International Journal of Engineering & Technology, 7 (4.5) (2018) 689-692

International Journal of Engineering & Technology

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

Research paper



An approach to achieve high efficiency for large volume data processing using multiple clustering algorithms

Sarada. B [1] *, Vinayaka Murthy. M [2], Udaya Rani. V [3]

[1] *Research scholar, REVA University, India*
[2] *Professor and Assistant director R&D, REVA University, India*
[3] *Associate Professor, School of C & IT,REVA University, India*
**Corresponding author E-mail: saradasaikonda@gmail.com*

Abstract

Now a days data is increasing exponentially daily in terms of velocity, variety and volume which is also known as Big data. When the dataset has small number of dimensions, limited number of clusters and less number of data points the existing traditional clustering algorithms will give the expected results. As we know this is the Big Data age, with large volume of data sets through the traditional clustering algorithms we will not be able to get expected results. So there is a need to develop a new approach which gives better accuracy and computational time for large volume of data processing. The Proposed new System Architecture is a combination of canopy, Kmeans and RK sorting algorithm through Map Reduce Hadoop frame work platform. The analysis shows that the large volume of data processing will take less computational time and higher accuracy, and the RK sorting does not require swapping of elements and stack spaces.

Keywords: *Big Data; Canopy Clustering; Hadoop; K-Mean Clustering; Data Processing Techniques; Mapreduce; Rk Sorting Algorithm.*

1. Introduction

Today, we live in the information age, and we are not running out of information particularly of data form because date is growing exponentially daily that is in terms of volume, variety, and velocity, therefore the existing clustering algorithm takes more time to pro- duce the results. To produce results in terms of less computational time and more efficiency one should think of something big and that is parallel programing. MapReduce is one of the programming designs for large volumes of datasets in parallel .MapReduce with HDFS can be used to handle the big data ,which is commonly known as Hadoop .Once the file is placed into HDFS it can be read n number of times.

2. Overview of clustering algorithms

Clustering is the task of dividing data points into a number of groups such that data points in the same groups are more similar to other data points in the same group than those in other groups. These groups are called clusters.
Types of Clustering
Clustering can be divided into two different types.

- Hard Clustering: Each data point either completely belongs to the cluster or not.
- Soft Clustering: Here putting instead of each data point into separate clusters, a probability or likelihood of that data point to be in those clusters is assigned

Types of clustering algorithms:
Models follow different set of rules to define the "similarity" among the data points. There are hundred clustering algorithms known. Few of them are as follows

Connectivity models: In this model the data points near in the data space will have more similarity hen data points which are far away These models can follow two approaches. 1) Starts with classifying all the data points into clusters and then merging them as the dis- tance decreases, 2) Classifying all the data points into single cluster and then partition as the distance increases.

- • Centroid models: These clustering models are iterative clustering algorithms in which similarity is measured by the minimum distance between the data points and cen- troid.
- • Distribution models: These clustering models are based on the probability that all the data points in the cluster belongs to the same distribution
- • Density models: These type of clustering models searches for various densities of data points in the data space. It isolates different density regions and assigns the data points with in these regions in the same cluster.

3. Existing clustering algorithms

Clustering algorithms are the best examples for unsupervised learn- ing algorithm. It is a simple approach to group data points or ob- jects. Here the groups are called clusters. The objects are data points which are in the cluster are similar than those in the other clusters.

3.1. Canopy clustering algorithm

Input: Dataset
Output: number of clusters
The algorithm uses two threshold values T1 and T2 Where T1 is loose distance and T2 is tight distance Where T1>T2
The steps involved in canopy clustering algorithm are

Copyright © 2018 Sarada.B et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Step 1: Randomly select any data point from the data Set as a canopy center
Step 2: Find the distance to all other points in the data Set from the canopy center.
Step 3: If the distance calculated is less than the T1 then put data points into a canopy
Step 4: Remove from data set all the points which are less then T2
Step 5: Repeat the above step1 to step 4 until the dataset becomes empty
Step 6: Feed the output as input K-mean clustering Algorithm
Limitations
Accuracy is low, but has the great advantage of its speed

