IJARIIT INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume3, Issue2)

A Distributed Approach for Development of DBSCAN Clustering

Ghanshyam Dewta[1]
[1]*CSE, SSTC, CSVTU, Bhilai, Chhattisgarh, India*

Rajesh Tiwari[2]
2CSE, SSTC, CSVTU, Bhilai, Chhattisgarh, India

Abstract: Clustering plays an outstanding role in data mining applications such as scientific data exploration, information retrieval and text mining, web analysis, bioinformatics, and many others DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a density-based clustering algorithm. The key idea of the DBSCAN algorithm is that for each data point in a cluster, the neighborhood within a given radius has to contain at least a minimum number of points. We have proposed improved evenhanded workload allocation using hierarchical (Tree-Based) approach for constructing of data cluster

Keywords: Clustering, DBSCAN, Data Mining, Parallel Processing Languages.

I. INTRODUCTION

Parallel processing is used for dividing a large program into subprograms to parallel execute them for achieving a high-level degree of parallelism. Parallelism is very useful for reducing execution time and get the task done in less time. Parallel sorting is an application of parallelism in which sorting is done in many processors at the same time. Sorting program is divided in all the processors and executed separately then their resultant output merged down for final results.

There are several parallelism techniques available through which higher degree of parallelism can be achieved and also practices can be performed for exploring the ability of parallel processing and its applications. Those are Hybrid CUDA, Open MP, and MPI.

CLUSTERING

Clustering is a method of grouping similar types of data. This is very useful method applied in various applications. The K-means clustering and DBSCAN (Density-Based Spatial Clustering of Applications with Noise) clustering are the two most commonly used clustering techniques which are grouped the data together based on different criteria.

Data clustering is a popular unsupervised clustering approach for automatically finding classes, concepts, or group of patterns. Clustering involves dividing a set of objects into a specified number of clusters. The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering. There are many types of data that often occur in cluster analysis such as interval-scaled variables, binary variables, nominal, ordinal, mixed and ratio variables. Clustering is a challenging field of research in which its potential applications pose their own special requirements. Our project improves the performance of DBSCAN data mining algorithm by use of multi-core processing over large data set. Henceforth our research area is majorly parallel processing and data mining.

DATA MINING (CLUSTERING)

Clustering, or cluster analysis, group's data objects based only on information found in the data that describes the objects and their relationships. The goal is that the objects within a group be similar (or related) to one another and different (or unrelated to) from the objects in different groups. The quality of clustering is determined by distinctiveness of these groups, as well as homogeneity within a single group

OpenCL

OpenCL is an industry standard framework for programming computers composed of a combination of CPUs, GPUs, and other processors. These so-called heterogeneous systems have become an important class of platforms, and OpenCL is the first industry standard that directly addresses their needs. First released in December of 2008 with early products available in the fall of 2009, OpenCL is a relatively new technology.

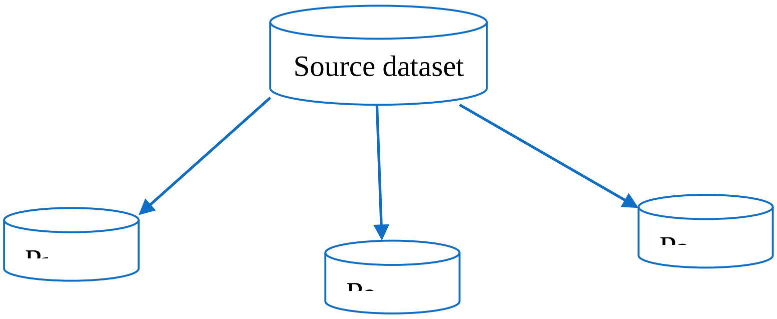

Fig. 1: Datasets division between processor

Parallel handling is broadly utilized term as a part of the field of software engineering in late decades, it is however not new in everyday life. All things considered, when there is some issue or assignment which is unrealistic for a solitary individual, we have a tendency to coordinate and complete the errand. For instance: In a marriage, there are heaps of things to deal with, such as providing food, adornment, customs, and so on. We relegate every undertaking to various people or gathering of individuals. Along these lines, we partition the entire marriage capacity in little modules and assign those errands to various gatherings they finish their individual undertaking independent of the reality of that whether the other gathering has done its undertaking or not. So also, in software engineering when we have an extensive recompense issue and in the event that we have a multi-center machine, we partition it into little modules, then distribute these modules to separate processors. Those modules executed on various processors and afterward, the general result is joined to produce the last yield.

