Data Partitioning Technique to Enhance DBSCAN Clustering Algorithm

Safaa O. Al-Mamory
College of Business Informatics ,University of Information Technology and Communications
salmamory@uoitc.edu.iq
Esraa Saleh Kamil
esraa@itnet. uobabylon.edu.iq
University of Babylon

Abstract

Among density- based clustering techniques ,DBSCAN is a typical one because it can detect clusters with widely different shapes and sizes, but it fails to find clusters with different densities and for that we propose a new technique to enhance the performance of DBSCAN on data with different densities ,the new solution contains two novel techniques ,one is the separation (partitioning) technique that separate data into sparse and dense regions, and the other is the sampling technique that produce data with only one density distribution. the experimental results on synthetic data show that the new technique has a clustering
Keywords: Density-based clustering , DBSCAN, different densities , separation, sampling. result better than that of DBSCAN.

الخلاصة

من بين تقنيات التجميع المعتمدة على الكثافة , تعتبر DBSCAN تقنية نموذجية لأنها تستطيع ايجاد مجموعات ذات احجام واشكال مختلفة , لكنها لا تستطيع ايجاد مجموعات ذات كثافات مختلفة ولهذا فقد اقترحنا تقنيه جديده لتحسين اداء DBSCAN مع البيانات ذات الكثافات المختلفة , الحل الجديد يتضمن تقنيتين جديدتين ,الأولى هي تقنية الفصل (التقسيم) التي تفصل البيانات الى مناطق متناثرة وكثيفة ,والأخرى هي تقنية أخذ عينات تنتج بيانات ذات توزيع كثافة واحد فقط . وقد اوضحت النتائج التجريبية على البيانات الاصطناعية أن التقنية الجديدة تمتلك نتيجة تجميع أفضل من تقنية تجميع DBSCAN .

الكلمات المفتاحية : التجميع المعتمد على الكثافة , DBSCAN , الكثافات المختلفة , الفصل , اخذ العينات .

1. Introduction

Recently, data have been increased in observable form, where it became received from many equipment's and companies. In order to deal with this huge amount of data, we need to separate data into smaller understandable groups.

Clustering is one of the important methods to analyze data by grouping them based on similarity between data points or based on density distribution , In general there are four major types of clustering based on its application : partition- based, hierarchical –based methods, density- based methods and grid- based methods (Miller, 2008).

DBSCAN (Ester, 1996) is one of the effective density based clustering algorithms that can detect clusters with different shapes and sizes ,but it fails to detect clusters with different densities, this failure resulted from being used a global density threshold (***minpts***) on all the data points .

In the last decade a large number of methods were proposed to handle the problem of different densities in DBSCAN . In this paper, we introduce a new method to enhance DBSCAN to find clusters with different densities by adopting a novel data partitioning technique ,where we separate data based on the levels of density containing on the data . This partitioning is achieved by applying two constraints on the data points on global and local levels ,by this process, the points will be recognized as dense points and sparse points .Then taking sample from the dense points by applying a new sampling technique .

By these processes , we will provide to DBSCAN data with only one density level so that it can find clusters from it perfectly. The remaining dense points after sampling (other density level) will be clustered with the core points resulted from DBSCAN by using KNN (k-nearest neighbors) algorithm. The experimental results show that the performance of DBSCAN improved to large extent .

The remaining of the paper is ordered as follows , Section 2 presents related work , a general overview on DBSCAN will be presented in Section 3 ,our proposed algorithm will completely implemented and illustrated in Section 4 , experimental results on synthetic data were applied and tested in Section 5 , and Section 6 will discuss the main conclusions.

2.Related Works

The drawback of DBSCAN in finding efficient clusters when there are different densities in data urged many researchers to develop methods to handle the problem of dealing with these types of data.

By partitioning data into several density levels (Xiong , 2012) have been developed a new method to analyze the characteristics of these density levels ,and determining ***Eps*** and ***Minpts*** differently to each density level and ultimately applying DBSCAN multiple times ,one time for each level of density .

According to (Liu, 2012) ,the exploitation of the spatial proximity and attributes similarity in spatial clustering will improve this clustering to detect clusters in the spatial domain by employing a modified density – based clustering method ,they conclude that the objects that belong to same cluster ,detected by their method, has a proximity in the spatial domain and similarity in an attribute domain.