3.2. k-mean clustering algorithm

K-means clustering algorithm is very simple and easy to understand. The steps involved in this algorithm are:
Step1: Randomly select the centroids and place them in space, which are temporary means of the cluster.
Step2: Calculate the Euclidean distance between each data point and cluster center. And then assign the data points to cluster centroid whose distance is minimum.
Step3: Recalculate the centroids for each cluster and replace by respective cluster centroid.
Step4: If there is no reassignment of the data point then go to next step otherwise go to step2
Step5: End
Limitations
Some of the drawbacks of existing k-mean algorithm through liter- ature survey are:

1) A review of uncertainty handling formalisms by A. Hunter and.

S. Parsons [6].In this paper computation time is reduced but initial centroids are selected randomly.

2) An overview from a database perspective by M. S. Chen, J. Han, and P.

S. Yu. [4]. in this paper author proposed the initial centroid algorithm to avoid selection of random centroid.

3) Efficient k-mean clustering algorithm for reducing the time complexity by D.Napoleon, P.Ganga Lakshmi. The authors say that reducing the time complexity is expensive for high dimensional data sets [3].
4) Overcoming the Defects of k-Means Clustering by using Canopy Clus- tering Algorithm by Ambika.s .s and Kavitha.G[1]. Avoided random selec- tion of centroid by using canopy clustering algorithm.

4. Proposed system architecture

The main aim of the proposed System is to find the initial values of centroids that is K value for K-mean clustering algorithm and stud- ying the space complexity and time complexity on Hadoop and MapReduce platform.
The Modules used in proposed system are

1) Big Data
2) Canopy Clustering Algorithm
3) k-Mean Clustering Algorithm
4) RK Sorting Algorithm

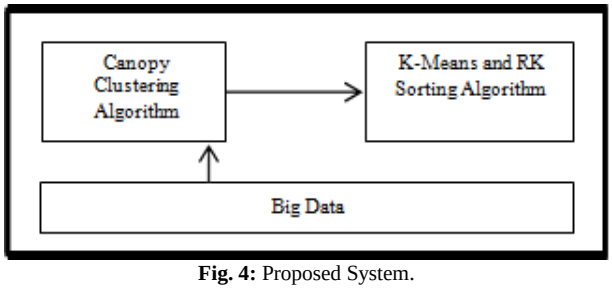


Fig. 4: Proposed System.

Data Sets:
Big data in terms of velocity, variety, volume growing exponential- ly daily.
Canopy Clustering Algorithm:
The execution time is less but accuracy is low .The results of this algorithm are a number of canopies which are the cluster centers for the given dataset.
K-Mean Clustering Algorithm:
The execution time of K-Mean clustering Algorithm given by O (nkdi) where n is the number of data points ,k is the number of clusters, i is the number of iterations needed to converge and d is the dimensions.
When the value of n and d increases then it is time consuming process or it is not applicable .In order to overcome the canopy clustering algorithm is used which is also called as pre clustering algorithm. In the Proposed system the canopy clustering the data first "coarse" or pre clustering, K value and then use K-means clustering algorithm to get the "fine" clustering

4.1. RK (Ramakrishna) sorting algorithm

RK sorting algorithm in the proposed approach is used to give ranking which does not require swapping of elements and stack space [14].
RKSORTING (A [0…N-1])
INPUT: An array A [] of n elements OUTPUT: Sorted array of n elements
Step1. Read how many elements for array i.e. n Step2. Read all n elements
Step3. for i=0 to n-1 do Step4. Key = a[i]
Step5. K=0
Step6. for j=0 to n-1 do
Step7. if(i= =j)then continue from step6 with next j value
Step8. If (key>a[j])
K++
else
If (key==a[j] && i>j) K++
end for (end of inner for loop)
Step9. B[k] =key
end for(end of outer for loop) Step10. Sorted elements are in array B. Step11. End

4.2. Results and analysis

Table 1: Computational Time and Accuracy of Canopy Algorithm
Canopy Clustering:

Number of segments	Accuracy	computational time
1000	0.68	0.24
5000	0.71	0.39
10000	0.71	0.47
20000	0.74	0.54
40000	0.79	0.82
60000	0.75	0.98
80000	0.81	1.34
100000	0.84	1.98