II. RELATED WORK

This section primarily reflects the comparison and contrast of the above-reviewed literature regarding the different DBSCAN variations and modifications. It identifies the similarities and differences among the various research works on the DBSCAN algorithm enhancements.

1. The principle goal of this paper is information mining procedure is to concentrate data from an expansive information set and change it into a justifiable structure for further utilize [8]. Grouping is a primary errand of exploratory information investigation and information mining applications. Bunching is the undertaking of collection an arrangement of items in a manner that articles in the same gathering (called a bunch) are more like each other than to those in different gatherings (groups)[5]. There are distinctive sorts of bunch model: Connectivity models, Distribution models, Centroid models, Density models, Subspace model, Group models and Graph-based models. Grouping should be possible by the diverse calculations, for example, various leveled, parceling, lattice, thickness and diagram based calculations [2][1].

2. The cooperation between sensor hubs, sink hub, base station and assailants in WSNs was examined, after which a half and a half, interruption location, thickness based fluffy settler aggressive bunching algorithm (D-FICCA) theoretic recognition and grouping component was proposed [10][16]. This framework joins DBSCAN-based thickness grouping with fluffy sets components. Thusly, the proposed D-FICCA adjusts to the base station specialist to rein-power identification capacity against approaching assaults that may bring about blockage and downtime in WSN correspondence as an aftereffect of flooding parcels[3][4]. This methodology based adjusted ICA creates as an aftereffect of the ceaseless self-gaining from earlier assaults and the conduct in the fluffy learning basic leadership procedure to beat the aggressor [5][6][9].

3. The approach proposed in this paper describes the new way of intrusion detection using k- medoid clustering algorithm and certain modifications of it. The algorithm specified a new way of selection of initial medoid and proved to be better than K-means for anomaly intrusion detection [12][14]. The algorithm conveys the idea of data mining technologies which is certainly a good field and popular area of research in intrusion detection. The proposed approach is having many advantages over the existing algorithm which mainly overcomes the disadvantages of dependency on initial centroids, dependency on the number of cluster and irrelevant clusters [11][13]. The algorithm is able to sort out these problems and has been able to provide high detection rates and less false negative rate. The algorithm has many advantages but there are few disadvantages which need to be focused. The detection rate can for probe and user to root attack can be further enhanced by an efficient method of clustering which is our future work. [20][21].

4. This paper illustrates Intrusion detection is one of the major fields of research and researchers are trying to find new algorithms for detecting intrusions. Clustering techniques of data mining is an interested area of research for detecting possible intrusions and attacks, Incremental K-means and DBSCAN are two critical and well-known grouping methods, for now, 's extensive element databases (Data distribution centers, WWW et cetera) where information are changed aimlessly form[15][17]. The execution of the incremental K-implies and the incremental DBSCAN are diverse with each other in view of their time investigation qualities. Both calculations are an effective contrast with their current calculations regarding time, expense and exertion. In this paper, the execution assessment of incremental DBSCAN bunching calculation is actualized and above all it is contrasted and the execution of incremental K-implies grouping calculation and it additionally clarifies the qualities of these two calculations in view of the progressions of the information in the database. This paper likewise clarifies some sensible contrasts between these two most mainstream grouping calculations. This paper utilizes an air contamination database as a unique database on which the examination is performed [16][18][19].

III. PROPOSED METHOD

We have proposed better-balanced workload distribution using hierarchical (Tree-Based) approach for constructing of data cluster.

IV. EXPERIMENTAL STEPS

The Software Development Kit is a decent approach to finding out about DBSCAN bunching, we can perform this by utilizing taking after strides. This changed the above-composed program into the parallel code by utilizing the library capacities. These capacities are utilized to duplicate information from host to gadget and the other way around, change execution from CPU to GPU.

V. RESULTS

The present part of the proposition portrays the exploratory results. Comes about have been measures by utilizing the different emphases. Careful exploratory work with results has been plate During the trial we gauged the same size of lattice repeat much time for the yield estimation and at some point watched that the time will shift concerning network size is bigger for little grid measure CPU takes less time contrasted with GPU, all operation perform beneath appeared in unthinkable frame and correlation chart. The portrays that the K-implies bunching frames regular shapes groups and the DBSCAN grouping structure diverse states of bunches. The second correlation lies between the ideas of these two bunching calculations. If there should arise an occurrence of K-means grouping the aggregate number of bunches must be predefined however if there should be an occurrence of DBSCAN grouping the groups are framed in view of the new coming information, there is no compelling reason to predefine the quantity of bunches used in the resulting areas of this part of the theory.

Original Data	Time (in msec)
500	53750
600	52324
700	50750
800	48000
900	43540
1000	41045
1100	40250
……	……..