When using spatial index together with grid technique the result is GMDBSCAN ,this method uses space dividing technique and consider each grid as a separate part ,then it estimates independent ***Minpts*** for every grid (part) based on its density ,after that it applies multiple DBSCAN on each grid ,and finally it uses distance –based method to amend boundary (Xiaoyun, 2008)**.**

In the same context Ren *et al.*,(Ren *et al.,* 2012) proposed a new method called DBCAMM that developed DBSCAN. Firstly, by replacing Euclidian distance by Mahalanobies distance metric ,this metric is associated with the distribution of the data and secondly, by introducing a method to combine sub-clusters by using the information of the density of the sub-cluster.

In the form of constraints, Ruiz *et al.*,(Ruiz , *et al.*,2010) presented a density –based semi supervised clustering method ,where they use DBSCAN to produce temporary clusters ,which then combined by applying two types of constrains Must-link constraint and cannot-link constraint .

The concept of nearest neighbors of each data point where developed by (Levent, 2003***)***, their idea was to compute the number of neighbors that shared by each pair of points. This novel definition of similarity helps in removing noise and outliers ,recognizing core points ,and creating clusters around the core points .

GRPDBSCAN ("Grid-based DBSCAN algorithm with referential parameters") is another solution to the problem of different densities ,which merged both of grid partition technique and multi-density based clustering algorithm ,also it generates ***Eps*** and ***Minpts*** automatically to enhance DBSCAN (Darong , 2012)**.**

Spatial clustering based on Delaunay triangulation (ASCDT) was presented by (Deng, 2011), it uses a new technique to cut the edges of Delaunay triangulation in two levels : global level and local level, to find clusters spatially ,where in the global effect ,the clustering is applied on the features that are well separated ,then after the global effect is removed, the local effect is considered .

Other approaches head toward some of data structures to handle the problem ,such as using two rounds of minimum spanning trees in two phases to detect separated clusters whether it was separated by distance or separated by density in the first phase , in the next phase it performs partitioning on the sub groups resulted from the first phase called touching clusters by comparing cuts in the two rounds of minimum spanning trees (Zhong, 2010).

Zhang and Xu (Zhang&Xu, 2013) introduced a new technique depends on four concepts: Contribution ,grid technique , migration-coefficient, and tree index structure to optimize the performance of DBSCAN to be able to discover clusters with different densities . This optimization carried out by, firstly, using grid technique to reduce the time where the algorithm will be efficient for large databases. Secondly, the optimization of the clustering results is fulfilled by expressing the density of the grid based on the concept of contribution . Thirdly, the improving of the clustering quality will be done by focusing on boundary points using migration coefficient.

In M-DBSCAN(Multi density-DBSCA***)*** (Amini,2014), neighbors didn't found with a constant radius ϵ ,instead the determination of neighboring radius is performed based on the data distribution around the core using standard deviation and mean values. To get the clustering results, M-DBSCAN is applied on a set of core-mini clusters where each core-mini cluster represents a virtual point lies in the center of that cluster. The ϵ value of DBSCAN is replaced by local density cluster in M-DBSCAN, in this algorithm, the clusters are extended by adding core-mini clusters that have similar mean values with a little difference determined by the standard deviation of the core.

3. Preliminary

Density-based clustering considers that each point of a cluster , must has number of neighbors less than a specified radius ***Eps*** . DBSCAN clustering algorithm (Ester, 1996) is one of the more effective algorithms in clustering data based on their density that is implemented to find clusters and noise points for spatial data .

Ester et al. (Ester *et al*., 1996),mentioned that what make us identify the clusters from noise points is the high contrast in density within and outside the cluster ,where the density within a cluster is more higher than the density outside it.

The cluster in DBSCAN is identified as a set of points that are highly density connected ,where each core point in a cluster must has number of neighbors greater than a specified density threshold called ***minpts*** (minimum number of points) within the determined ***Eps*** (Ren *et al*., 2012)

The main requirements of DBSCAN are to determine the two parameters ***Eps*** and ***minpts*** of each cluster and start with one arbitrary point of that cluster, then all the points that are density reachable from the specified point could be retrieved by using the suitable parameters . If the arbitrary point is a core point then there is a cluster and will be given a cluster id , if the arbitrary point is a border point , where the border point is an on-dense pattern has a distance less than or equal to ***Eps*** from a dense pattern , then that point has

no density reachable points from it and DBSCAN will progress to the next point in the dataset .Since DBSCAN uses global parameters , It merges two clusters of different density only if these clusters close to each other.

There is a basic approach to determine the parameters of this algorithm which represented by looking at the behavior of the distance from a point to its k-nearest neighbors which is called k-dist , This process could be briefly described as the process of taking k-nearest neighbors for each point and sorted these values increasingly and then plotted them .We expect to see a sharp change at the value of k-dist that corresponds to a suitable value of Eps and the value of k will represent *minpts* parameter.

