Performance evaluation of clustering algorithms on trajectories data

Sweta Kumari[1] , Mrs. Varsha Singh[2]
Electrical Department , NIT Raipur
[1]M.Tech.(C.T.),[2]Asst.Prof.(Electrical Department)
NIT Raipur, Chhattisgarh, INDIA

***Abstract*: Clustering is a data mining technique used to place data elements into related groups without advance knowledge of the group definitions. In this paper, trajectories data are use to evaluate the performance of clustering algorithms on the factor of time parameter. We propose the time- based clustering algorithm that adapts the agglomerative and DBSCAN clustering algorithms for trajectory data. We present experimental results that show the performance and accuracy of clustering algorithms.**

***Keywords*: clustering, DBSCAN, agglomerative hierchical, trajectories database**

I. INTRODUCTION

Data mining can be performed on data represented in quantitative, textual, or multimedia forms. Data mining applications can use different types of parameters to examine the data. They include association, sequence or path analysis, classification, clustering and forecasting. Data mining has become increasingly common in both the public and private sectors [1]. In this, paper we are discussing performance of clustering algorithms. Cluster analysis has been widely used in numerous applications, including market research, pattern recognition, data analysis, and image processing. In business, clustering can help marketers discover distinct groups in their customer bases and characterize customer groups based on purchasing patterns. Clustering is also called data segmentation in some applications because clustering partitions large data sets into groups according to their similarity [2] . Performance of clustering algorithms is evaluated among k-mean, hierarchical, SOM and expectation maximization on many factors such as size of data set, type of data set and type of software [5]. Clustering moving object trajectory data is thus an appealing research direction to fulfil the needs of many applications. In general, clustering is defined as the division of data into groups of similar objects. K-mean clustering algorithm adapts time-based parameter for trajectory data [7]. In this paper, evaluating the performance of agglomerative and DBSCAN clustering algorithms on the basis of time-based parameter for trajectory data. Agglomerative hierarchical clustering also called bottom-up approach starts with each object forming a separate group. It successively merges the objects or groups that are close to one another, until all of the groups are merged into one or until all a termination condition holds. DBSCAN clustering method have been developed based on the notion of density because most of the partitioning methods cluster objects based on the distance between objects their general idea is to continue growing the given cluster as long as the density in the “neighborhood” exceeds some threshold.

II. CLUSTERING ALGORITHMS

Clustering can be considered the most important unsupervised learning problem; so, as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data. A cluster is therefore a collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters. So, the goal of clustering is to determine the intrinsic grouping in a set of unlabeled data. All clustering algorithms will produce clusters, regardless of whether the data contains them as shown in fig 1.Clustering is widely used in many applications including pattern recognition, dense region identification, customer purchase pattern analysis, web pages grouping, information retrieval, and scientific and engineering analysis.

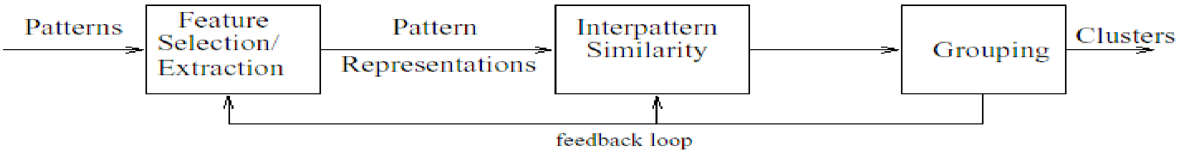

Fig 1: Stages in clustering

To study the performance of the clustering algorithms with moving object date sets, we have chosen agglomerative hierarchical, DBSCAN and it is discussed below.

A) Agglomerative hierarchical clustering

A hierarchical clustering method works by grouping data objects into a tree of clusters. Hierarchical clustering methods can be classified as either agglomerative or divisive, depending on whether the hierarchical decomposition is formed in a bottom-up (merging) or top-down(splitting) fashion [3-4]. Agglomerative clustering treats each data point as a singleton cluster, and then successively merges clusters until all points have been merged into a single remaining cluster as shown in fig2 .

