Volume 3, Issue 3, March 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Local Density Differ Spatial Clustering in Data Mining

Richa Sharma
M.Tech. Student
Computer Science
Galgotia Inst. of Engg. & Tech.
India

Bhawna Malik
Professor & Head
Computer Science
Galgotia Inst. of Engg. & Tech.
India

Anant Ram
Associate Professor & Head
Computer Science
GLA, Mathura,
India

Abstract: Clustering in data mining is a discovery process that groups a set of data objects so that the inter-cluster similarity is minimized and intra- cluster similarity is maximized. In presence of noise and outlier in high dimensional data base it is a difficult task to find out the clusters of different shapes, sizes and differ in density. Density based clustering algorithms like DBSCAN finds the clusters based on density property but still within the same cluster the major density difference may exist due to the only minimum point value. In this paper we propose a Local density differ clustering algorithm which capable to handle the local density variation within the cluster. It calculates the density variance in its surrounding and if any core object that have the density variance less than a given threshold value K than that core object can start the formation of cluster. The proposed clustering algorithm generates more density based homogenous cluster in comparison to DBSCAN.

Keywords: Core objects, minimum points, density variance, radius, threshold value K.

I. INTRODUCTION

Clustering is a technique through which similar data objects are collected into one group named as cluster. Clustering is performed via various density based algorithms like DBSCAN, DENCLUE etc. These density based algorithms are effective in finding homogeneous and arbitrary shape clusters of any size. But these algorithms fail to identify the clusters with varied densities within a cluster. Like for example DBSACN algorithm which is based on density based algorithms manage to find a effective homogeneous single cluster but it fails to detect distinct clusters of varied density within a cluster.

So a **local density differ DBSCAN algorithm** is proposed which is the based on the concept of DBSCAN algorithm which detect the clusters of varied densities within a cluster. LDD-DBSCAN algorithm checks for the density variation of every neighbor of the core object within the given radius value ε.

If the value of density variance is less than of a core objects with respect to it is ε-neighborhood is less than a specified threshold value k. Then only a core object is allowed for expansion otherwise the object is simply added into the cluster. As result good homogenous density varied clusters are formed.

Rest of the paper is organized as follows. Section 2 presents related work on density based clustering technique for high dimensional database. Section 3 discuses the existing DBSCAN clustering algorithm and required modification to get better clustering results. The proposed modification along with algorithm is discussed in section 4. Experimental results are presented in section 5. Finally, Section 6 presents conclusion and future work.

II. RELATED WORK

DBSCAN[6] (Density based spatial clustering application with noise) is a basic density based algori thm is able to detect arbitrarily shaped clusters by one single pass over the data. To do so, DBSCAN uses the fact, that a density connected cluster can be detected by finding one of its core object's p and computing all objects which are density-reachable from p. The retrieval of density-reachable objects is performed by iteratively collecting directly density-reachable objects. DBSCAN checks the e-neighborhood of each object p in the database. If Ne (p) of an object p consists of at least MinPts objects, i.e., if p is a core object, a new cluster C containing all objects of Ne (p) is created. Then, the e-neighborhood of all objects q Î C, which has not yet been processed, is checked. If object q is also a core object, the neighbors of q, which are not already assigned to cluster C, are added to C and their e-neighborhood is checked in the next step. This procedure is repeated until no new object can be added to the current cluster C. Then the algorithm continues with an object, which has not yet been processed, trying to expand a new cluster. The computational complexity of the algorithm is O (n log n).Where n is the number of objects to be clustered.

