Exploring Data Clustering using DBCAN

Alka Shrivastava[1] Suman Swarnkar[2]
[1,2]Bharti College of Engineering & Technology, Durg, India

Abstract— Clustering is demarcated as a technique in which data are alienated into different groups in such a way that items in every one group stake more similarity than with other items in other groups. Data clustering is a renowned technique in numerous areas of computer science and associated domains. DBSCAN is the illustrious method density centered data clustering method. Several implementation of DBSCAN has been done which made improvement in different aspects of DBSCAN and improved the performance. In this paper we will explore existing approach of numerous researchers meant for improving DBSCAN performance. In this paper we will focus on distributed approach applied for DBSCAN.

Key words: TPR, FNR, ML, NIDS

I. Introduction

Data clustering is a data mining practice that assemblies data into evocative subclasses, termed as clusters, such that it curtails the intra-differences and get the most out of inter-differences of these subclasses. Renowned algorithms for data clustering include DBSCAN, K-medoids, Kmeans, BIRCH, WaveCluster and STING. These algorithms have been castoff in numerous methodical areas for example satellite image segmentation, noise sieving and outlier detection, unsupervised text document clustering, and clustering of bioinformatics data. Earlier data clustering algorithms have been unevenly classified into four classes: partitioning-based, hierarchy based, grid-based, and density-based.

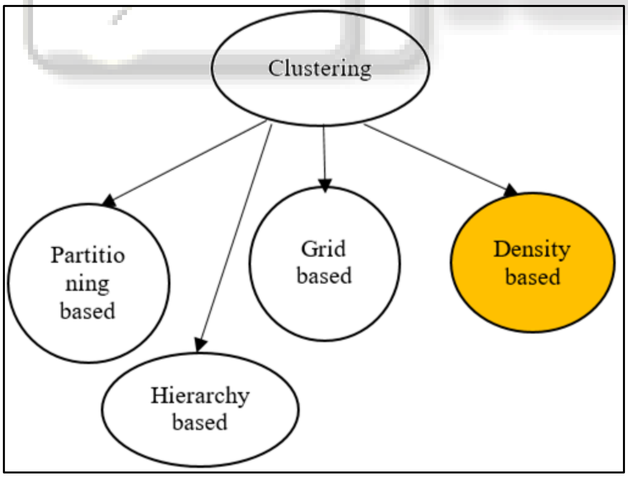

Fig.-1 Data Clustering Approach

DBSCAN (Density Based Spatial Clustering of Applications with Noise) is a density centered clustering algorithm. The concept behind DBSCAN algorithm is that for every data point in a cluster, the vicinity within a specified radius (eps) has to comprise at least a minimum number of points (minpts), i.e. the density of the neighborhood has to outstrip some threshold.

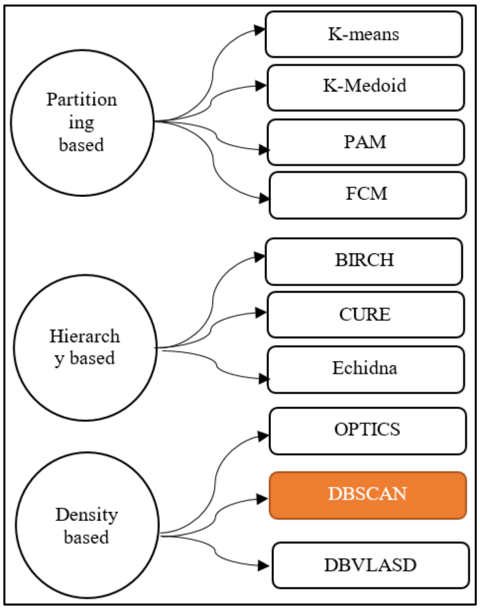

Fig. 2: Data Clustering Taxonomy

DBSCAN algorithm was projected by Martin Ester and others in 1996. DBSCAN algorithm entails only two factors: Eps & MinPts. Eps frights with a random preliminary point that has not been stayed. This point’s neighborhood is salvaged, and if it comprises adequately many points, a cluster is underway. MinPts is the minutest number of points needs to form a cluster. Eps is a prevalent parameter for DBSCAN algorithm but shrewd the value of Eps is time consuming and difficult. For calculating Eps.

```
DBSCAN (D, eps, MinPts)
C = 0
for each unvisited point p in dataset D
    mark P as visited
    NeighborPts = regionQuery (P, eps)
    If sizeof(NeighborPts) < MinPts
        mark P as NOISE
    else
        C= next cluster
        expandCluster (P, NeighborPts, C, eps, MinPts)
add P to cluster C
for each point P’ in NeighborPts
    if P’ is not visited
        mark P’ as visited
        NeighborPts’ = regionQuery (P’, eps)
        If sizeof (NeighborPts’) >= MinPts
            NeighborPts = Neighborpts joined with NeighborPts’
    if P’ is not yet member of any cluster
        add P’ to cluster C
regionQuery (P, eps)
return all points within P’s eps-neighbourhood
```

Further in this paper in section-II we have discussed different literature, in section-III we have given tabular association among different literature, in section-IV we will discuss advantage and disadvantages of different implementation of DBSCAN, at last we will conclude our survey.