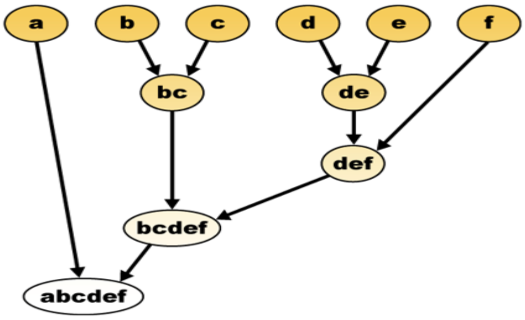

Fig 2**:** Agglomerative hierarchical clustering

Steps of agglomerative hierarchical algorithm

1. Compute the proximity matrix containing the distance between each pair of patterns. Treat each pattern as a cluster.

2. Find the most similar pair of clusters using the proximity matrix. Merge these two clusters into one cluster. Update the proximity matrix to reflect this merge operation.

3. If all patterns are in one cluster, stop. Otherwise, go to step2.

The advantages of the hierarchical clustering algorithms are the reason this algorithm was chosen for discussion. These advantages include:

- Embedded flexibility regarding a level of granularity.
- Ease of handling of any forms of similarity or distance.
- Consequently applicability to any attributes types.
- Hierarchical clustering algorithms are more versatile.

B) DBSCAN clustering

Many clustering algorithms determine clusters based on Euclidean or Manhattan distance measures. Those algorithms based on this distance measures gives clusters in spherical shape with similar size and density. DBSCAN clustering is methods which forms clusters of arbitrary shape. It is a density-based clustering algorithm because it finds a number of clusters starting from the estimated density distribution of corresponding nodes. DBSCAN is one of the most common clustering algorithms and also most cited in scientific literature. This algorithm is particularly suited to deal with large datasets, with noise, and is able to identify clusters with different sizes and shapes [6].

. DBSCAN requires two parameters

- Eps: Maximum radius of the neighborhood
- MinPts: Minimum number of points in an Eps neighborhood

.

The clustering process is based on the classification of the points in the dataset as core points, border points and noise points and on the use of density relations between points directly density reachable, density reachable, density connected[Ester 1996] to form the clusters.

Core points: The points that are at the interior of a cluster are called core points. A point is an interior point if there are enough points in its neighborhood.

Border points: Points on the border of a cluster are called border points. NEps(p): {q belongs to D | dist(p,q) <= Eps}

Noise points: A noise point is any point that not a core point or a border point.

Directly Density-Reachable: A point p is directly density-reachable from a point q with respect to Eps, MinPts if p belongs to NEps(q) |NEps (q)| >=MinPts

Density-Reachable: A point p is density-reachable from a point q with respect toEps, MinPts if there is a chain of points p1, …, pn, p1 = q, pn = p such that pi+1 is directly density-reachable from pi

Density-Connected: A point p is density-connected to a point q with respect to Eps, MinPts if there is a point o such that both, p and q are density reachable from o with respect to Eps and MinPts.

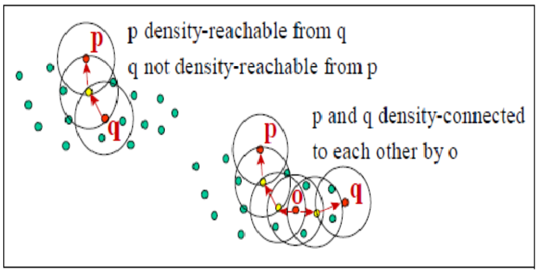
p
q
p density-reachable from q
q not density-reachable from p
p
o
q
p and q density-connected
to each other by o

Fig 3: density-reachable and density-connected

Algorithm: The algorithm of DBSCAN is as follows (M. Ester, H. P. Kriegel, J. Sander, 1996)

- Arbitrary select a point p
- Retrieve all points density-reachable from p with respect to Eps and MinPts.
- If p is a core point, a cluster is formed.
- If p is a border point, no points are density-reachable from p and DBSCAN visits the next point of the database.
- Continue the process until all of the points have been processed.