4. The Proposed Algorithm

As we mentioned above ,the new algorithm consists of four major steps started by separating data into two groups based on their density ,taking sample from one of the resulted groups ,cluster sampled data and sparse data with DBSCAN clustering algorithm ,and finally clustering the remaining dense data with KNN clustering algorithm ,Figure.1 shows the whole steps of the proposed system.

4.1 Separation of Data

To separate data ,Spatial proximity relationships among objects of data should be constructed, as discussed by Liu *et al* .(Liu *et al*., 2012),who mentioned that the methods of finding spatial proximity are divided into two types of relations: Distance relations such as k-nearest neighbors and Eps-neighborhood method , and topological relations such as Delaunay triangulation method , in this paper we use distance relations represented by k-nearest neighbors and (global and local means) constraints to get spatial proximity relationship among objects.

After constructing k-nearest neighbors for each object ,the next step is to apply global mean constraint and local mean constraint ,these constraints are used to determine the level of density of a point in the data. Suppose that our data set called D then the global

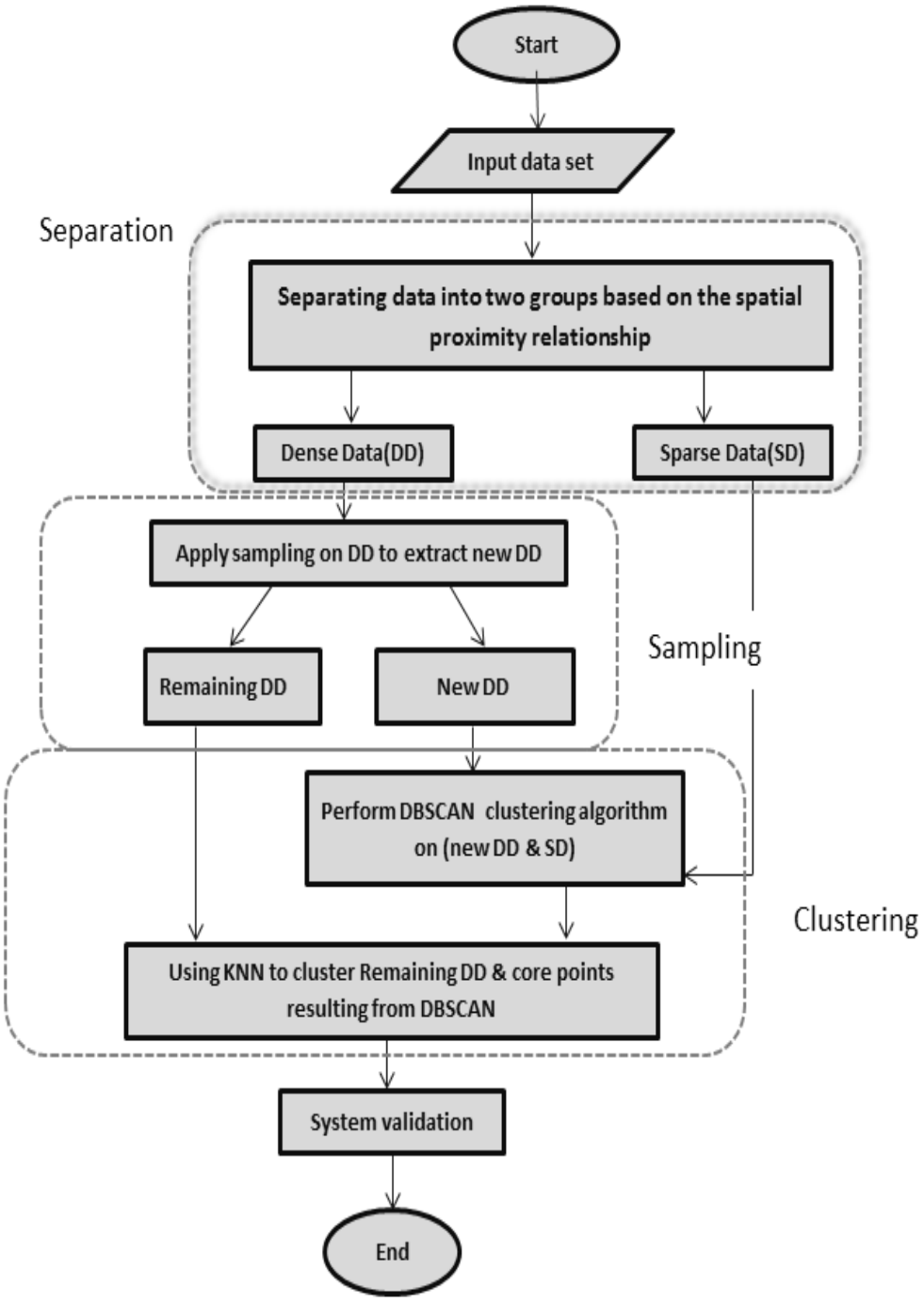

Figure (1): Architecture of proposed system

mean GM (D) represents the mean value of all values in the Sparse matrix and is defined as :

