An Enhanced Method of Clustering for Big Data Mining using K-Means

[1]Dr.K.P.N.V.Satya Sree, [2]Dr.S.M Roy Choudri
[1,2]Professor
[1,2]Department of Computer Science & Engineering
[1,2] Usha Rama College of Engineering and Technology Telaprolu, Krishna Dt., A.P, India

Abstract : For to examining big data Clustering is a basic information mining and apparatus. There are troubles for applying grouping procedures to huge information twosome to new difficulties that are raised with huge information. As Big Data is alluding to TBs and PBs of information and grouping calculations are accompanied great computational outlays, the inquiry is the means by which to adapt to this issue and how to convey bunching methods to enormous information and get the outcomes in a sensible time. K-Means which is a standout amongst the most utilized bunching strategies and K-Means in view of MapReduce is considered as a propelled answer for substantial dataset grouping. Be that as it may, the executing time is as yet an obstruction because of the expanding quantity of iterations when there is an expansion of dataset extent and number of groups. This paper exhibits another approach for diminishing the quantity of emphases of K-Means calculation which can be connected to expansive dataset grouping. And furthermore this technique introduce plan to redress the issues related with k-implies significantly with the centroid choice issue. This strategy can likewise ensure the base computational time and increment in the exactness of results.

IndexTerms **- K-means. Big Data, Data Mining, Clustering.**

I. INTRODUCTION

Previous decades have seen an emotional augment in our ability to assemble data from various sensors, contraptions, in different associations, from free or related applications. This data surge has outpaced our ability to process, separate, store and fathom these datasets. For to separate examples frame the informational collections the understand innovation is information mining. Information Mining is an effective new innovation to remove concealed prescient data from vast databases. It causes organizations to canter around the most basic information in their data appropriation focuses. Data mining devices foresee future examples and practices with which agents can make proactive, learning driven decisions. The modernized, up and coming examination offered by Data Mining [1] moved past the examination of the past events gave by audit gadgets. Interestingly, enormous information mining posture new difficulties of huge information have root in its five vital qualities:

Figure 1: Five Views of Big Data.

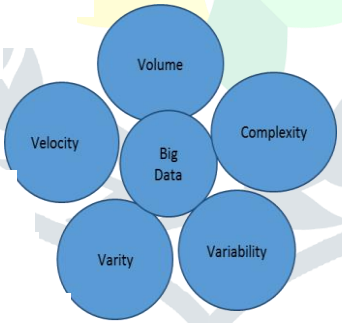
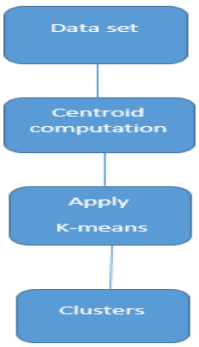


Figure 1 shows the characteristics of big data or the challenges which pose by big data and those are explained below.

- Volume: The earliest is Volume and an illustration is the un-structured information gushing In type of web-based social networking and it enlargements question, for example, how to decide the importance Inside huge information volumes and how to examine the important information to create Valuable data.
- Velocity: Data is engulfing at rapid and it must be managed in sensible time. Reacting rapidly to information speed is one of the difficulties in enormous information.
- Variety: Another testing issue is to oversee, blend and administer information that originates from various sources with various details, for example, email, sound, unstructured information, social information, video and so on.
- Variability: Inconsistency in information stream is another test. For instance in online networking it could be day by day or occasional pinnacle information oodles which marks it harder to bargain and deal with the information uniquely when the information is unstructured.
- Complexity: Records is originating from various sources and have diverse structures; thus it is important to interface and connect connections and information linkages or you observe your information to be crazy rapidly.

Figure 2: K-Means Work Flow.