The algorithm **OPTICS[7]** (Ordering Points to Identify the Clustering Structure) is an extension of the density-connected clustering notion of DBSCAN by hierarchical concepts. In contrast to DBSCAN, OPTICS does not assign cluster memberships but computes an ordering in which the objects are processed and additionally generates the information. This information consists of two values for each object, the core-distance and the reach ability -distance. It starts with an arbitrarily chosen object p □ D,assigns a reach ability of ∞ to object p and expands the cluster order if the core-distance of p is smaller than the specified parameter. The expansion is worked out by inserting each object q □ □ N□ □ (p) into apriority queue. The priority queue stores that object first, having the minimum reach ability to all already processed objects. The heap structure is maintained which updates the reach ability of the objects that are already in the priority queue if their according values decrease. The next object to be inserted in the cluster ordering is always the first object of the priority queue. If the core distance of this object is smaller or equal to □ , all objects in the □-neighborhood are again inserted into or updated in the priority queue. If the priority queue is empty and there are still some not yet processed objects in D, another not yet handled object in D is chosen to further expand the cluster ordering. The computational complexity of the algorithm is O (n 2). Where n is the number of objects to be clustered

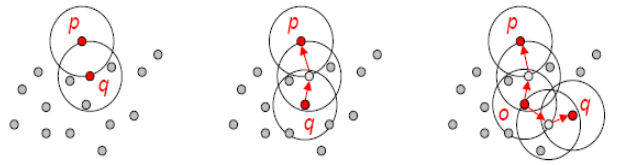
DENCLUE[8] (Clustering Based on Density Distribution Function) is another density-based algorithm. The basic idea of DENCLUE is to model the overall object density analytically as the sum of influence functions of the data objects. The influence function can be seen as a function, which describes the impact of a data object within its neighborhood. Then, by determining the maximum of the overall density function can identify clusters. The algorithm allows a compact mathematical description of arbitrarily shaped clusters in high-dimensional data sets and is significantly faster than the other density based clustering algorithms. Moreover, DENCLUE produces good clustering results even when a large amount of noise is present. As in most other approaches, the quality of the resulting clustering depends on an adequate choice of the parameters. In this approach, are two important parameters, namely s and ξ. The parameter s determines the influence of an object in its neighborhood and ξ describes whether a density-attractor is significant. Density-attractors are local maxima of the overall density function. The computational complexity of the algorithm is O (n log n). Where n is the number of objects to be clustered.

The clustering techniques described above try to find clusters of variable sizes and shapes. But these algorithms fail to detect the varied clusters within a cluster. Proposed algorithm detects good homogeneous density varied clusters within a cluster.

III. INTRODUCTION TO DBSCAN ALGORITHM

DBSCAN (Density-Based Spatial Clustering of Application with noise) [6] relies on density based notion of clusters and is designed to discover clusters of arbitrary shape as well as to distinguish noise. DBSCAN can cluster objects as well as spatially extended objects according to their spatial and non-spatial dimensions. Density based clustering is based on the fact that clusters are of higher density than its surroundings. In the following, the basic definitions of density-based clustering are presented.

- The ε-neighborhood of an object p, denoted by N_ε (p), is defined as N_ε (p) = {q ∈ D | dist (p, q) ≤ ε}.
- If the ε-neighborhood of an object p contains at least minimum number, MinPts of objects, and then the object is called a core object i.e. an object p is core if | N_ε (p)| ≥ MinPts.
- An object p is directly density reachable from an object q with respect to ε and MinPts if p ∈ N_ε (q) and N_ε (q) ≥ MinPts. As shown in figure 3.1(a).
- An object p is density-reachable from an object q with respect to ε and MinPts if there is a chain of objects $p_1 \ldots p_n$, p_1= q, p_n= p such that p_{i+1} is directly density-reachable from p_i. As shown in figure 3.1(b).
- An object p is density-connected to an object q with respect to ε and MinPts if there is an object o such that both, p and q are density reachable from o with respect to ε and MinPts. As shown in figure 3.1(c).



(a) *p* is directly density-reachable from *q*, but *q* is not directly density-reachable from *p*.

(b) *p* is density-reachable from *q*.

(c) *p* and *q* are density-connected.

Figure 3.1: Direct Density Reachable, Density Reachable & Density Connected Objects

- Let D be a database of objects. A cluster C with respect to ε and MinPts is a non-empty subset of D satisfying the following conditions:

1) ∀p, q : if p ∈ C and q is density-reachable from p with respect to ε and MinPts, then q ∈ C. (Maxi1mality).

2) ∀p, q∈C: p is density-connected to q with respect to ε and MinPts(Connectivity).