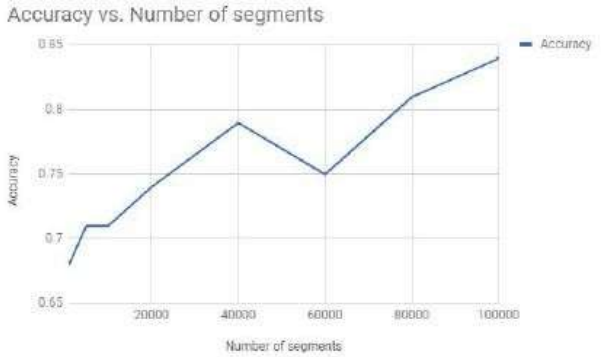


Fig. 1: Canopy Clustering – Accuracy vs. Segments.

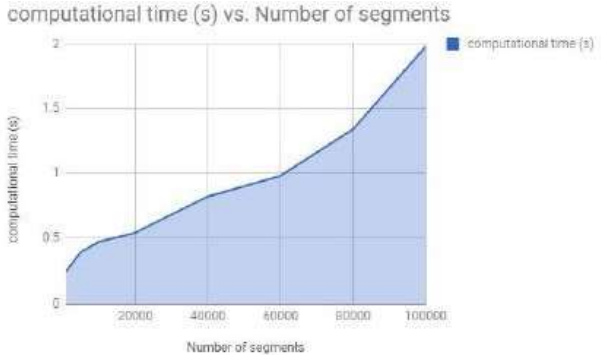


Fig. 2: Canopy Clustering – Computational Time vs. Segments.

Table 2: Computational Time and Accuracy of Kmeans Algorithm

K means Clsutering

Number of segments	Accuracy (%)	computational time (s)
1000	0.48	0.39
5000	0.54	0.52
10000	0.58	0.63
20000	0.63	0.75
40000	0.69	0.91
60000	0.71	1.21
80000	0.77	1.69
100000	0.8	2.43

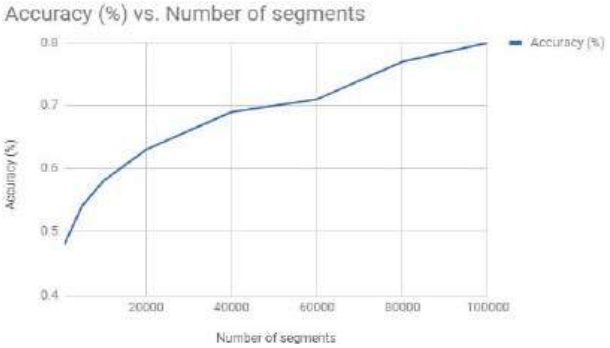


Fig. 3: Kmeans Clustering – Accuracy vs. Segments.

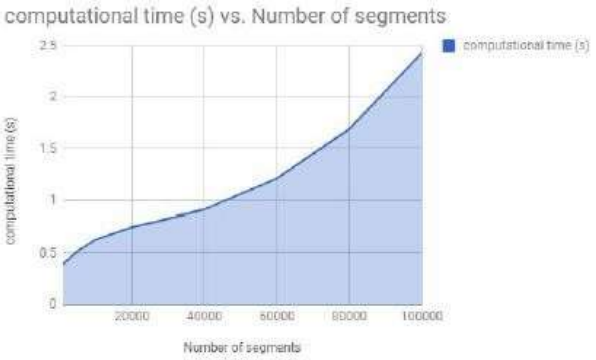


Fig. 4: Kmeans Clustering – Computational Time vs. Segments.