The traditional techniques of data mining, failed handle the data which is in semi structured and unstructured format. For to handle those kind of formats we need to modify the existing data mining methodologies, in this paper we are concentrating on the clustering techniques which are more suitable to mine and extract patterns from unstructured data. Presently, more number of peoples are concentrating on the k- means clustering because of its simplicity and adaptability of the algorithm. In contrast, k-means does not support all kind of data that is unstructured data semi structured data. In order to handle those modification of k-means is required in this paper we are concentrating on how to enhance the k-means which will able to handle all kinds of data. Figure-2 demonstrations the working mechanism of a k-means where initially k-means takes a data set and computes the centroid normally it will randomly select a value in a data set as a centroid and then computes the clusters. The remainingarticle is prepared as tracks section-II compacts with allied work, section-III describes the projected methodology, section-IV shows the experimental setup, section-V illustrates the results and compared with the existing mechanisms and finally section-VIcompletes the paper.

II. RELATED WORK

Arshad Muhammad Mehar et al.[1] built up another technique in view of internal validation measures, in order to locate an ideal estimation of k which, can give more steady groups. The proposed bunch legitimacy measure is utilized to figure the extent of basic articles in every pair of groups.

KA Abdul Nazeer et al.[2] introduced an upgraded k-means that includes sorting the information set and parceling the sorted information set into "k" number of sets which, brought about better starting centroids along these lines enhancing the precision of this algorithm. The algorithm converges speedier contrasted with traditional algorithm of K-Means. The main drawback of this algorithm is the estimation of k (number of sought groups) still should be given as an information.

Shi Na et al.[3] discussed an enhanced k-means calculation keeping in mind the end goal to tackle the issue of ascertaining the Euclidean separation between every information article and all group focuses in every emphasis, which expands the running time. In this approach a straightforward information structure is utilized to store some data in each emphasis, which can then be utilized as a part of the following cycle.

Shuhua Ren et al.[4] displayed a calculation CV-k-means i.e. coefficient of variety k-means algorithm. This helped in lessening the impacts of immaterial qualities brought on by taking Euclidean separation as the comparability measure by presenting variety coefficient weight vector. The main problem is that the quantity of craved clusters (k) is to be given as an information.

Kunhui Lin et al.[6] displayed an upgraded k-means clustering paper that optimized the starting focuses in light of data dimensional density which, affirm that these underlying focuses have the greatest contrast between groups. This algorithm is implemented on the Hadoop platform (MapReduce programming model). This methodology helped in enhancing the steadiness of the K-means clustering.

Anupama Chadha et al.[8] exhibited a calculation that does not require K (number of bunches) as an information. It expelled the reliance on K which, is in some cases exceptionally hard to foresee as it requires domain knowledge. The work is restricted to numeric information set as it were.

MadhuYedla et al.[9] proposed a paper on K-means possession in cognizance the end goal to locate the better initial centroids and thus lessens time unpredictability. The primary thought was that if the information point stays inside same bunch then the essentialinvolvedness lessens from O(k) to O(1). Subsequently the aggregate time unpredictability diminish to half i.e. for allotting the information directs it decreases toward O(nk) rather than O(nkl) which, brought about aggregate time taken to be O(nlogn). The limitation of this methodology was that the initialising the value of K was still required.

Z.Min et al.[10] acquainted a calculation with conquer the impediment of k-means++ approach by picking least change test as the principal starting grouping focus which, won't just dispose of the effect of confined focuses, additionally explores demonstrate that the enhanced calculation have bunching consequence of a moderately steady and better dependability and precision. The issues of such algorithm contains (a) Time utilization issue brought by the complexity of the approach used (b) how to keep away from regular computation issues if there should be an occurrence of a lot of information, and so on.

Soumi Ghosh et al.[11] showed a comparative study between KM and FCM based on the number of samples and K. The experimental results show that the K-means algorithm is far better than FCM as it takes more time in performing fuzzy measure calculations which, results in increase in its time complexity and hence effects the result. Hence, no doubt FCM produces as good results as produced by KM close results but the time complexity is comparatively still high.

III. PROPOSED METHOD

It is plainly that the K-Means bunching algorithms an outstanding grouping strategy in any case, two hindrances exist for extensive datasets grouping. The principal impediment is computational intricacy of separation counts, which ascertains removes

between information tests to clusters. This trouble can be overwhelmed by applying the MapReduce model to disseminate calculations to numerous specialists in a dispersed situation. Be that as it may, this hindrance is still there when information estimate increments exponentially.

