Heterogeneous Density Based Spatial Clustering of Application with Noise

J. Hencil Peter[†] **and A.Antonysamy** [††],

[†]Research Scholar St. Xavier's College, Palayamkottai Tamil Nadu, India
[††]Principal St. Xavier's College, Kathmandu, Nepal

Summary
The DBSCAN [1] algorithm is a popular algorithm in Data Mining field as it has the ability to mine the noiseless arbitrary shape Clusters in an elegant way. As the original DBSCAN algorithm expand the Cluster based on the core object condition, it doesn't have the intelligence to mine the clusters which have different densities and these clusters may or may not be separated by the sparse region. In this paper we propose a new algorithm for mining the density based clusters and the algorithm is intelligent enough to mine the clusters with different densities. For every new cluster expansion, homogeneity core object's density range (start and end value) will be obtained using a function and based on the range values, cluster(s) will be allowed to expand further. To improve the performance of the new algorithm and without loosing the quality of Clusters, we have used the Memory Effect in DBSCAN Algorithm [7] approach. The new algorithm's output and performance analysis shows that proposed solution is superior to the existing algorithms.

Keywords:
Density Different Cluster(s), Variance Density DBSCAN, Heterogeneous Density Clusters.

1. Introduction

Data mining is a fast growing field in which clustering plays a very important role. Clustering is the process of grouping a set of physical or abstract objects into classes of similar objects [2]. Among the many algorithms proposed in the clustering field, DBSCAN is one of the most popular algorithms due to its high quality of noiseless output clusters. Mining the clusters with similar and different densities are the two subcategories of density based clustering. For dealing with different densities in the same data base, few algorithms are already proposed and most of the existing algorithms take very sensitive parameters as input to determine the density variance. Hence, if we make a minor change in the input parameter(s), algorithm will give the unpredicted output. In this paper, a simple and user-friendly parameter acceptable DBSCAN algorithm has been introduced to deal with different density clusters in the same data set.

Manuscript received August 5, 2010
Manuscript revised August 20, 2010

Rest of the paper is organised as follows. Section 2 gives the brief history about the related works in the same area. Section 3 gives the introduction of original DBSCAN and section 4 explains the proposed algorithm. After the new algorithm's explanation, section 5 shows the Experimental Results and final section 6 presents the conclusion and future work associated with this algorithm.

2. Related Works

The DBSCAN (Density Based Spatial Clustering of Application with Noise) [1] is the basic clustering algorithm to mine the clusters based on objects density. In this algorithm, first the number of objects present within the neighbour region (Eps) is computed. If the neighbour objects count is below the given threshold value, the object will be marked as NOISE. Otherwise the new cluster will be formed from the core object by finding the group of density connected objects that are maximal w.r.t density-reachability. The cluster formed by the DBSCAN algorithm will have wide variation inside each cluster in terms of density.

The OPTICS [4] algorithm adopts the original DBSCAN algorithm to deal with variance density clusters. This algorithm computes an ordering of the objects based on the reachability distance for representing the intrinsic hierarchical clustering structure. The Valleys in the plot indicate the clusters. But the input parameters ξ is critical for identifying the valleys as ξ clusters.

The DENCLUE [5] algorithm uses kernel density estimation. The result of density function gives the local density maxima value and this local density value is used to form the clusters. If the local density value is very small, the objects of clusters will be discarded as NOISE.

The CHAMELEON [6] is a two phase algorithm. It generates a k-nearest graph in the first phase and hierarchical cluster algorithm has been used in the second phase to find the cluster by combining the sub clusters.

The DDSC (A Density Differentiated Spatial Clustering Technique) [3] and EDBSCAN (An Enhanced Density Based Spatial Clustering of Application with Noise) [8] are the extension of DBSCAN algorithm, gives solution to

handling different densities. The DDSC algorithm takes very sensitive parameter for variance density clusters and even a very minimum change in the parameter will give wrong result. The other algorithm EDBSCAN expands the cluster based on the Relative Core Object condition. Homogeneity Index (HI) and Density Variance are the two important parameters which determine the density variance.

