Survey on DBSCAN Algorithm in Distributed Data Mining using Map Reduce

Patel Chaitali[1] Akhilesh Bansiya[2]
[1,2]Department of Computer Science & Engineering
[1,2]Vedica Institute of Technology, Bhopal, India

Abstract— Data mining is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining, also called knowledge discovery in databases. Clustering is a grouping of objects into classes such as object in same cluster is similar and objects in different clusters are dissimilar. Clustering can also be used for anomaly detection. In the data mining we use DBSCAN algorithm for clustering method. DBSCAN is one the density based algorithm but most of the time lacking in the performance, run time complexity and not gives proper output in the multi density dataset. To overcome from this problem many algorithms are developed and here present literature survey on those algorithms. This paper also gives the details of different types of DBSCAN Algorithm.

Key words: Data Mining, Map-Reduce, Big Data

I. INTRODUCTION

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Data mining, also called knowledge discovery in databases, in computer science, the process of discovering interesting and useful patterns and relationships in large volumes of data. The field combines tools from statistics and artificial intelligence (such as neural networks and machine learning) with database management to analyze large digital collections, known as data sets.

The iterative process consists of the following steps:

- Data cleaning: Dirty data can cause confusion for the mining procedure, resulting in unreliable result.
- Data integration: at this stage, multiple data sources, often heterogeneous, may be combined in a common source.
- Data selection: at this step, the data relevant to the analysis is decided on and retrieved from the data collection.
- Data transformation: also known as data consolidation, it is a phase in which the selected data is transformed into forms appropriate for the mining procedure.
- Data mining: it is the crucial step in which clever techniques are applied to extract patterns potentially useful.
- Pattern evaluation: in this step, strictly interesting patterns representing knowledge are identified based on given measures
- Knowledge representation: is the final phase in which the discovered knowledge is visually represented to the user.

II. INITIATION OF A DBSCAN ALGORITHM

To find a cluster, DBSCAN starts with Database D,Core point (q), Border point (p), Minimum no of points in cluster (Minpts) and Radius(Eps).

A. Definition 1: Minimum number of neighbor points

(Eps-neighborhood of a point) The Epsneighborhood of a point p, denoted by NEps(P), is defined NEps(P) = {q E D I dist(p,q) _< Eps}.

B. Definition 2: Directly density-reachable

A point p is directly density-reachable from a point q wrt. Eps, MinPts

NEps (p) = {q€D | dist(p,q)>=Minpts}

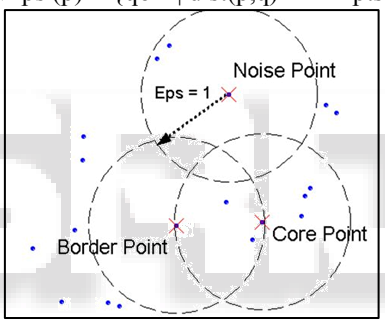

Fig. 1: core points and border points

C. Definition 3: Density-reachable

A point p is density reachable from a point q wrt. Eps and MinPts if there is a chain of points Pl ..... Pn, Pl = q, Pn = P such that Pi+l is directly density-reachable from Pi. This relation is transitive, but it is not symmetric.

Directly density reachable points are core point q and border point p.

- Core point: Minimum numbers of points are needed within Eps-neighborhood.

|NEps (q)| > = Minpts

- Border Point: Eps-neighborhood of border point has less point than the Eps of core point.

p€ NEps (q)

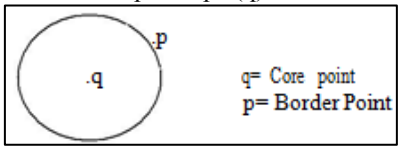

Fig. 2: CP & BP

D. Definition 4: Density-connected

A point p is density connected to a point q wrt. Eps and MinPts if there is a point o such that both, p and q are density-reachable from o wrt. Eps and MinPts. Density-connectivity is a symmetric relation.

