails real time classification using k-means algorithm. International Journal of Engineering & Technology, Vol. 7, No. 1.2, pp. 96-100. ISSN 2227-524X.

[19] Nafis, Md Tabrez and Biswas, Ranjit. (2018). A Secure Clustering Technique for Unstructured and Uncertain Big Data,in*Progress in Advanced Computing and Intelligent Engineering, Advances in Intelligent System and Computing* – 564, edited by Khalid Saeed, NabenduChaki, BibudhenduPati, SambitBakshi, Durga Prasad Mohapatra, Springer Nature Singapore,Vol.2, Part-III, pp 459-466.

[20] RatnaBabu, P; Bhanu Prakash Battula. (2018).A novel k-nearest neighbor distance based under sampling for improved opinion mining on skewed data using random forest. International Journal of Engineering & Technology, Vol.7, No.1.8, pp. 62-66. ISSN 2227-524X.

[21] SagarImambi, S., P.Vidyullatha, P., Santhi, M.V.B.T. and Haran Babu, P.(2018). Explore Big Data and Forecasting Future Values using Univariate Arima Model in R.International Journal of Engineering & Technology, Vol. 7, No. 2.7, pp. 1107-1110. ISSN 2227-524X.

[22] Sakthivel, N K; Gopalan, N P; Subasree, S. (2018). Parallel framework based gene signature-hierarchical random forest cluster for predicting human diseases. International Journal of Engineering & Technology, Vol. 7, No. 2.27, pp. 12-16. ISSN 2227-524X.https://doi.org/10.14419/ijet.v7i2.27.12103.

[23] Sankaramalladi, Bhima; Srinivas Prasad. (2017). big data life cycle: security issues, challenges, threat and security model. International Journal of Engineering & Technology, Vol.7, No.1.3, pp. 100-103. ISSN 2227-524X.https://doi.org/10.14419/ijet.v7i1.3.9666.

[24] Seenu, Aaluri;Kameswara Rao, M. (2018). A Novel Privacy Preserving Data mining using improved decision tree and KP-ABE on High Dimensional Data. International Journal of Engineering & Technology, Vol.7, No.2.7, pp. 515-519. ISSN 2227-524X.https://doi.org/10.14419/ijet.v7i2.7.10874.

[25] Shirkhorshidi, Ali Seyed., Aghabozorgi, Saeed,Wah, Teh Ying.,Herawan, Tutut. (2014). Big Data Clustering A Review,Proceedings of the International Conference on Computational Science and Its Applications ICCSA 2014: Computational Science and Its Applications–ICCSA (2014) pp 707-720.

[26] Shobana, G; Prakash Battula, Bhanu. (2018). an under sampled k-means approach for handlingimbalanced data using diversified distribution. International Journal of Engineering & Technology, Vol.7, No.1.8, pp. 113-117. ISSN 2227-524X.

[27] Simmons, G.F. (1963).*Introduction to Topology and Modern Analysis.*McGraw Hill, New York.

[28] Suhailan, S et al. (2018). A hybrid model of ordinal ranking-based clustering using G+Rank K-Means. International Journal of Engineering & Technology, Vol. 7, No. 2.15, pp. 41-44. ISSN 2227-524X.https://doi.org/10.14419/ijet.v7i2.15.11209.

[29] S. Shraddha Bollamma, K; Manishankar, S; V. Vishnu, M. Optimizing the performance of hadoop clusters through efficient cluster management techniques. International Journal of Engineering & Technology, Vol.7, No.2.31, pp. 19-22. ISSN 2227-524X.

[30] Tremblay,J.P. & Manohar,R. (1987), *Discrete Mathematical Structures with Applicationsto Computer Science.* McGraw Hill Int. Ed.

[31] Tryon. (1939). *Cluster Analysis.* McGraw-Hill Publishers, New York.

[32] VishwanathBrahmane, Anilkumar; Murugan, R. (2018). Parallel processing on Big Data in the context of Machine Learning and Hadoop Ecosystem: A Survey. International Journal of Engineering & Technology, Vol.7, No.2.7, pp. 577-588. ISSN 2227-524X.

[33] Zhang, T., Ramakrishnan, R., Livny, M. (1996). BIRCH: An efficient data clustering method for very large database. In: SIGMOD Conference, pp. 103–114.https://doi.org/10.1145/233269.233324.

[34] Zhao, W., Ma, H., He, Q. (2009). Parallel k-means clustering based on MapReduce. In: Cloud Computing, pp. 674–679.https://doi.org/10.1007/978-3-642-10665-1_71.