III. EXPERMENTAL RESULT

We have implemented the two clustering algorithms agglomerative hierarchical and DBSCAN clustering algorithm in dot net and performed the experiments on a normal desktop computer. We have kept some parameters of the simulation as constant and vary few parameters and measured the performance as shown in table 1, the number of clusters was changed in each case and in each case number of noise was measured.

TABLE 1: SUMMARY OF RESULT

Clustering Algorithm	Agglomerative Hierarchical		DBSCAN	
No of Records	Cluster	Noise	Cluster	Noise
150	13	75	31	4
350	25	90	62	6
750	41	200	95	10
1000	58	412	214	12

The performance graph is measured between agglomerative hierarchical and DBSCAN clustering algorithm in terms of number of cluster is formed with respect to time as shown in fig 4. .

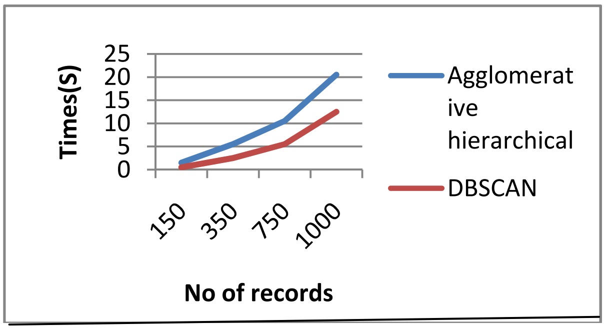

Fig 4: Graph for agglomerative vs. DBSCAN

IV. CONCLUSION

In this paper performance evaluation of a two clustering algorithm is established. Clustering algorithms are attractive for the task of class identification in spatial databases. In this work, focus has been made over the comparison of clustering algorithms i.e. DBSCAN and Agglomerative hierarchical. DBSCAN relies on a density-based notion of clusters. It requires only one input parameter and supports the user in determining an appropriate value for it.In this paper the performance evaluation of Agglomerative hierarchical and DBSCAN clustering algorithm is established. This performance measure and compare the performance with the existing Agglomerative hierarchical and DBSCAN clustering is also presented in this paper clearly. The proposed technique is implemented using open source technology dot net frame and dataset is selected for the experiment.

REFERENCES

1. CRS Report RL31798 "Data Mining: An Overview" by Jeffrey W. Seifert. Congressional Research Service December 16, 2004
2. Naresh kumar Nagwani and Ashok Bhansali, "An Object Oriented Email Clustering
   Model Using Weighted Similarities between Emails Attributes", International Journal of Research and Reviews in Computer science (IJRRCS), Vol. 1, No. 2, June 2010.
3. Reynaldo J. Gil-Garc´_a1, Jos´e M. Bad´_a-Contelles2 and Aurora Pons-Porrata "A General Framework for Agglomerative Hierarchical Clustering Algorithms" June 2006.
4. Parul Agarwal, M. Afshar Alam, Ranjit Biswas" Analysing the agglomerative hierarchical Clustering Algorithm for Categorical Attributes" International Journal of Innovation, Management and Technology, June 2010
5. Osama Abu Abbas "Comparisons Between Data Clustering Algorithms" The International Arab Journal Of Information Technology , July 2008.
6. Angeline Christobel . Y ,Dr. Sivaprakasam "A Study on the Performance of Classical Clustering Algorithms with Uncertain Moving Object Data Sets" *International Journal of Computer Science and Information Security,Vol. 9, No. 4,April, 2011*
7. Hoda M. O. Mokhtar1 Omnia Ossama2 Mohamed El-Sharkawi3 "A Time Parameterized Technique for Clustering Moving Object Trajectories International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.1, No.1, January 2011