E. Definition 5: Density reachable points.

Point p is referred as density reachable from another point q in order to Eps and Minpts. If there is a connected chain of point p_1 to p_i, p_1=q, p_i=p such as p_{n+1} is directly density reachable from p_n.

F. Definition 6: Noise

Any point that is neither core point nor border point and as well as not belongs to any of the cluster is called noise point.

```
DBSCAN (SetOfPoints, Eps, MinPts)
// SetOfPoints is UNCLASSIFIED
ClusterId := nextId(NOISE);
FOR i FROM 1 TO SetOfPoints.size DO
Point: = SetOfPoints.get (i);
IF Point.CiId = UNCLASSIFIED THEN
IF ExpandCluster (SetOfPoints, Point,
ClusterId, Eps, MinPts) THEN
ClusterId: = nextId (ClusterId)
END IF
END IF
END FOR
END;
ExpandCluster (SetOfPoints, Point,
ClusterId, Eps, MinPts)
Add p to cluster C
        For each point p' in N
        If p' is unvisited
                Mark p' visited
 N'=regionQuery (p', Eps)
 If sizeof (N')>= Minpts
        N=N' combine to N
      If p' is not in any cluster than add p' to cluster C..
```

1) Leads of DBSCAN Algorithm

- DBSCAN can remove noise from the dataset.
- DBSCAN can find arbitrary shape cluster.
- DBSCAN has a notion of noise, and is robust to outliers

2) Drawback of DBSCAN Algorithm

- Multi density dataset are not complete by DBSCAN.
- Run time complexity is high.

III. TECHNIQUE OF DBSCAN ALGORITHM

A. FDBSCAN (Fast DBSACN algorithm)

The DBSCAN clustering algorithm is one of the few clustering algorithms that allows us to find clusters within clusters, like the ones shown below.

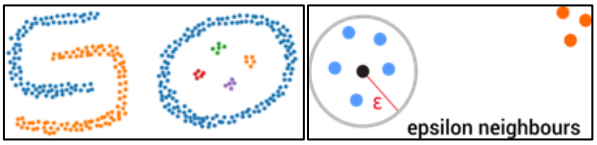

Fig. 3: Techniques

B. ODBSCAN (Optimized density based clustering algorithm)

FDBSCAN Algorithm selected representative objects as seed objects approach during the cluster expansion has been used.

C. VDBSCAN (Varied Density Based Spatial Clustering of Applications with Noise)

The basic idea of VDBSCAN is that, before adopting traditional DBSCAN algorithm, some methods are used to select several values of parameter Eps for different densities according to a k-dist plot. It calculates and stores k-dist for each project and partition the k-dist plots. The number of density is given by k-dist plot.

D. ST-DBSCAN (Spatial-Temporal Density Based Clustering)

ST-DBSCAN algorithm can cluster spatial-temporal data according to non-spatial, spatial and temporal attributes. In order to solve the conflicts in border objects it compares the average value of a cluster with new coming value.

E. Incremental DBSCAN Algorithm

Incremental DBSCAN algorithm is capable of adding points in to bulk to existing set of clusters. In this algorithm data points are added to the first cluster using DBSCAN algorithm and after that new clusters are merged with the already available clusters to come up with the modified set of clusters.

1) Invention for the Incremental DBSCAN

- New points added are clustered using DBSCAN.
- New data points which intersect with old data points are determine.
- For each intersection point from the new data set use incremental DBSCAN algorithm to determine new cluster membership.
- Clusters membership of the remaining new points then updated.

REFERENCES

[1] Derya Birant, and Alp Kut. ST-DBSCAN: An algorithm for clustering spatial-temporal data Data Knowledge. Eng. (January 2007)
[2] M.Parimala, Daphne Lopez, N.C. Senthilkumar, A Survey on Density Based Clustering Algorithms for Mining Large Spatial Databases.
[3] Wei Wang, Shuang Zhou, Bingfei Ren, Suoju He" Improved VDBSCAN with Global Optimum K" ISBN: 978-0-9891305-0-9 ©2013 SDIWC

