IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-6, Issue-4 E-ISSN: 2347-2693

Examination of Clustering Techniques using Genetic Algorithm

S.Ramya [1*], N.Subha[2]

[1] Department of Computer Science, KNG Arts College (W) Autonomous, Thanjavur, India
[2] Department of Computer Science, KNG Arts College (W) Autonomous, Thanjavur, India

[*]*Corresponding Author: ramyavam@gmail.com*

Available online at: www.ijcseonline.org

Received: 14/Apr/2018, Revised: 20/Apr/2018, Accepted: 25/Apr/2018, Published: 30/Apr/2018

***Abstract*—** Bunch investigation is utilized to order comparative protests under same gathering. It is a standout amongst the most critical data mining techniques. In any case, it neglects to perform well for big data because of enormous time many-sided quality. For such situations parallelization is a superior approach. MapReduce is a prevalent programming model which empowers parallel handling in an appropriated domain. Be that as it may, a large portion of the clustering calculations are not "normally parallelizable" for example Genetic Algorithms. This is thus, because of the successive idea of Genetic Algorithms. This paper acquaints a system with parallelize GA based clustering by expanding hadoop MapReduce. An examination of proposed way to deal with assess execution picks up regarding a consecutive calculation is displayed. The investigation depends on a genuine huge data set.

***Keywords*—** Big Data, Clustering, Davies-Bouldin Index, Distributed processing, Hadoop MapReduce , Heuristics, Parallel Genetic Algorithm.

I. INTRODUCTION

Clustering is a famous method utilized for ordering data set into gatherings. Data focuses under specific gathering share comparative highlights. It is broadly utilized for design acknowledgment, data mining and so forth. Numerous systems have been formulated for group examination, for example, K-implies, fluffy c implies and so forth. However the vast majority of the traditional strategies either bargains speed of execution for clustering precision or create poor outcomes. For example, some clustering calculations stuck at nearby optima. To accomplish internationally ideal arrangement, it requires repeating over all conceivable clustering. As the quantity of cycles is exponential in data measure, for huge data sets the vast majority of such strategies would come up short. To handle this we make a move to the heuristics. Heuristics utilizes reasonable approach to acquire close ideal arrangements. Under heuristics we bargain the exactness to accomplish impressive speed ups. Rather than accomplishing an exact outcome heuristic goes for accomplishing an attractive close ideal answer for accelerate the procedure. Genetic calculation, is one such procedure. It emulates the Darwinian's foremost of "Survival of the fittest" to locate the ideal arrangement in seek space. Be that as it may, Genetic Algorithms neglect to stay aware of big-data because of colossal time intricacy. Big data is a term used to address data sets of substantial sizes. Such data sets are past the likelihood to oversee and process inside middle of the road passed time. For such a situation parallelization is a superior approach.

Hadoop MapReduce is a parallel programming method expand on the systems of Google application motor MapReduce. It is utilized for preparing expansive data in a conveyed domain. It is very adaptable and can be fabricate utilizing ware equipment. Hadoop MapReduce parts the information data into specific measured pieces and procedures these lumps all the while over the group. It in this manner lessens the time unpredictability for tackling the issue by circulating the handling among the bunch hubs. In this paper we propose a strategy to actualize clustering utilizing genetic calculation in a parallel form utilizing hadoop MapReduce. To do as such we broaden the coarse grained parallel model of genetic calculations and play out a two stage clustering on the data-set. This two stage clustering approach is acknowledged by misusing the hadoop MapReduce engineering. Whatever remains of the paper is composed as follows in area 2 we give a review of genetic calculations. Area 3 clarifies the MapReduce show and talks about the hadoop MapReduce. Segment 4 exhibits the procedure we formulated to parallelize genetic calculation based clustering by expanding hadoop MapReduce. Area 5 and 6 portray the criteria we utilized for conveying parallelized GA on hadoop MapReduce and also the consequences of experimentation.