Table 1

Incremental Data	Time (in msec)
100	12480
200	24643
300	38943
400	52530
500	60930

Table 2

Below chart represents the time vs. original datasets incremental DBSCAN clustering.

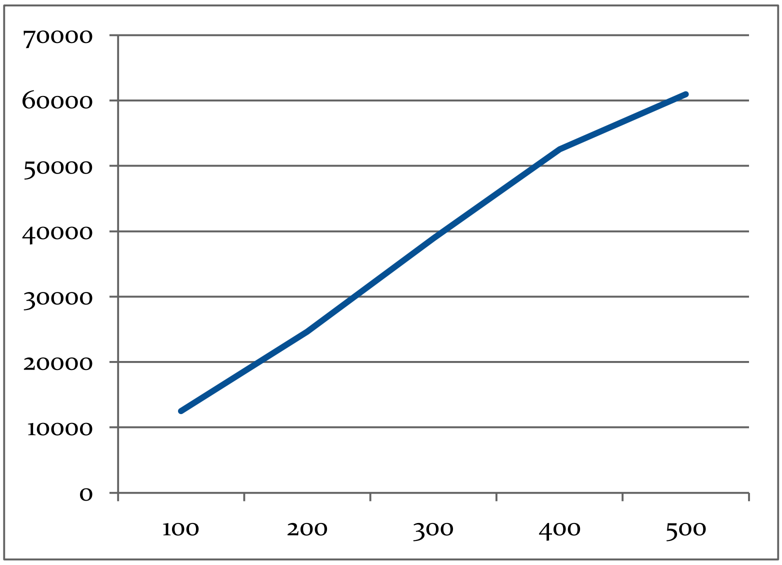

Chart 1

V. CONCLUSION AND FUTURE WORK

This paper the execution assessment of a proposed incremental DBSCAN bunching calculation is set up. This paper additionally coherently thinks about the attributes of incremental DBSCAN and incremental K-implies grouping calculations. It likewise looks at the execution of these two calculations when they are connected to constant element databases. Subsequently, the incremental K-implies grouping performs superior to the incremental DBSCAN bunching as for time investigation. In this paper, the execution assessment of a proposed incremental DBSCAN grouping calculation is built up. This paper additionally intelligently looks at the attributes of incremental DBSCAN and incremental K-implies bunching calculations. It additionally thinks about the execution of these two calculations when they are connected to continuous element databases. Therefore, the incremental K-implies bunching performs superior to the incremental DBSCAN grouping regarding time examination.

REFERENCES

[1] A. M. Fahim, A. M. Salem, F. A. Torkey, and M.A. Ramadan,"Density Clustering Based on Radius of Data (DCBRD)," World Academy of Science, Engineering, and Technology 2006.

[2] L. Kaufman and P. J. Rousseeuw, "Finding Groups in Data. An Introduction to Cluster Analysis," Wiley, 1990

[3] R. Ng and J. Han, "Efficient and Effective Clustering Method for Spatial Data Mining," Proc. Of the International Conference

[4] On Very Large Data Bases, Santiago, Chile, 1994, pp.144-155.

[5] G. Sudipto, R. Rastogi and K. Shim, "CURE: An Efficient Clustering Algorithm for Large Databases," Proc. Of the 1998 ACM SIGMOD International Conference on Management of Data, Seattle, WA, 1998, pp.73-84.

[6] G. Karypis, E. H. Hanand, V. Kumar, "Chameleon: Hierarchical Clustering using Dynamic Modelling," Computer, Aug 1999,
vol. 32, pp.68-75.

[7] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan, "Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications," Proc. of the ACMSIGMOD '98 International Conference on Management of Data, Montreal, Canada, 1998, pp.94-105.

[8] C. H. Cheng, A. W. Fu, and Y. Zhang, "Entropy-Based Subspace Clustering for Mining Numerical Data," Proc. of the 5th International Conference on Knowledge Discovery and Data Mining, San Diego, CA, 1999, pp.84-93.

[9] G. Sheikholeslami, S. Chatterjee, and A. Zhang, "Wave Cluster: A Multi-Resolution Clustering Approach for Very Large Spatial Databases," Proc. of the 24th International Conference on Very Large Databases, San Francisco, CA, 1998, pp.428-439.

[10] A. M. Fahim, A. M. Salem, F. A. Torkey, and M.A. Ramadan,"Density Clustering Based on Radius of Data (DCBRD)," World Academy of Science, Engineering, and Technology 2006.

[11] L. Kaufman and P. J. Rousseeuw, "Finding Groups in Data. An Introduction to Cluster Analysis," Wiley, 1990.