Table 3: Computational Time and Accuracy of Integration Canopy and K Means

Integration of canopy and K means Clustering:

Number of segments	Accuracy (%)	computational time (s)
1000	0.62	0.2
5000	0.66	0.31
10000	0.7	0.4
20000	0.73	0.61
40000	0.79	0.72
60000	0.83	0.83
80000	0.87	1.69
100000	0.89	1.81

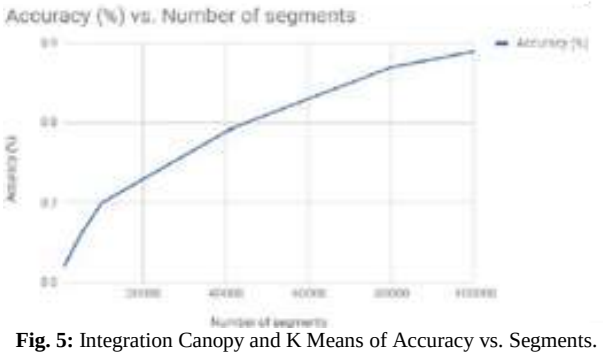


Fig. 5: Integration Canopy and K Means of Accuracy vs. Segments.

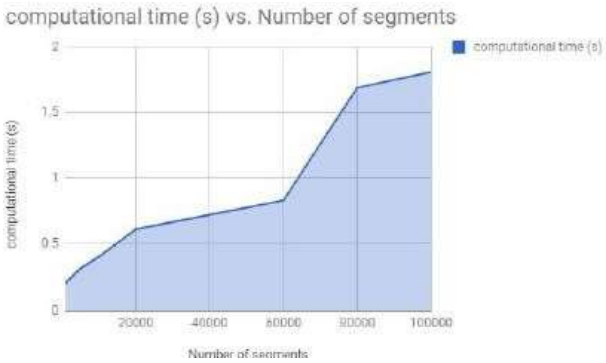


Fig. 6: Integration Canopy and Kmeans Computational vs. Segments.

Table 4: Computational Time and Accuracy of Proposed System Integration of Canopy, Kmeans, RKSorting Clustering:

Number of segments	Accuracy (%)	computational time (s)
1000	0.72	0.23
5000	0.75	0.29
10000	0.73	0.32
20000	0.79	0.37
40000	0.83	0.56
60000	0.84	0.76
80000	0.91	0.93
100000	0.93	1.11

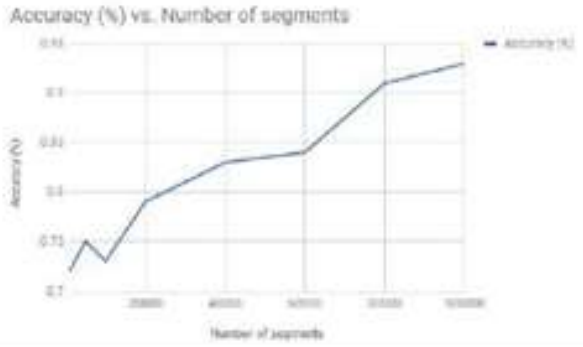


Fig. 7: Proposed System Accuracy vs. Segments.

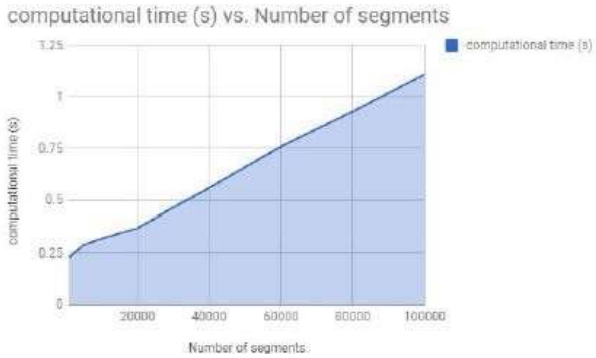


Fig. 8: Proposed System Time vs. Segments.

The data is growing in terms of volume, variety and velocity .The behavior of each clustering algorithm is analyzed through MapReduce and Hadoop platform which uses parallel processing technique. Here we considered the simulated social data of size one lakh with twelve attributes'

The figures 1,2and Table1 shows that as the dataset increases the time increases and Accuracy is not consistent for Canopy clustering algorithm ,which shows that canopy clustering algorithm alone will not be able to give better accuracy results.

In figure 3, 4 and Table2 which is Kmeans clustering algorithm shows as the data increases accuracy and computational time increases.