II. GENETIC ALGORITHMS

Genetic Algorithm is a nature motivated heuristic approach utilized for taking care of hunt based and improvement issues. It has a place with a class of transformative calculations. In GAs we develop a populace of competitor arrangements towards an ideal arrangement. GA mimics nature based procedures of hybrid, change, choice and legacy to get to an ideal arrangement. Under GA we actualize the law of survival of the fittest to streamline the hopeful arrangements. The system of GA advances in the accompanying way:

Step 1:- Initial populace of competitor arrangements is made

Step 2:- Each individual from the populace is appointed a wellness esteem utilizing suitable wellness work

Step 3:- Parents are chosen by assessing the wellness

Step 4:- Offspring are made utilizing generation administrators i.e. hybrid, change and choice on guardians

Step 5:- New populace is made by choosing posterity in light of wellness assessment

Step 6:- Steps 3,4,5 are rehashed until the point that an end condition is met

III. MAPREDUCE PARADIGM

MapReduce programming worldview includes disseminated preparing of huge data over the bunch. Under this worldview the information data is spitted as indicated by the square size. The data split is performed by the info organize. These parts are doled out a particular key by the record peruser and in this way a key, esteem combine is created. Key, esteem sets are then subjected to a two stage preparing. This two stage preparing includes a map stage and a diminish stage. The engineering of an essential MapReduce worldview is portrayed in figure 1. The map stage is made out of a mapper or a map routine (). Map stage is executed in the mapper of every hub. The diminish stage is made out of a reducer or a decrease routine (). After getting the mapped comes about reducer plays out the rundown activities to produce last outcome.

Map Phase:

- The mapper gets the key-esteem sets created by the record peruser
- The mapper plays out the disseminated calculation to process the key-esteem combines and produces the mapping brings about type of middle of the road key-esteem sets
- The middle of the road key-esteem sets are then passed on to the reducer

Diminish Phase:

- The mapped consequences of the mapper are rearranged
- The rearranged comes about are at that point passed on to the fitting reducer for additionally preparing
- Combined yield of the considerable number of reducers fills in as the last outcome

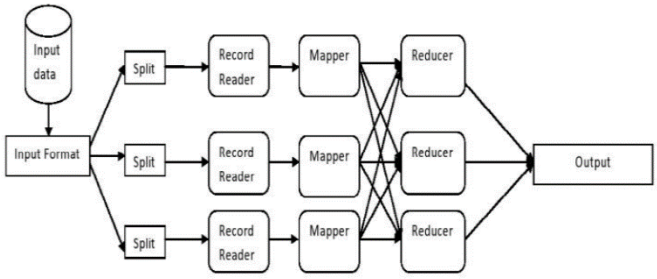

Figure 1: MapReduce paradigm

A. HADOOP MAPREDUCE

Hadoop MapReduce is a programming model which utilizes the MapReduce worldview for handling. It is enlivened by the Google application motor MapReduce. It takes into account enormous adaptability by utilizing product equipment. MapReduce utilizes HDFS (hadoop appropriated document framework) which is another part of hadoop system for putting away and recovery of data. The preparing time is lessened by part the data set into squares relying on the piece estimate. The square size is normally 64mb or 128mb. This split data is then prepared parallely finished the group hubs. MapReduce in this way gives a conveyed way to deal with unravel complex and long issues.

IV. PARALLEL GENETIC ALGORITHMS

In the accompanying segments we examine a few systems ordinarily utilized for parallelizing GA ,. At that point, we propose a tweaked way to deal with execute Clustering construct parallel GA in light of hadoop MapReduce.

i. Parallel executions

Parallel execution of GA is acknowledged utilizing two normally utilized models as:

- Coarse-grained parallel GA
- Fine-grained parallel GA

Under first model every hub is given a populace split to process. The people are then moved to other hub after map stage. Relocation is utilized to synchronize the arrangement set. In the second model every individual is given to a different hub as a rule for wellness assessment. Neighboring hubs speak with each other for choice and remaining tasks.

A. CUSTOMIZED PARALLEL IMPLEMENTATION FOR CLUSTERING USING HADOOP MAPREDUCE

In this sub area we propose the organization of GA we utilized for clustering based issues. Alongside this we talk about our altered way to deal with abuse Coarse-grained parallel GA display. This approach effectively actualizes GA construct clustering in light of hadoop MapReduce. Core of this approach lies in playing out a two staged clustering in mapper and after that, in the reducer.To start, the info data set is part as indicated by the square size by the information arrange. Each split is given to a mapper to play out the First stage clustering. The main stage mapping consequences of every mapper are passed on to a solitary reducer to play out the Second stage mapper. We in this manner, are utilizing various mappers and a solitary reducer to actualize our clustering based parallel GA. The engineering of proposed demonstrate is delineated in Figure 2.