- Let $C_{1,}$…..C_k be the clusters of the dataset D with respect to parameters ε and $MinPts_i$, i = 1, . . . , k. Then noise is defined as the set of objects in the database D not belonging to any cluster C_i i.e. Noise= {p ∈ D | ∀i: p ∉ C_i}.
- A border object is not a core object, but it falls within the ε-neighborhood of a core object.

Given a dataset, ε and MinPts as input, DBSCAN searches for clusters by checking the ε-neighborhood of each object in the dataset. If the ε-neighborhood of an object p contains more than MinPts, a new cluster with p as core object is created. DBSCAN then iteratively collects directly density-reachable objects from these core objects. The process terminates when no new objects can be added to any cluster.

IV. ROPOSED ALGORITHM

A. Description of the proposed algorithm

It is based on the concept that it computes the density variation of a core object with respect to the densities of all its ε-neighborhood. If the density variance of a core objects with respect to its ε-neighborhood is less than a specified threshold value k. Then only a core object is allowed for expansion otherwise the object is simply added into the cluster.

In addition to DBSCAN the following definitions are required in LDD-DBSCAN to allow the considerable density variation within the same cluster and wide density variation with other clusters.

Defnition.1 (Density variance): The density variance of an object o in connection to the ε-neighborhood (o) is defined as follows:

$$DensityVariance[o] = \frac{\sum_{x \in N_\varepsilon(o)} (|N_\varepsilon(o)| - |N_\varepsilon(x)|)^2}{\text{Total number of objects in } N_\varepsilon(o)}$$

It is denoted by Density variance (o). If the variance of a core object o is less than specified threshold value, i.e. Density variance (o) ≤ k Then it may allowed for expansion in the cluster.

Defnition.2 (Density Similar Object): An object o is said to be a density similar object if the density variance of the objects with it ε-neighborhood (o) is less than the specified threshold value k.

Definition.3 (Relative Density similar core objects): An object o is said to be density similar core object if it follows the following two conditions:

(i) Object o must be core object i.e. $|N_\varepsilon(o)| \geq \mu$ (MinPts).

(ii) Object o must be density similar object i.e. Density variance (o) ≤ k, where k is the threshold value.

The user-specified parameter k is used to detect the density differ clusters, not only separated by the sparse region but also separated by the region which does not have the considerable density variation. If any core object which is having either the sparse region or the dense region in its ε-neighborhood then the density variance of the core object will be greater than the specified threshold value k.

b) LDD-DBSCAN Algorithms (D, ε , k, μ)

1. Initially all objects are unclassified.
2. For each unclassified object o ∈ D.
3. If Core (o) then
4. Generate new Cluster ID & Assign the clusterID to o.
5. Insert o into the Queue.
6. While Queue ≠ Empty.
7. Extract front object y from the Queue.
8. Calculate R={x ∈ D | dist (y, x) ≤ ε}.
9. For each object x ∈ R .
10. If x is unclassified and Relative Density similar core object.
11. Then insert x into Queue.
12. If x is unclassified or noise.
13. Then assign the clusterID to x.
14. End For.
15. End while.
16. Else o is noise
17. End for

C) Formation of cluster

The proposed algorithm starts to form a cluster by selecting the Core object; it inserts the selected Core object into the Queue. It pops out the front object from the seed list i.e. Queue. It calculates all the ε-neighborhood of that Core object and finds out

the all the Relative Density Similar Core objects. It inserts all the Relative Density Similar Core objects for further expansion in the Queue, if still unclassified. The process continued until all the objects classified or declared as noise.

v. Experimental evaluation

To compare the performance of the proposed algorithm, we have also implemented the well known DBSCAN algorithm. Java is used as a language to implement the algorithms. The performances of the above two algorithms are evaluated by using the 2-Dimensional synthetic dataset. The 2-Dimensional synthetic dataset is containing 4000 objects in 2-Dimensional plane. We performed the experiments by using different values of parameters.

The fig 1-2, shows the clusters detected by the DBSCAN and LDD-DBSCAN for the mentioned parameters value. The common parameters, i.e. μ and ε are having the same values for both the algorithms. In figure 1 due to global parameter μ and ε, DBSCAN detects only one cluster, because it cannot handle the density variation within a cluster.