The second hindrance is the quantity of rounds which essentially increments when the quantity of test information increments. This issue might be tackled by utilizing two-phase K-Means calculation or K-Means plus plus calculation [19]. K-Means plus plus comprises of two stages: the initial step is to choose better beginning mid points and the second step is the KMeans.

K-meansalgorithm is utilized for input space division into a few subspaces. It is iterative, data allocating estimation that doles out n discernments to exactly one of k bundles.k is picked from the earlier before the calculation begins. Each group is characterized by centroid. In the fundamental adaptation centroid is figured as a mean of all information directs having a place toward the bunch. The calculation continues as takes after:

Improved k means for accurate clusters

Input:

D = do1, do2… don / set of n data objects.

K // it is the desired number of clusters.

Output:

A set of K number of clusters.

Step-1 Centroid computations

1 Data set D

2 Select a data point C_i

3 Start computing distance from C_i to all other data points $x_1, x_2 \ldots\ldots x_n$

4 Distance $d(x_i, C_i)=Sqrt((C_{i1}-x_{i1})^2)+ (C_{i2}-x_{i2})^2))$

5 Apply quick Sort ()

Step-2 Apply K-means

6 Pick K- Initial Centroids based on the distances divide the sorted data points into k number of – equalPartitions.

7 Recalculate the centre of each cluster only based on the data in the cluster.

8 Repeat line 6 & line 7 until convergence

9When the new cluster centres are the same as the cluster centres obtained in previous iteration, output the clustering results;

IV. Experimental Setup

The calculation is assessed utilizing an independent machine with 16 GB RAM, 1 TB HDD, Intel fifth era processor and Ubuntu 16.04 LTS machine is utilized Hadoop delineate writing computer programs is utilized to do the trials. The proficiency of proposed calculation is assessed by leading examinations on five counterfeit informational collections, three genuine datasets down stacked from the site UCI and two microarray informational indexes (two yeast informational collections) downloaded from http://www.cs. washington.edu/homes/kayee/group.

The genuine informational collections utilized:

1. Iris plants database (n = 149, d = 5, K = 3)
2. Glass (n = 218, d = 8, K = 5)
3. Wine (n = 179, d = 12, K = 2)

The genuine microarray informational collections utilized:

The yeast cell cycle information demonstrated the vacillation of articulation levels of roughly 6001 qualities more than two cell cycles (18 time focuses).

1. The main subset comprises of 384 qualities whose articulation levels crest at various time directs relating toward the five periods of cell cycle.
2. The subsequent subset comprises of 230 qualities relating to four classifications in the MIPS catalogue. The four classes (DNA combination and replication, association of centrosome, nitrogen and sulphur digestion, and ribosomal proteins) were appeared to be reproduced in bunches from the yeast cell cycle information

V. Results & Comparison

This section presents a comprehensive view of the proposed method and the existing methods that k-means and k-means++. And here made comparison based on the performance metrics that total computation time, centroid selection time and accuracy of each method on different data sets.

Figure 3 describes the proportional analysis of traditional k-means clustering algorithm, K-means plus plusalgorithm and proposed method with respect to the total cluster formation time that is the time taken for centroid selection time and the cluster formation time. Here proposed method out performed that the two existing techniques.

Figure 3: Total Time for formation of Cluster.