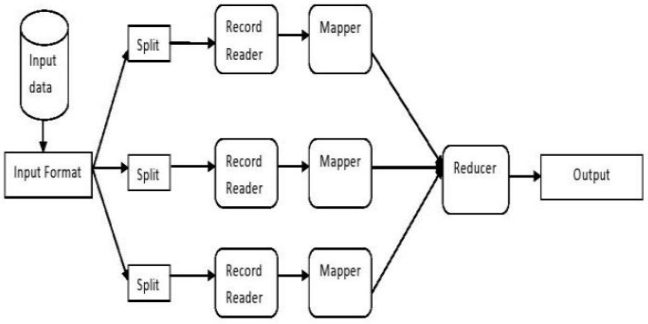

Figure 2: Parallel Implementation For Clustering Using Hadoop MapReduce

a) FIRST PHASE CLUSTERING:

- Population instatement:

Subsequent to accepting the information split every mapper frames the underlying populace of people. Every individual is a chromosome of size. Each fragment of the chromosome is a centroid. Centroids are arbitrarily chosen data focuses from the got data split. For each datum point in every chromosome clustering is performed. For this data point in the got data set doled out to the bunch of the nearest centroid.

- Fitness assessment
- For assessing wellness we are registering the Davies-Bouldin list of every person. Davies-Bouldin file is the proportion of bury bunch dissipate to the intra group partition.

The bury group diffuse of a bunch C_i is figured as

$$S_i = \frac{1}{T_i}\sum_{j=1}^{T_i} \left\| X_j - A_i \right\|_p \qquad (1)$$

Here, A_i is the centroid point, X_j is the bunch point, T_i is the group measure, p is 2 as we are computing the Euclidian separation.

The intra bunch partition of two centroids Ai and Aj is figured as

$$M_{i,j} = \left(\left(\sum_{k=1}^{n} \left| a_{k,i} - a_{k,j} \right|^p \right)^{\frac{1}{p}} \right) \qquad (2)$$

Here, k is the quantity of measurement of the data point and estimation of p is 2. Presently the Davies-Bouldin file is

$$DB = \frac{1}{N}\sum_{i=1}^{N} D_i \qquad (3)$$

Where D_i:

$$Di = \max_{j:\, i \neq j} \left\{ \frac{S_i + S_j}{M_{i,j}} \right\} \qquad (4)$$

- Mating and Selection:

For mating we are utilizing traverse and transformation methods. For traverse we are utilizing number juggling traverse with 0.7% likelihood. This produces one posterity from two guardians. The centroid of the posterity is the number juggling normal of the relating centroid of guardians. For transformation swap change is connected with 0.02%.

Under swap change we take 9's compliment of the data focuses. The posterity from more established populace are chosen to populate another populace. For choice we are utilizing Tournament choice strategy. Under competition determination the individual is chosen by playing out a competition in view of wellness assessment among a few people picked indiscriminately from the populace.

- Termination:

Another populace as created replaces the more seasoned populace. This populace would again frame a more up to date populace utilizing mating and determination strategy. This entire strategy would be rehashed and again until the point that the end condition is met. Under the proposed approach this is accomplished by finishing the predefined number of emphasess. The fittest individual of the last populace of every mapper is passed on as the outcome to the reducer. The reducer at that point performs Second stage clustering on the mapping consequences of all mapper.

b) SECOND PHASE CLUSTERING:

- Reducer frames another chromosome by joining the chromosomes got from every mapper
- This recently made chromosome is investigated. Those centroids for which intra group partition is not as much as the edge, their separate bunches are combined. For two groups the edge is figured as entirety of 20% the intra bunch partition and most extreme of the biggest separation of a bunch point from centroid among the two groups. Centroid of this recently made group is the math mean of centroids of unique bunches.

The edge calculation:

$$T = (0.2XM_{i,j}) + \max(D_i, D_j)$$

Here T is the edge, $M_{i,j}$ is the intra bunch partition of the groups C_i and C_j , D_i *and* D_j are the separation of farthestest purposes of the bunches C_i and C_j from their particular centroids

- Above expressed process is rehashed until the point that all centroids of the chromosome have a bury bunch division more prominent than limit esteem.
- The last chromosome contains area of centroid of ideal groups.