The most of the Density Based algorithms accepts very sensitive parameters for working on different density clusters. Even if we give the right density parameter values, it will not be able to deal with different range of densities and this may vary based on the nature of data base. So this paper introduces a function to handle the density variance. The function fHR() takes the core object and MinObjs as input, and gives the given object's start and end density range values. The cluster expansion will happen based on the new density range values.

3. Introduction to DBSCAN Algorithm

The working principles of the DBSCAN algorithm are based on the following definitions:

Definition 1: Eps Neighbourhood of an object p

The Eps Neighbourhood of an object p is referred as NEps(p), defined as

$NEps(p) = \{q \in D \mid dist(p,q) <= Eps\}$.

Definition 2: Core Object Condition

An Object p is referred as core object, if the neighbour objects count >= given threshold value (MinObjs). i.e. $|NEps(p)| >= MinObjs$

Definition 3: Directly Density Reachable Object

An Object p is referred as directly density reachable from another object q w.r.t Eps and MinObjs if

$p \in NEps(q)$ and

$|NEps(q)| >= MinObjs$ (Core Object condition)

Definition 4: Density Reachable Object

An object p is referred as density reachable from another object q w.r.t Eps and MinObjs if there is a chain of objects p1,…,pn, p1=q, pn=p such that pi+1 is directly density reachable from pi.

Definition 5: Density connected object

An Object p is density connected to another object q if there is an object o such that both, p and q are density reachable from o w.r.t Eps and MinObjs.

Definition 6: Cluster

A Cluster C is a non-empty subset of a Database D w.r.t Eps and MinObjs which satisfying the following conditions.

For every p and q, if $p \in$ cluster C and q is density reachable from p w.r.t Eps and MinObjs then $q \in C$.

For every p and q, $q \in C$; p is density connected to q w.r.t Eps and MinObjs.

Definition 7: Noise

An object which doesn't belong to any cluster is called noise.

The DBSCAN algorithm finds the Eps Neighbourhood of each object in a Database during the clustering process. Before the cluster expansion, if the algorithm finds any non core object, it will be marked as NOISE. With a core object, algorithm initiate a cluster and surrounding objects will be added into the queue for the further expansion. Each queue objects will be popped out and find the Eps neighbour objects for the popped out object. When the new object is a core object, all its neighbour objects will be assigned with the current cluster id and its unprocessed neighbour objects will be pushed into queue for further processing. This process will be repeated until there is no object in the queue for the further processing.

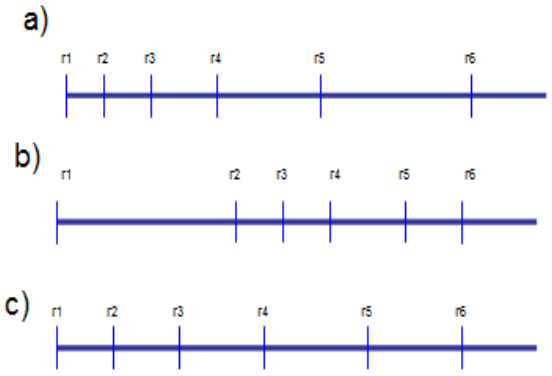
4. Proposed Solution

A new algorithm has been proposed in this paper to overcome the problem of dealing with different density clusters. To maintain the density variance between the clusters, following definition has been proposed.

Definition 1: Homogenity core object density range

Homogeneity core objects density range is defined as the range of core objects density value which has the start and end values. i.e. if an object o is a core object, the bounding start and end values are the core object density range.

To find the homogeneity core object density range, function fHR(o, MinObjs) has been introduced and this function is highly configurable that takes one core object and MinObjs as argument. In the existing algorithms, homogeneity density variance has been calculated using the user specified parameters and it can't be customised to give the different density range depends on the nature of data base. Following diagram shows three different possible density ranges.



The different ranges of densities (r1=MinObjs)

In the above diagram, minimum density will be the MinObjs parameter value which is always equal to r1 as the algorithm should not loose the originality. Then the higher densities can be splitted into different range based on the nature of the database. If we need different range of density values, the function needs to be customised with the appropriate formula.