II. Literature survey

I this paper we have gone through different literature some of them are briefed below:

Xiangliang Zhang et. al. said that proposed algorithm STRAP aims at clustering data streams with evolving data distributions. STRAP confronts the arriving items to the current AP model, storing the outliers in a reservoir and monitoring the ratio of outliers using the PH change point detection test. Upon triggering the PH test, the clustering model is rebuilt from the current one and the reservoir using WAP \. The key issue here was to build the change indicator, monitored by the PH test, in order to preserve the computational cost vs accuracy tradeoff. In this paper, we monitored the ratio of outliers over a sliding window and adapted the change detection threshold in real-time. The proposed approach STRAP was theoretically analyzed on guaranteeing acceptable distortion loss when exemplars slightly drift from the already selected ones, on consuming small amount of memory with little variation, and on requiring acceptable computing time that depends on the complexity of underlying distribution. The performance of STRAP in clustering quality and efficiency is empirically validated on KDD'99 benchmark problem and the URLs stream [IEEE 2013].

Jieming Shi et. al. studied for the first time the problem of Densitybased Clustering Places in Geo-Social Networks (DCPGS). Our clustering model extends the density-based clustering paradigm to consider both the spatial and social distances between places. We defined a new measure for the social distance between places, considering the social ties between users that visit them. Also said that our measure is shown to be more effective and efficient to compute, compared more complex ones based on node-to-node graph proximity. We analyzed the effectiveness of DCPGS via case studies and demonstrated that DCPGS can discover clusters with interesting properties (i.e., barrier-based splitting, spatially loose clusters, clusters with fuzzy boundaries), which cannot be found by merely using spatial clustering. Besides, we designed two evaluation measures to quantitatively evaluate the social quality of clusters detected by DCPGS or competitors, called social entropy and community score, which also confirm that DCPGS is more effective than alternative approaches.[IEEE 2017].

Kamran Khan et. al. said that Data Mining is all about data analysis techniques. It is useful for extracting hidden and interesting patterns from large datasets. Clustering techniques are important when it comes to extracting knowledge from large amount of spatial data collected from various applications including GIS, satellite images, X-ray crystallography, remote sensing and environmental assessment and planning etc. To extract useful pattern from these complex data sources several popular spatial data clustering techniques have been proposed. DBSCAN (Density Based Spatial Clustering of Applications with Noise) is a pioneer density based algorithm. It can discover clusters of any arbitrary shape and size in databases containing even noise and outliers. DBSCAN however are known to have a number of problems such as: (a) it requires user's input to specify parameter values for executing the algorithm; (b) it is prone to dilemma in deciding meaningful clusters from datasets with varying densities; (c) and it incurs certain computational complexity. Many researchers attempted to enhance the basic DBSCAN algorithm, in order to overcome these drawbacks, such as VDBSCAN, FDBSCAN, DD_DBSCAN, and IDBSCAN. In this study, we survey over different variations of DBSCAN algorithms that were proposed so far. These variations are critically evaluated and their limitations are also listed [IEEE 2014].

Chetan Dharni et. al. Analyzed that the varied density clusters and time complexity are the main problems which has been improved in many DBSCAN variations. The time complexity is reduced to O(N) with the help of indexing in DBSCAN algorithm. The problem of varied density and increasing dimensionality effect the performance of DBSCAN algorithm. Still it needs more improvement [IJCST 2013].

Fang Huang et. al. said that Density-based spatial clustering of applications with noise (DBSCAN) is a density-based clustering algorithm that has the characteristics of being able to discover clusters of any shape, effectively distinguishing noise points and naturally supporting spatial databases. DBSCAN has been widely used in the field of spatial data mining. This paper studies the parallelization design and realization of the DBSCAN algorithm based on the Spark platform, and solves the following problems that arise when computing macro data: the requirement of a great deal of calculation using the single-node algorithm; the low level of resource-utilization with the multi-node algorithm; the large time consumption; and the lack of instantaneity. The experimental results indicate that the proposed parallel algorithm design is able to achieve more stable speedup at an increased involved spatial data scale [MDPI 2017].