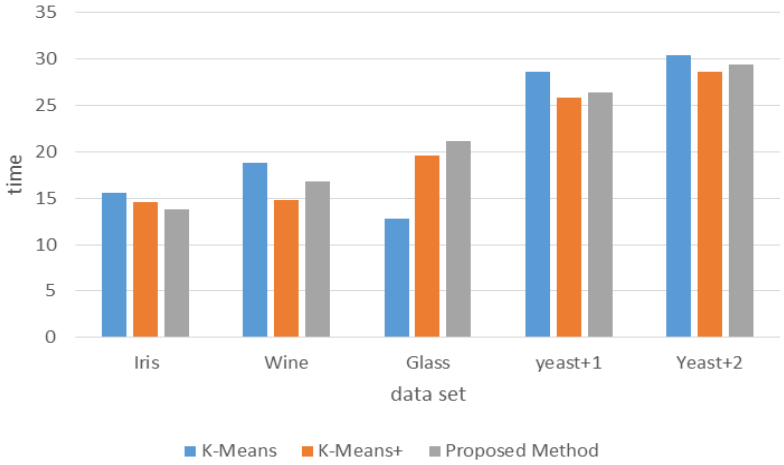


Figure 4 describes the proportional analysis of traditional k-means clustering algorithm, K-means plus + algorithm and proposed method with respect to the Accuracy. Here proposed method out performed that the two existing techniques.

Figure 4: Accuracy Comparison.

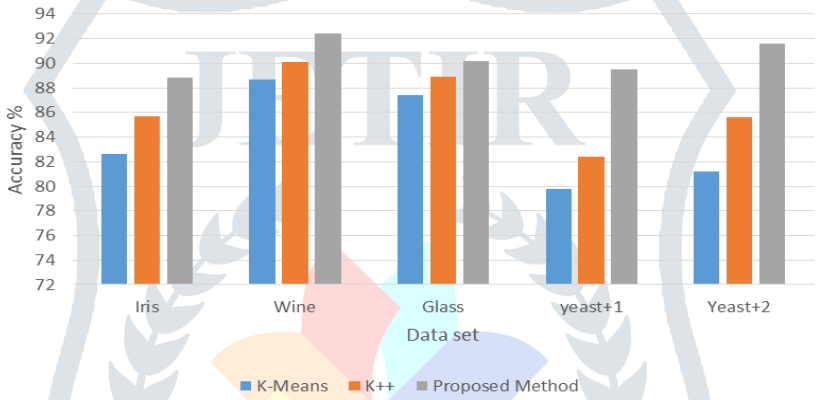
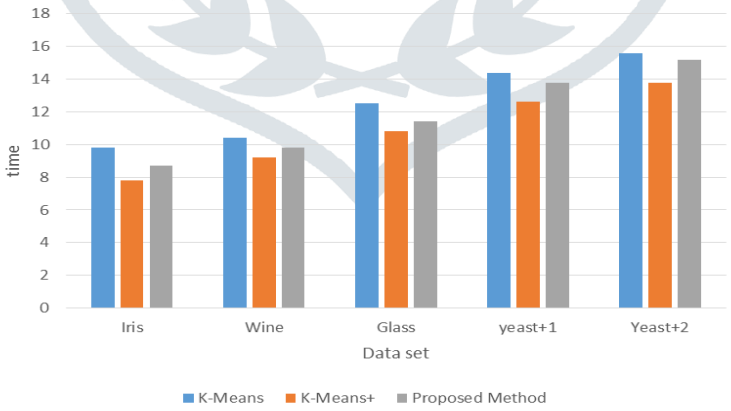


Figure 5 describes the proportional analysis of traditional k-means clustring algorithm, K-means plus plus algorithm and proposed method with respect to the centroid computation time. Here proposed method out performed that the two existing techniques.

Figure 5: Computation Time for Each Centriod.



VI. Conclusion

Here in this method of modified k-means method works well in all aspects because here the major advantage is identifying the accurate centroids in a very faster manner with the parallel processing programming method map-reduce, because this programming model is a fitting decision for huge dataset grouping employments. Contrasted with the past strategies, for example, grouping utilizing k-implies k-means++ this new strategy can be considered as more critical in light of the fact that it can work with whole datasets, essentially diminish executing time, and give high precision. The test comes about demonstrates that the proposed strategy is better.

References

[1] .M. Mehar , K Matawie and A Maeder, "Determining an OptimalValue of K in K-means Clustering" in IEEE InternationalConference on Bioinformatics and Biomedicine,2013: pp. 51-55.