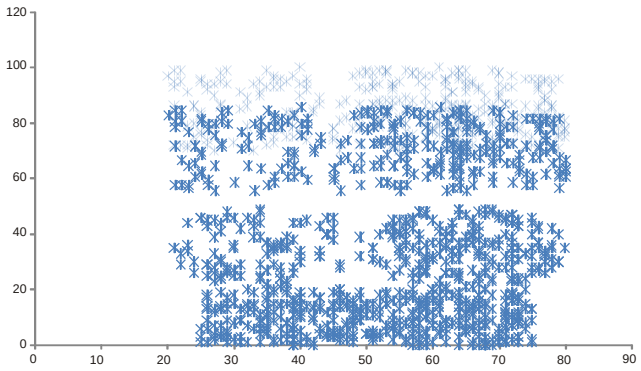


FIG 1 Cluster generated by DBSCAN algorithm for the values of, μ= 15and ε=0.5.

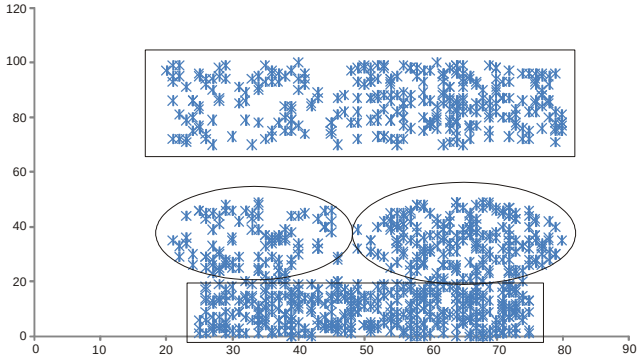


Fig 2 clusters generated by LDD-DBSCAN algorithm for the values of ε = 0.5 , μ = 15 and k=0.25 .

In addition to two global input parameters used by DBSCAN algorithm, LDD-DBSCAN algorithm which uses one more parameter i.e. K, four clusters (as shown in figure 2). This shows that LDD- DBSCAN algorithm is able to handle the density variations that exist within the cluster. The clusters detected by LDD-DBSCAN algorithm are having considerable density variation within clusters. The detected clusters are not only separated by the sparse region but also separated by the regions, having the density variations. This illustrates that LDD-DBSCAN outperform the DBSCAN, especially in the case of density variation within the clusters.

VI.CONCLUSION

In this paper we proposed LDD-DBSCAN, an enhancement of DBSCAN algorithm. The proposed algorithm can find cluster that represent relatively uniform regions. A parameter K is used to limit the amount of allowed local density variations within a cluster. The future work can be focused on to reduce the time complexity of algorithm and to determine the value parameter K automatically for better clustering of any given dataset.

VII. REFERENCES

[1]. Jain, A.K. Dubes, R.C. "Algorithm for clustering data" Printice Hall Englewood cliffs NJ.1998.
[2]. Han, J. and Kamber, M. "Data Mining: Concepts and Techniques". Morgan Kaufman, 2001.
[3]. B.Borah, D.k. Bhatacharyya "DDSSC: A Density Differentiated Spatial Clustering Technique ". JOURNAL OF COMPUTERS, VOL. 3, NO. 2, FEBRUARY 2008.

[4]. Lian Duan, Deyi Xiong, Jun Lee and Feng Guo "A local density based spatial clustering algorithm with noise" 2006 IEEE International Conference on Systems, Man, and Cybernetics October 8-11, 2006, Taipei, Taiwan.

[5]. B. Borah' and D.K. Bhaftacharyya2 "A Clustering Technique using Density Difference" IEEE - ICSCN 2007, MIT Campus, Anna University, Chennai, India. Feb. 22-24, 2007.

[6]. Ester, M. Kriegel, H.-P. Sander, J. and Xu, X. . "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise". In Proc. KDD, 1996.

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

[8]. Hinneburg, A., Keim, D. 1998. DENCLUE: An efficientapproach to clustering in large multimedia data sets withnoise. In proceedings of 4th International Conference onKnowledge Discovery and Data Mining. pp. 58–65.