Jieming Shi et. al. said that Spatial clustering deals with the unsupervised grouping of places into clusters and finds important applications in urban planning and marketing. Current spatial clustering models disregard information about the people who are related to the clustered places. In this paper, we show how the density-based clustering paradigm can be extended to apply on places which are visited by users of a geo-social network. Our model considers both spatial information and the social relationships between users who visit the clustered places. After formally defining the model and the distance measure it relies on, we present efficient algorithms for its implementation, based on spatial indexing [SIGMOD 2014].

Ali Seyed Shirkhorshidi et. al. said that Clustering is an essential data mining and tool for analyzing big data. There are difficulties for applying clustering techniques to big data duo to new challenges that are raised with big data. As Big Data is referring to terabytes and petabytes of data and clustering algorithms are come with high computational costs, the question is how to cope with this problem and how to deploy clustering techniques to big data and get the results in a reasonable time. This study is aimed to review the trend and progress of clustering algorithms to cope with big data challenges from very first proposed algorithms until today's novel solutions. The algorithms and the targeted challenges for producing improved clustering algorithms are introduced and analyzed, and afterward the possible future path for more advanced algorithms is illuminated based on today's available technologies and frameworks [springer 2014].

Ilias K. Savvas et. al. implemented a parallel version of the well-known DBSCAN was presented and implemented using MPI. The results obtained from different concrete examples proved that were identical with the results produced by the application of the original sequential technique. In addition, the time complexity reduced dramatically and the experimental results shown that the algorithm scales in a very efficient manner [IEEE 2016].

III. COMPARISON

S. No.	Author/Title/Year/Publication	Method Used	Description
1.	Xiangliang Zhang et. Al./Data Stream Clustering with Affinity Propagation/ /HAL-Inria 2014	Affinity Propagation (AP) algorithm	Clustering data streams with evolving data distributions. STRAP confronts the arriving items to the current AP model.
2.	Jieming Shi et.l./Density-based Place Clustering in Geo-Social Networks/2014/SIGMOD	DBSCAN	Studied for the first time the problem of Density based Clustering Places in Geo-Social Networks (DCPGS). Clustering model extends the density-based clustering paradigm to consider both the spatial and social distances between places.
3.	Saif Ur Rehman et. Al./DBSCAN: Past, Present and Future/2016/IEEE	DBSCAN	Presented the summary information of the different enhancement of density-based clustering algorithm called the DBSCAN.
4.	SANJAY CHAKRABORTY et. Al./Analysis and Study of Incremental DBSCAN Clustering Algorithm/2014/ARXIV	DBSCAN	Propose an improve DBSCAN clustering approach which provides better and fastest result compare to the existing DBSCAN clustering algorithm up to some certain point of change in the original database.
5.	Xiangliang Zhang et. Al./ Data Stream Clustering With Affinity Propagation/IEEE 2016	STRAP	Presented STRAP algorithm combines AP with a statistical change point detection test; the clustering model is rebuilt whenever the test detects a change in the underlying data distribution. Besides the validation on two benchmark data sets, the presented algorithm is validated on a real-world application.
6.	Ali Seyed Shirkhorshidi et. Al/Big Data Clustering: A Review/Springer 2014	MapReduce, Parallel Clustering	This study is aimed to review the trend and progress of clustering algorithms to cope with big data challenges from very first proposed algorithms until today's novel solutions.
7.	Zexuan Ji et. Al./Interval-valued possibilistic fuzzy C-means clustering algorithm/Elsevier 2013	Fuzzy C-means clustering	Author compared the proposed algorithm with five fuzzy clustering approaches, including the FCM, PCM, PFCM, IFCM and IPCM, on two-dimensional Gaussian data sets and four multi-dimensional benchmark data sets. Also applied these clustering techniques to segment the brain magnetic resonance images and natural images. Results of this paper show that the proposed IPFCM algorithm is more robust to outliers and initializations and can produce more accurate clustering results.
8.	K.Kameshwaran, K.Malarvizhi/ Survey on Clustering Techniques in Data Mining/ IJCSIT 2014	-	Author said that Clustering can be done by the different algorithms such as hierarchical, partitioning, grid, density and graph based algorithms. Hierarchical clustering, which is connectivity based clustering. Partitioning clustering is the centroid based clustering. Distribution based clustering model most closely related to statistics is based on distribution models. Density-based clustering, clusters are defined as areas of higher density than the remainder of the data set.
9.	Ilias K. Savvas, and Dimitrios Tselios/Parallelizing DBSCAN Algorithm Using MPI/IEEE 2016	Parallel DBSCAN using MPI	The time complexity reduced dramatically and the experimental results shown that the algorithm scales in a very efficient manner.