[2] K A Abdul Nazeer, S D Madhu Kumar, "Enhancing the k-meansclustering algorithm by using a O(n logn) heuristic method forfinding better initial centroids" in Second International Conferenceon Emerging Applications of Information Technology, 2011: pp.261-264
[3] S. Na, L.Xumin and G.yong, "Research on k-means ClusteringAlgorithm" in IEEE Third International Symposium on IntelligentInformation Technology and Security Informatic,2010:pp. 63-67.
[4] S. Ren, A. Fan, "K-means Clustering Algorithm Based on Coefficient of Variation" in IEEE 4th International Congress onImage and Signal Processing, Vol. 4, 2011 :pp. 2076-2079.
[5] K. M. Kumar, Dr. A. R.M. Reddy, "A Fast K-Means ClusteringUsing Prototypes for Initial Cluster Center Selection", IEEE 9thInternational Conference on Intelligent Systems and Control(ISCO),2015:pp. 1-4.
[6] K.Lin, X.Li, J. Chen, Z. Zhang, "A K-means Clustering withOptimized Initial Center Based on Hadoop Platform" in The 9thInternational Conference on Computer Science & Education,2014:pp. 263-266.
[7] J. Xie, S. Jiang 2010, "A simple and fast algorithm for global Kmeansclustering" in IEEE Second International Workshop onEducation Technology and Computer Science, Vol. 2, 2010:pp. 36-40.
[8] A.Chadha, S. Kumar, "An Improved K-Means ClusteringAlgorithm: A Step Forward for Removal of Dependency on K" inIEEE International Conference on Reliability, Optimization andInformation Technology, 2014:pp. 136-140.
[9] M. Yedla, S. R. Pathakota, T M Srinivasa, "Enhancing K-meansClustering Algorithm with Improved Initial Center" in InternationalJournal of Computer Science and Information Technologies, Vol. 1(2), 2010:pp. 121-125.
[10] Z. Min, Kai-fei, "Improved research to k-means initial clustercenters" in Ninth International Conference on Frontier of ComputerScience and Technology, 2015:pp. 349-353.
[11] S. Ghosh, S.K. Dubey, "Comparative Analysis of K-Means andFuzzy C-Means Algorithms" in International Journal of AdvancedComputer Science and Applications, Vol. 4, 2013:pp. 35-39.
[12] R. Suganya, R. Shanthi, "Fuzzy C- Means Algorithm- A Review" inInternational Journal of Scientific and Research Publications, Vol2(11), 2012:pp. 1-3.
[13] http://archive.ics.uci.edu/ml/datasets/Iris
[14] S. Banerjee, A. Choudhary, S. Pal, "Empirical Evaluation of KMeans,Bisecting KMeans, Fuzzy C-Means and Genetic K-MeansClustering Algorithms" in IEEE International WIE Conference onElectrical and Computer Engineering,2015:pp. 168-172.
[15] L. Chen ; S. Chen, "New shadowed fuzzy C-means algorithm forimage segmentation" in 3rd International Conference on Informative&Cybernatics for Computational Social Systems,2016:pp 43-46.
[16] A. Gupta ; H. Shivhare ; S. Sharma, "Recommender system using fuzzy c-means clustering and genetic algorithm based weightedsimilarity measure" in International Conference on Computer,Communication & Control,2015:pp 1-8.
[17] Philip Chen, C.L., Zhang, C.-Y.: Data-intensive applications, challenges, techniques andtechnologies: A survey on Big Data. Information Sciences (in press, 2014)
[18] Barioni, M.C.N., Razente, H., Marcelino, A.M.R., Traina, A.J.M., Traina, C.: Open issues for partitioning clustering methods: an overview. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 4, 161–177 (2014)