$$\mathbf{GM}=\frac{1}{n\times m}\sum_{i=1}^{n}\sum_{j=1}^{m}\boldsymbol{dist}(\boldsymbol{p_i},\ \boldsymbol{p_j}\ )$$

where n represents the number of objects in D and $\boldsymbol{dist}(\boldsymbol{p_i},\boldsymbol{p_j}) \neq \mathbf{0}$ in sparse matrix.

The purpose of computing global mean is to identify the average neighborhood value in the global level (the relationship of the point with all other points in the data), After computing global mean we need to compute local mean of each object p_i in the data , local mean of object $\boldsymbol{p}$ represents the mean value of all the values that are in the same row of object $\boldsymbol{p}$ in the sparse matrix(the average value of k nearest neighbors of object $\boldsymbol{p}$), Local mean(LM) was computed by the following formula :

$$\mathbf{LM}(\boldsymbol{p_i})=\frac{\mathbf{1}}{\boldsymbol{k}}\sum_{l=1}^{k}\boldsymbol{dist}(\boldsymbol{p_i},\ \boldsymbol{p_l})$$

Where k represents number of neighbors to the point $\boldsymbol{p_i}$ in sparse matrix ,By computing local mean we identify the neighborhoods of the point(the relationship of point with only points that are directly incident to point $\boldsymbol{p_i}$) (Liu et al., 2012)**.**

In order to separate objects according to regions' densities ,the proposed method supposes that each point has local mean greater than global mean then it lies in sparse region and labeled as **SD** ,so any point fulfills this condition will be considered as sparse point, on the other hand any point has local mean smaller than global mean will considered to be lie in the dense region and labeled as **DD**, from this we could extract two definitions that illustrate our proposed method :

Definition 1: (dense point), any point has local mean smaller than global mean

LM(pi) < GM(D)

Definition 2: (sparse point),any point has local mean greater than global mean

LM(pi) > GM(D)

The following algorithm illustrates the whole procedure of separation.

Algorithm Separation (D, k)

Input: data set D, number of k nearest neighbors k
Output: two groups of data DD and SD
1.Begin
2. Compute distance matrix DM for all points in D .
3. Compute sparse matrix SM from DM.
4. Compute GM (SM) where
5. GM=$\frac{1}{n\times k}\sum_{i=1}^{n}\sum_{j=1}^{k}dist(p_i,\ p_j\ )$
6. For each point $\boldsymbol{pi}$ in SM
7. compute LM ($\boldsymbol{p_i}$)
8. if LM($\boldsymbol{p_i}$) > GM then
9. {SD}={SD}+ $\boldsymbol{p_i}$
10. else
11. {DD}={DD}+ $\boldsymbol{p_i}$
12. End for
13. End.

By completing these steps ,we will be able to determine how many levels of density are in the data, and the experiments show that this method can determine more than two levels of density by applying the previews constraints .

4.2 Sampling Based on Density

Sampling is a technique of choosing a representative part of a population in order to determine some characteristics of the whole population .In this step, sampling is applied on dense data only to get part of the data that has density similar to the density of sparse data.

In this paper ,we propose a new sampling technique depending on the global mean mentioned previously , after separating the data into two groups : DD and SD ,the sampling will be applied on DD in order to get a sample having similar density to that of SD. The new sampling technique is done by computing the following :

Algorithm Sampling (DD,SD)

Input : the set of dense points DD and the set of sparse points SD.
Output : subset of dense points .
1.Begin
2. Compute GM(DD).
3. Compute GM(SD).
4. While (GM(DD) < GM(SD))
5. select point ***p*** from DD randomly .
6. remove ***p*** from DD.
7. re-compute distance matrix and sparse matrix for DD.
8. compute new GM(DD).
9. GM(DD) = new GM(DD).
10. END While .
11.END.

4.3.Clustering Sparse Data by DBSCAN.