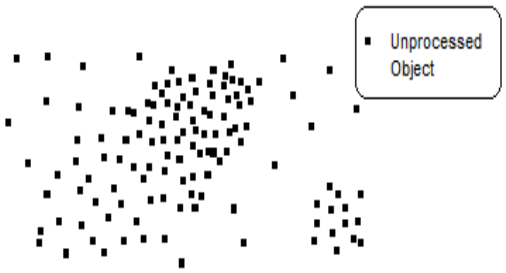
To improve the performance of the algorithm, Memory Effect in DBSCAN Algorithm [7] approach has been applied. So there are two types of Regionquery functions have been introduced in this algorithm namely, LongRegionQuery and ShortRegionQuery. First LongRegionQuery function will be called to get the region objects present in Eps neighbours as well as 2*Eps neighbours surrounded by the given object. Later all the unprocessed objects present in the Eps neighbour region will be processed using the ShortRegionQuery function call.

Proposed Algorithm (HDBSCAN – Heterogeneous DBSCAN)

1) Input D, Eps, MinObjs
2) Initialize all objects Cluster ID field as UNCLASSIFIED.
3) For each UNCLASSIFIED object o ∈ D
4) Call LongRegionQuery function with D, Eps and o parameters to be obtain InnerRegionObjects and OuterRegionObjects.
5) Set the object o Density to InnerRegionObjects.Count
6) IF o is a core object THEN
7) Call the function fHR(o, MinObjs) to Get the Homogenity Core Density Range for the Current core object o.
8) Get the ClusterID for the new Cluster.
9) Assign the new Cluster ID to all the objects exist in the InnerRegionObjects if the Object.ClusterID ∈{UNCLASSIFIED, NOISE} OR (Object.Density >= HomogenityCoreDensityStart AND Object.Density <= HomogenityCoreDensityEnd).
10) For each object T ∈ InnerRegionObjects where T.Density == NOTSET
11) Call ShortRegionQuery function with InnerRegionObjects, OuterRegionObjects, Eps and Object T to obtain the ShortRegionobjects.
Set T.Density=ShortRegionObjects.Count
13) If T is Homogenity Core Density Object
14) Push every objects SRO ∈ ShorRegionObjects to SeedQueue if the object SRO is not previously added to the queue AND SRO.ClusterID == UNCLASSIFIED
15) Assign ClusterID to all the objects exist in ShorRegionObjects if the Object.ClusterID ∈{UNCLASSIFIED, NOISE} OR Object is a Homogenity Core Density Object.
16) End If
17) End For
18) Process steps 19-22 until Seed Queue is Empty.
19) Get the next Object n from the SeedQueue.
20) If the n.Density == NOTSET
21) Call LongRegionQuery function with D, Eps and n Parameters to obtain InnerRegionObjects and OuterregionObjects.
22) Repeat the steps 9-18
23) End If
24) Else
25) Mark o as NOISE
26) End If
27) End for

First the algorithm start with LongRegionQuery function call to obtain the Neighbour objects (InnerRegionobjects and OuterRegionObjects) and the InnerRegionObjects count will be assign to the current object's Density field. In this algorithm LongRegionQuery and ShortRegionQuery has been used to improve the speed. The ShortRegionQuery takes the return array objects of LongRegionQuery function and will not process the whole Data set in the subsequent iteration. Thus the performance improvement has been guaranteed when the Eps value is reasonably insensitive. Once the LongRegionQuery function call gets executed, all the InnerRegionObjects whose distance <= Eps will be processed in the subsequent iterations using the ShortRegionQuery function. When the initial core object is obtained, new cluster id gets generated and all the InnerRegion Objects which are satisfies the condition (NOISE OR UNCLASSIFIED or Homogenity core object density range) get assigned with the new Cluster ID. Then the Homogenity density range will be obtained using the function fHR(o, MinObjs).