V. EVALUATION CRITERIA & RESULTS

We looked at the execution of the proposed calculation with a successive calculation. Both the calculations were executed for a sum of 500 emphases with traverse likelihood of 6% and transformation likelihood of 0.25%. The proposed calculation was executed on a multi-hub group with a sum of 5 hubs each running hadoop v1.2.1 on a Ubuntu 13.0 under vmware virtual machine with an allocated RAM of 2 GB, hard circle of 250 GB and two assigned handling centers. Equipment arrangement of the bunch is appeared in Table 1. The consecutive calculation was executed on a solitary hub with design appeared in table 2. To assess execution we quantified the exactness accomplished and add up to execution time. Execution time was estimated utilizing framework clock. The data set utilized for this analysis speaks to the differential directions of Europe map. It comprises of 169308 examples and 2 measurements.

Table 1: Hardware Configurations

Nodes	CPU	RAM	Hard Disk
Node 1	Intel core i3-370m	4GB DDR 3	640 GB
Node 2	Intel core i3-370m	4GB DDR 3	640 GB
Node 3	Intel core i7-2630qm	6GB DDR 3	640 GB
Node 4	Intel core i5-3230m	4GB DDR 3	1 TB
Node 5	Intel core i5-4200u	4GB DDR 3	1 TB

Table 2: Hardware Specifications

CPU	Intel core i3-370m
RAM	4GB DDR 3
Hard Disk	640 GB

Figure 3 and 4 shows the result on total execution time and accuracy achieved for the proposed algorithm and a sequential algorithm. The total execution time is highly reduced by using parallel genetic algorithm. A speed up of 80% was observed for PGA with a clustering accuracy of 92%. This shows that proposed algorithm considerably speeds up the clustering process for big datasets.

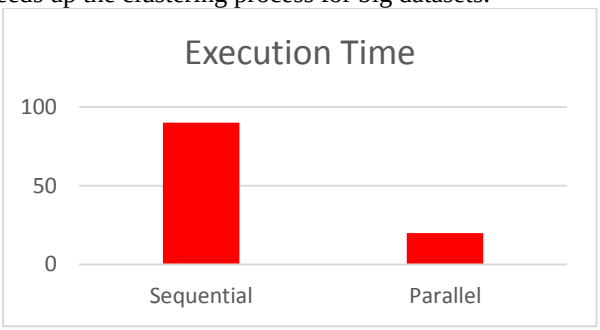

Figure 3: Execution Time

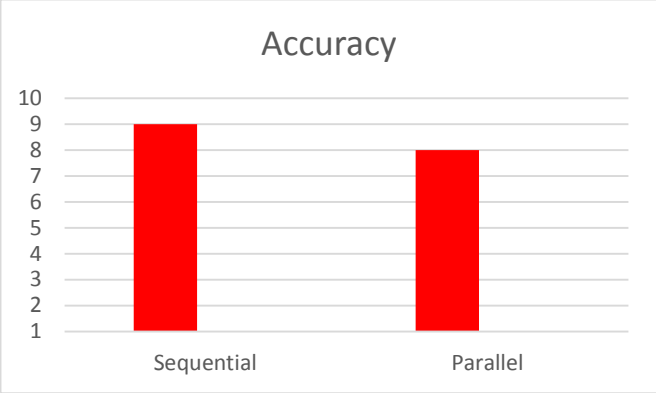

Figure 4: Accuracy

VI. CONCLUSION

This Paper Introduces a Novel Technique to Parallelize GA based clustering. For this, we have Customized Hadoop MapReduce by Implementing a Dual Phase Clustering. The accelerate in view of evaluation are introduced. In Future, we hope to improve upon the Accuracy and Enhance the Speed Gains

REFERENCES

[1] R.T.Ng, Jiawei Han, "CLARANS: a method for clustering objects for spatial data mining", IEEE Transactions on Knowledge and Data Engineering, Vol. 14, No. 5, PP. 1003 – 1016, 2002.

[2] G.Biswas, J.B.Weinberg, D.H.Fisher, "ITERATE: a conceptual clustering algorithm for data mining, IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), Vol. 28, No. 2, PP. 219 – 230, 1998.

[3] Yan Yang, Hao Wang, "Multi-view clustering: A survey", Big Data Mining and Analytics, Vol. 1, No. 2, PP. 83 – 107, 2018.