[12] R. Ng and J. Han, "Efficient and Effective Clustering Method for Spatial Data Mining," Proc. Of the International Conference on Very Large Data Bases, Santiago, Chile, 1994, pp.144-155.

[13] G. Sudipto, R. Rastogi, and K. Shim, "CURE: An Efficient Clustering Algorithm for Large Databases," Proc. Of the 1998 ACM SIGMOD International Conference on Management of Data, Seattle, WA, 1998, pp.73-84.

[14] G. Karypis, E. H. Hanand, V. Kumar, "Chameleon: Hierarchical Clustering using Dynamic Modelling," Computer, Aug 1999, vol. 32, pp.68-75.

[15] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan, "Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications," Proc. of the ACMSIGMOD '98 International Conference on Management of Data, Montreal, Canada, 1998, pp.94-105.

[16] C. H. Cheng, A. W. Fu, and Y. Zhang, "Entropy-Based Subspace Clustering for Mining Numerical Data," Proc. of the 5th International Conference on Knowledge Discovery and Data Mining, San Diego, CA, 1999, pp.84-93.

[17] G. Sheikholeslami, S. Chatterjee, and A. Zhang, "Wave Cluster: A Multi-Resolution Clustering Approach for Very Large Spatial Databases," Proc. of the 24th International Conference on Very Large Databases, San Francisco, CA, 1998, pp.428-439

[18] Margaret H. Dunham, "Data Mining: Introductory and Advanced Topics", ISBN: 0130888923, published by Pearson Education, Inc., 2003.

[19] B. Borah and D. K. Bhattacharyya, "An Improved Sampling-Based DBSCAN for Large Spatial Databases," presented in the international Conference on Intelligent Sensing and Information Processing, Chennai, India, January 2004.

[20] A. Ram, A. Sharma, A. S. Jalall, R. Singh, and A. Agrawal, "An Enhanced Density-Based Spatial Clustering of Applications with Noise," 2009 IEEE International Advance Computing Conference (IACC2009) Patiala, India, 6-7 March 2009.

[21] B. Borah and D. K. Bhattacharyya, "A Clustering Technique using Density Difference," IEEE - ICSCN 2007, MIT Campus, Anna University, Chennai, India. Feb. 22-24, 2007. pp. 585-588.

[22] X. Y. Chen, Y. F. Min, Y. Zhao, and P. Wang, "GMDBSCAN: Multi-Density DBSCAN Cluster Based on Grid," IEEE International Conference on e-Business Engineering (ICEBE 2008).

[23] P. Viswanath, and V. S. Babu, "l-DBSCAN: A Fast Hybrid Density Based Clustering Method," Proceedings of the 18th International Conference on Pattern Recognition, ICPR 2006.

[24] D. Birant and A. Kut, "ST-DBSCAN: An algorithm for clustering spatial–temporal data," Data and Knowledge Engineering, Volume 60, Issue 1(January 2007), pp. 208-221, Year of Publication: 2007, ISSN: 0169-023X.

[25] S. Mahran and K. Mahar, “Using Grid for Accelerating Density-Based Clustering,” Computer and Information Technology, CIT2008, 8th IEEE International Conference on. 08/08/2008, ISBN: 978-1-4244-2357-6, Sydney, NSW.

[26] X. P. Yu, D. Zhou, and Y. Zhou, “A New Clustering Algorithm Based on Distance and Density,” presented in proceedings of International Conference on Services Systems and Services Management (ICSSSM-2005), Vol. 2.

[27] M. T. H. Elbatta and W. M. Ashour, “A Dynamic Method for Discovering Density Varied Clusters”, Published in International Journal of Signal Processing, Image Processing and Pattern Recognition, Vol. 6, No. 1, February 2013

[28] A. Ram, S. Jalal, A. S. Jalal and M. Kumar, "DBSCAN: A Density-based Algorithm for Discovering Density Varied Clusters in Large Spatial Databases", International Journal of Computer Applications (0975–8887), vol. 3, no. 6, (2010) June.

[29] M. N. Gaonkar and K. Sawant, “Auto Eps DBSCAN: DBSCAN with Eps Automatic for Large Dataset“, Published in International Journal on Advanced Computer Theory and Engineering (IJACTE), ISSN (Print): pp. 2319 – 2526, Volume-2, Issue-2, 2013. 238

BIOGRAPHY

Ghanshyam Dewta is a Post Graduate research scholar in Department of Computer Science and Engineering, SSGI, SSTC, Bhilai (CG), India. The author obtained his bachelor degree in Computer Science and Engineering. His research interest includes Distributed DBSCAN Clustering.