Hereafter only the Homogenity core objects whose density bound between the Homogeneity Core Object start and end range will be expanded further for the present cluster. While processing the entire object exists in the InnerRegionobject array, new unprocessed objects will be pushed into SeedQueue for the further processing and once the InnerRegionObjects are processed, next object will be popped out from SeedQueue and LongRegionQuery function call will be applied. This process will be continued until the SeedQueue become empty.



Unprocessed Dataset [Fig.1]

Processed Dataset [Fig.2] where Eps =15 and MinObjs = 3

5. Performance Evaluation:

The proposed algorithm has been implemented using Visual C++ (2008) on Windows Vista OS and tested using two dimensional Dataset. To know the real performance difference achieved in the new algorithm, we haven't used any additional data structures (like spatial tree) to improve the performance. As the ME approach has been used in the HDBSCAN algorithm, it is proved that the new density variance algorithm's performance is superior to the existing.

Total Objects	HDBSCAN (without using ME approach)	HDBSCAN (using ME approach)
150	0.001	0.001
320	0.018	0.027
535	0.031	0.069
735	0.057	0.092
1150	0.084	0.219
2550	0.162	0.503

Performance Difference Table (Run time is given in Seconds)

Above table shows the essential of using the ME approach in this algorithm. Though the sensitive Eps Value affect the performance even if we using ME approach, most of the cases we can achieve the better performance using ME approach.

6. Conclusion

In this paper, we have proposed a simple and fast DBSCAN algorithm to work with different density mixed cluster objects within the same dataset. The new algorithm accepts, same numbers of parameters as like original DBSCAN algorithm require. The new algorithms homogeneity core objects density range has been measured using the configurable function to support different range and Memory Effect approach has been used to improve the speed of the algorithm.

References

[1] Ester M., Kriegel H.-P., Sander J., and Xu X. (1996) "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise" In Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD'96), Portland: Oregon, pp. 226-231.

[2] J. Han and M. Kamber, Data Mining Concepts and Techniques. Morgan Kaufman, 2006.

[3] B. Borah, D.K. Bhattacharyya,: "DDSC, "A Density Differentiated Spatial Clustering Technique", Journal Of Computers, Vol. 3, No. 2, February 2008.

[4] M. Ankerst, M. Breunig, H. P. Kriegel, and J. Sander, "OPTICS: Ordering Objects to Identify the Clustering Structure, Proc. ACM SIGMOD," in International Conference on Management of Data, 1999, pp. 49–60.

[5] A. Hinneburg and D. Keim, "An efficient approach to clustering in large multimedia data sets with noise," in 4th International Conference on Knowledge Discovery and Data Mining, 1998, pp. 58–65.

[6] G. Karypis, E. H. Han, and V. Kumar, "CHAMELEON: A hierarchical clustering algorithm using dynamic modeling," Computer, vol. 32, no. 8, pp. 68–75, 1999.

[7] Li Jian; Yu Wei; Yan Bao-Ping; , "Memory effect in DBSCAN algorithm," Computer Science & Education, 2009. ICCSE '09. 4th International Conference on , vol., no., pp.31-36, 25-28 July 2009.

[8] Ram, A.; Sharma, A.; Jalal, A.S.; Agrawal, A.; Singh, R.; , "An Enhanced Density Based Spatial Clustering of Applications with Noise," Advance Computing Conference, 2009. IACC 2009. IEEE International , vol., no., pp.1475-1478, 6-7 March 2009.

J. Hencil Peter is Research Scholar, St. Xavier's College (Autonomous), Palayamkottai, Tirunelveli, India. He earned his MCA (Master of Computer Applications) degree from Manonmaniam Sundaranar University, Tirunelveli. Now he is doing Ph.D in Computer Applications and Mathematics (Interdisciplinary) at Manonmaniam Sundranar University, Tirunelveli. His interested research area is algorithms inventions in data mining.

Dr.A. Antonysamy is Principal of St. Xavier's College, Kathmandu, Nepal. He completed his Ph.D in Mathematics for the research on "An algorithmic study of some classes of intersection graphs". He has guided and guiding many research students in Computer Science and Mathematics. He has published many research papers in national and international journals. He has organized Seminars and Conferences in state and national level.