[19] Hadian, A., Shahrivari, S.: High performance parallel k-means clustering for diskresidentdatasets on multi-core CPUs. The Journal of Supercomputing, 1–19 (2014) 298 D.V. Hieu and P. Meesad
[20] Bharill, N., Tiwari, A.: Handling Big Data with Fuzzy Based Classification Approach.In: Jamshidi, M., Kreinovich, V., Kacprzyk, J. (eds.) Advance Trends in Soft Computing.STUDFUZZ, vol. 312, pp. 219–227. Springer, Heidelberg (2014)
[21] Chen, M., Mao, S., Zhang, Y., Leung, V.M.: Chapter 1. Introduction. In: Big Data, pp.1–10. Springer, Heidelberg (2014)
[22] Jain, A.K.: Data Clustering: 50 Years Beyond K-means. In: Daelemans, W., Goethals, B., Morik, K. (eds.) ECML PKDD 2008, Part I. LNCS (LNAI), vol. 5211, pp. 3–4.Springer, Heidelberg (2008)
[23] Stoffel, K., Belkoniene, A.: Parallel k/h-Means Clustering for Large Data Sets. In:Amestoy, P.R., Berger, P., Daydé, M., Duff, I.S., Frayssé, V., Giraud, L., Ruiz, D. (eds.) Euro-Par 1999. LNCS, vol. 1685, pp. 1451–1454. Springer, Heidelberg (1999)
[24] Dean, J., Ghemawat, S.: MapReduce: simplified data processing on large clusters. Commun.ACM 51, 107–113 (2008)
[25] Zhao, W., Ma, H., He, Q.: Parallel K-Means Clustering Based on MapReduce. In:Jaatun, M.G., Zhao, G., Rong, C. (eds.) Cloud Computing. LNCS, vol. 5931, pp. 674–679. Springer, Heidelberg (2009)
[26] Lin, C., Yang, Y., Rutayisire, T.: A Parallel Cop-Kmeans Clustering Algorithm Basedon MapReduce Framework. In: Wang, Y., Li, T. (eds.) Knowledge Engineering andManagement. AISC, vol. 123, pp. 93–102. Springer, Heidelberg (2011)
[27] Lv, Z., Hu, Y., Zhong, H., Wu, J., Li, B., Zhao, H.: Parallel K-means clustering ofremote sensing images based on mapReduce. In: Wang, F.L., Gong, Z., Luo, X., Lei, J.(eds.) Web Information Systems and Mining. LNCS, vol. 6318, pp. 162–170. Springer,Heidelberg (2010)
[28] Manning, C.D., Raghavan, P., Sch¨utze, H.: K-Means. In: An Introduction to InformationRetrieval. Cambridge University Press (2009)
[29] Guha, S., Rastogi, R., Shim, K.: CURE: an efficient clustering algorithm for large databases. SIGMOD Rec. 27, 73–84 (1998)
[30] Har-Peled, S., Mazumdar, S.: On coresets for k-means and k-median clustering. Presented at the Proceedings of the Thirty-Sixth Annual ACM Symposium on Theory of Computing, Chicago, IL, USA (2004)[15] Jain, A.K., Dubes, R.C.: Chapter 3. Clustering Methods and Algorithms. In: Algorithmsfor Data Clustering, vol. Computer Science. Prentice Hall (1988)
[31] Anchalia, P.P., Koundinya, A.K., Srinath, N.K.: MapReduce Design of K-Means ClusteringAlgorithm. In: 2013 International Conference on Information Science and Applications(ICISA), pp. 1–5 (2013)
[32] Dom, B.E.: An Information-Theoretic External Cluster-Validity Measure. In: TheEighteenth Conference on Uncertainty in Artificial Intelligence (UAI 2002), Alberta, Canada, pp. 137–145 (2012)
[33] Wagner, S.,Wagner, D.: Comparing Clusterings - An Overview. Institute of TheoreticalInformatics (2007)
[34] Xu, Y., Qu,W., Li, Z.,Min, G., Li, K., Liu, Z.: Efficient k-means++ Approximation with MapReduce. IEEE Transactions on Parallel and Distributed Systems PP, 1–10 (2014)
[35]UCI.You Tube Multiview Video Games Dataset, YouTube+Multiview+Video+Games+Dataset
[36] UCI. Daily and Sports Activities, http://archive.ics.uci.edu/ml/datasets/Daily+and+Sports+Activities