The figure 5, 6 and Table3 shows the Integration of canopy and kmeans clustering algorithm as the data increases accuracy increases and computational time is not consistent.

The figure 7, 8 and Table4 shows the Integration of canopy, kmeans and RK sorting algorithm or Proposed System Architecture as the data increases accuracy and computational time increases.

5. Conclusion

In this paper we have studied existing canopy, Kmeans and RK sorting algorithms for big data using MapReduce and Hadoop plat- form. And the proposed new technique ,the canopy algorithm is applied to the Big data and the output is given as the initial centers (the value of k)to K-mean clustering algorithm ,for each cluster RK sorting algorithm is used to give the rankings through MapReduce and Hadoop frame work which uses parallel processing technique.

Acknowledgement

I would like to express my special thanks of gratitude to my guide Dr.M.Vinayaka Murthy and Dr.Udaya Rani.V for their continuous support to do research paper and also my husband Saikonda Venkateswarlu who has helped with technical guidance to do the research paper.

References

[1] Ambika.s and Kavitha.G," Overcoming the Defects of K-means clustering by using Canopy Clustering Algorithm IJSRD |Vol. 4, Issue 05, 2016 | ISSN (online): 2321-0613.

[2] D. Napoleon & P. Ganga Lakshmi "An Efficient K-Means Clustering Algorithm for Reducing Time Complexity Using Uniform Distribution Data Points" IEEE, 2010, pp, 42-45.

[3] Dweepna Garg 1, Khushboo Trivedi 2, B.B.Panchal ," A Comparative study of Clustering Algorithms using MapReduce 2321-0613 in Hadoop" IJSRDt| Vol. 4, Issue 05,2016 | ISSN (online):

[4] M. S. Chen, J. Han, and P. S. Yu. IEEE Trans Knowledge and Data Engineering Data mining. An overview from a database perspective, 866-883, 1996.

[5] Ayman E. Kheer, Ahmed I. El Seddawy, Amira M. Idrees," Performance Tuning of K-Mean Clustering Algorithm a Step towards Efficient DSS", IJIRCST,ISSN: 2347-5552, Volume 2, Issue 6, November – 2014.

[6] A. Hunter and S. Parsons, "A review of uncertainty handling formalisms", Applications of Uncertainty Formalisms LNAI 1455, pp.8-37. Springer –Verlag, 1998.

[7] H.R. Shashidhar, G.T. Raju and M Vinayaka Murthy,"Efficient Estimation of Result Selectivity for Web Query Optimi zation", International Journal of Pure and Applied Mathematics, Volume 17 No. 7 2017, PP 193-205, ISSN:311-8080.

[8] H.R. Shashidhar, G.T. Raju and M Vinayaka Murthy, "Effective Cost Models for Web Query Optimization", International Journal of Pure and Applied Mathematics, Volume 117 No. 20, 17, PP 727-739, ISSN: 1311-8080.

[9] M Vinayaka Murthy "Survey On Web Query Optimization Trends and Future Research", International Conference On Advanced Material Technology 2016, Issue –V, pp 409 – 417, Elsevier Materials Today: Proceedings.

[10] M Vinayaka Murthy, "A Comparative Study on Mining the & Healthy Food Preferences of Women Clusters", Journal of Scientific Engineering Research, Vol 6, Issue 7, pp 126 131, 2017, ISSN: 2229-5518.

[11] "A Study of DM Techniques for CRM" at National Women's Science Congress, SB Arts & KCP Science College, Bijapur, on November 7 th – 9 th, 2008, Lilavathi -1, PP 65-74.

[12] "E – Governance Data analysis by Data Mining Algorithm" at the National Conference on E –Governance and its Application –alma nac '08' at Dayananda Sagar Institutions Bangalore – 78, on Nov.27 th – 28 th ,2008, PP -45.

[13] "Illustrations of k-means algorithm ",at UGC sponsored State level seminar on New Frontiers in the Development of science and technology,16 th &17th April 2009, BMS college for women - 2009.

[14] Ramakrishna,Assistant Professor Department of Computer Science, Reva Institute of Science and Management, Email: dcrrama@yahoo.co.in "New RK Sorting Algorithm".