[4] Ruizhi Wu, Guangchun Luo, Qinli Yang, Junming Shao, "Learning Individual Moving Preference and Social Interaction for Location Prediction", IEEE Access, Vol. 6, PP. 10675 – 10687, 2018.

[5] K.U. Malar, D. Ragupathi, G.M. Prabhu, "The Hadoop Dispersed File system: Balancing Movability And Performance", International Journal of Computer Sciences and Engineering, Vol.2, Issue.9, pp.166-177, 2014.

[6] Qiqi Zhu, Yanfei Zhong, Siqi Wu, Liangpei Zhang, Deren Li, "Scene Classification Based on the Sparse Homogeneous–Heterogeneous Topic Feature Model", IEEE Transactions on Geoscience and Remote Sensing, Vol. 56, No. 5, PP. 2689 – 2703, 2018.

[7] Guangwei Shi, Liying Ren, Zhongchen Miao, Jian Gao, Yanzhe Che, Jidong Lu, "Discovering the Trading Pattern of Financial Market Participants: Comparison of Two Co-Clustering Methods", IEEE Access, Vol. 6, PP. 14431 – 14438, 2018.

[8] Jianzhou Wang, Fanyong Zhang, Feng Liu, Jianjun Ma, "Hybrid forecasting model-based data mining and genetic algorithm-adaptive particle swarm optimisation: a case study of wind speed time series", IET Renewable Power Generation, Vol. 10, No. 3, PP. 287 – 298, 2016.

[9] Fen Miao, Nan Fu, Yuan-Ting Zhang, Xiao-Rong Ding, Xi Hong, Qingyun He, Ye Li, "A Novel Continuous Blood Pressure Estimation Approach Based on Data Mining Techniques", IEEE Journal of Biomedical and Health Informatics, Vol. 21, No. 6, PP. 1730 – 1740, 2017.

[10] Mauro De Sanctis, Igor Bisio, Giuseppe Araniti, "Data mining algorithms for communication networks control: concepts, survey and guidelines", IEEE Network, Vol. 30, No. 1, PP. 24 – 29, 2016.

[11] Daniele Casagrande, Mario Sassano, Alessandro Astolfi, "Hamiltonian-Based Clustering: Algorithms for Static and Dynamic Clustering in Data Mining and Image Processing", IEEE Control Systems, Vol. 32, No. 4, PP. 74 – 91, 2012.

[12] Xiangyang Li, Nong Ye, "A supervised clustering and classification algorithm for mining data with mixed variables", IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, Vol. 36, No. 2, PP. 396 – 406, 2006.

[13] Yuan He, Cheng Wang, Changjun Jiang, "Mining Coherent Topics With Pre-Learned Interest Knowledge in Twitter", IEEE Access, Vol. 5, PP. 10515 – 10525, 2017.

[14] Feng Zhang, Timwah Luk, "A Data Mining Algorithm for Monitoring PCB Assembly Quality", IEEE Transactions on Electronics Packaging Manufacturing, Vol. 30, No. 4, PP. 299 – 305, 2007.

[15] Byron Graham, Raymond Bond, Michael Quinn, Maurice Mulvenna, "Using Data Mining to Predict Hospital Admissions From the Emergency Department", IEEE Access, Vol. 6, PP. 10458 – 10469, 2018.

[16] A.Bernstein, F.Provost, S.Hill, "Toward intelligent assistance for a data mining process: an ontology-based approach for cost-sensitive classification", IEEE Transactions on Knowledge and Data Engineering, Vol. 17, No. 4, PP. 503 – 518, 2005.

[17] Chun-Hao Chen, Vincent S.Tseng, Tzung-Pei Hong, "Cluster-Based Evaluation in Fuzzy-Genetic Data Mining", IEEE Transactions on Fuzzy Systems, Vol. 16, No. 1, PP. 249 – 262, 2008.

[18] Tzung-Pei Hong, Chun-Hao Chen, Yeong-Chyi Lee, Yu-Lung Wu, "Genetic-Fuzzy Data Mining With Divide-and-Conquer Strategy", IEEE Transactions on Evolutionary Computation, Vol. 12, No. 2, PP. 252 – 265, 2008.

[19] D.A.Keim, C.Panse, M.Sips, S.C.North, "Visual data mining in large geospatial point sets", IEEE Computer Graphics and Applications, Vol. 24, No. 5, PP. 36 – 44, 2004.