Table 1:

IV. BACKGROUND

There are several implementations of DBSCAN algorithms.

El-Sonbaty et. al. has proposed a efficient density based algorithm which require preprocessing over input dataset to improve the performance of clustering, dataset partitioning used in pre-processing step. Next stage is merge stage for merging the nearest dense region which can be done by applying the DBSCAN separately on each partition in order to achieve final cluster.

A. Advantage

1) Algorithm is scalable.
2) In future parallel version of algorithm can be implemented which will improve the performance of algorithm.

B. Disadvantage

1) Before clustering preprocessing over input dataset is required.
2) I/O load may be augmented.

Adriano Moreira et. al. has proposed DBSCAN implementation which starts by identifying the k nearest neighbours of each point and identify the farthest k nearest neighbour (in terms of Euclidean distance) . The average of all this distance is then calculated. After that, for each point of the dataset the algorithm identifies the directly density-reachable points (using the Eps threshold provided by the user) and classifies the points into core or border points.

A. Advantage

1) Does not entail priori specification of number of clusters.
2) Capable to recognize noise data while clustering.

B. Disadvantage

1) Algorithm miss the mark in case of varying density clusters.
2) Performance degrades in the case of high dimension input dataset.

Duan et.al. has proposed Local Density Based Spatial Clustering Algorithm With Noise having time complexity is O(n). In the proposed algorithm, LOF (local outlier factor) is used to detect the noise and has overawed the problem of global density parameter.

A. Advantage

1) Easier for the user to prefer the suitable parameters.

B. Disadvantage

2) Cluster analysis is difficult.

Bing Liu has proposed Fast Density Based Clustering Algorithm for Large Databases in which objects are organized by certain dimensional coordinates. This algorithm uses global Eps parameter. If the value of Eps is less, then the few or single cluster containing all objects is formed and if the value of Eps is high, many trifling clusters are produced.

A. Advantage

1) Time complexity condensed.

B. Disadvantage:

1) The problem of diverse density clusters is not analyzed.

Zhang et. al has proposed A Linear DBSCAN Algorithm Based on LSH (Locality-Sensitive Hashing). The improvement of using LSH is that it decreases the time complexity and the scale of data. The time complexity of algorithm is O(NlogN).

A. Advantage

1) Comparative original DBSCAN time complexity reduces.
2) Outliers eradicated using LSH and nearest neighbor points are obtained.

B. Disadvantage

1) Preprocessing over input dataset is required.

Pathwary et. al. proposed A New Scalable Parallel DBSCAN Algorithm Using Disjoint-Set Data Structure To build clusters, a tree based bottom-up method is used. The disjoint-set data structure is cast-off to break the data access order and to achieve the merging efficiently.

A. Advantage

1) The master-slave method speeds up the process.

B. Disadvantage

1) Upsurges the I/O load and also effects the cost.

V. Conclusion

Data mining has pulled in a consideration in the data innovation and in the public eye, because of the wide accessibility of a huge amount of data and requires for extricating such data into valuable data. Clustering is the data mining technique. There are several implementation of DBSCAN which we have discussed in earlier section of paper along with advantages and disadvantages.

Existing parallel implementations of DBSCAN clustering algorithm espouse a master-slave tactic which can simply grounds an unbalanced workload and therefore result in low parallel competence.

References

[1] Xiangliang Zhang et. Al. Data Stream Clustering with Affinity Propagation HAL-Inria 2014

[2] Jieming Shi et.l. Density-based Place Clustering in Geo-Social Networks 2014 SIGMOD

[3] Saif Ur Rehman et. Al. DBSCAN: Past, Present and Future 2016 IEEE

[4] SANJAY CHAKRABORTY et. Al. Analysis and Study of Incremental DBSCAN Clustering Algorithm 2014 ARXIV

[5] Xiangliang Zhang et. Al. Data Stream Clustering With Affinity Propagation IEEE 2016

[6] Ali Seyed Shirkhorshidi et. Al Big Data Clustering: A Review Springer 2014

[7] Zexuan Ji et. Al. Interval-valued possibilistic fuzzy C-means clustering algorithm Elsevier 2013

[8] K.Kameshwaran, K.Malarvizhi Survey on Clustering Techniques in Data Mining IJCSIT 2014

[9] Ilias K. Savvas, and Dimitrios Tselios Parallelizing DBSCAN Algorithm Using MPI IEEE 2016