In this step we will perform DBSCAN on the data to construct clusters, not all the data will be used , instead we will use only sparse data resulted from original sparse points in the dataset and sparse points that have been constructed from sampling process.

The main purpose is to provide DBSCAN with data with one density level where DBSCAN works well, so we will override the difference in density that is found in the data . As it is well known , to apply DBSCAN the values of its parameters (***minpts***, ***Eps***) should be determined ,the determination of these parameters depends on a prior knowledge about the data ,and consequently effect on the final result of clustering ,so to determine the value of ***Eps*** we used k-dist (k-distances) plot where the value of ***Eps*** could be determined where there is sharp change in the curve of the plot (knee).

From this step the number of clusters will be known and we will get three types of points: Core points, border points ,and noise where each core point will be given cluster id that defines to which cluster it does belong , core points will be passed to the next step of clustering with noise points.

4.4. Clustering of Dense Data and Core Points by KNN

KNN introduced by (Fix & Hodges, 1951) as one of the simplest algorithms ,the aim of this clustering method is to separate data based on the assumed similarities between various classes.

KNN is a non-parametric(does not consider distribution of the data) and lazy(does not make any generalization using the training data) learning algorithm ,one of the considerations used by KNN is that it considers data in a metric space, it also considers that each of the training data contain vector and class attribute and it able to work with any number of classes.

In this paper ,we exploit this procedure with little modification where the cluster id assigned to core points in DBSCAN clustering algorithm(Subsection 3.3) will be used as a label to cluster data instead of the real class label of data points , KNN will be used to cluster dense data remaining from sampling(Subsection 3.2) with core points resulted from DBSCAN(Subsection 3.3),the following steps illustrate the steps of our modified version of KKN:

- Determine K (number of nearest neighbors).
- Take point of dense data and compute the distance between this point and all core points resulted from DBSCAN .
- Select K smallest distances .
- Determine the cluster id that is most frequent among k-nearest neighbors and give the dense point to the cluster with that common cluster id.

In order to reduce time complexity of this algorithm we use KD tree(K dimensional tree) data structure to find k nearest neighbors to each point in the dense data ,KD tree is a special case of binary search tree that split points in a way where each level in the tree represents one dimension of the data .

Algorithm *KNN clustering*(core points, dense points ,k)

Input: core points (training data),dense points (test data) ,k (number of nearest neighbors).
Output: n clusters.
1.Begin
2. for each dense point d in data do
3. compute dist (core ,d)
4. select D_d the set of closest core points to dense point d.
5. cluster id(d)=$argmax \sum(d) \in D_d \ \ I(v = cid)$
6. end.
7.End.

5.Experimental results

For the test purposes, several two-dimensional synthetic datasets are used to prove the benefits of new proposed algorithm, all the synthetic datasets have the characteristics of difference in shape, difference in size, and difference in density that is the focus of this paper.

Many clustering algorithms cannot produce effective clustering on these datasets with the mentioned characteristics .Thus in this paper we evaluate our proposed algorithm with these datasets .Table.1 shows the accuracy of separating each dataset. Clustering results in Figure.2 show the performance of the new algorithm on squares dataset , X dataset, Gaussian dataset , jain dataset, and fish dataset respectively . The parameters ***minpts*** and ***Eps*** of DBSCAN are selected by using k-dist plot that could be used to determine the range of values that the ***Eps*** could belong to, the accuracy of clustering is the percentage of objects that belong to cluster that is labeled correctly .

Table 1: Accuracy of separation

Datasets	**Size of dataset**	**Accuracy**
Squares (DS1)	**600**	**98.00%**
X (DS2)	**305**	**92.13%**
Gaussian (DS3)	**1100**	**91.91%**
Fish (DS4)	**2100**	**94.09%**
Jain (DS5)	**373**	**96.77%**

The observations show that the new algorithm can perform effective clustering on data with different shapes, sizes, and densities according to the selected values represented by ***k*** that used to separate data , ***minpts*** and ***Eps*** used in DBSCAN and the number of neighbors used in KNN.

In order to validate the new proposed algorithm it compared with original DBSCAN ,we compute the accuracy of two algorithms on each dataset, the observation illustrates that the accuracy of the new algorithm is better than the accuracy of DBSCAN under the same parameters (***minpts*** , **Eps**) and that means that the new algorithm enhanced the performance of DBSCAN to some extent.

The results in Table. 2 show that the accuracy of the proposed algorithm enhanced for most of the datasets , especially for fish dataset which has more overlapped densities and that represent a challenge for the clustering process by DBSCAN , and this enhancement proves the affectivity of the proposed algorithm .

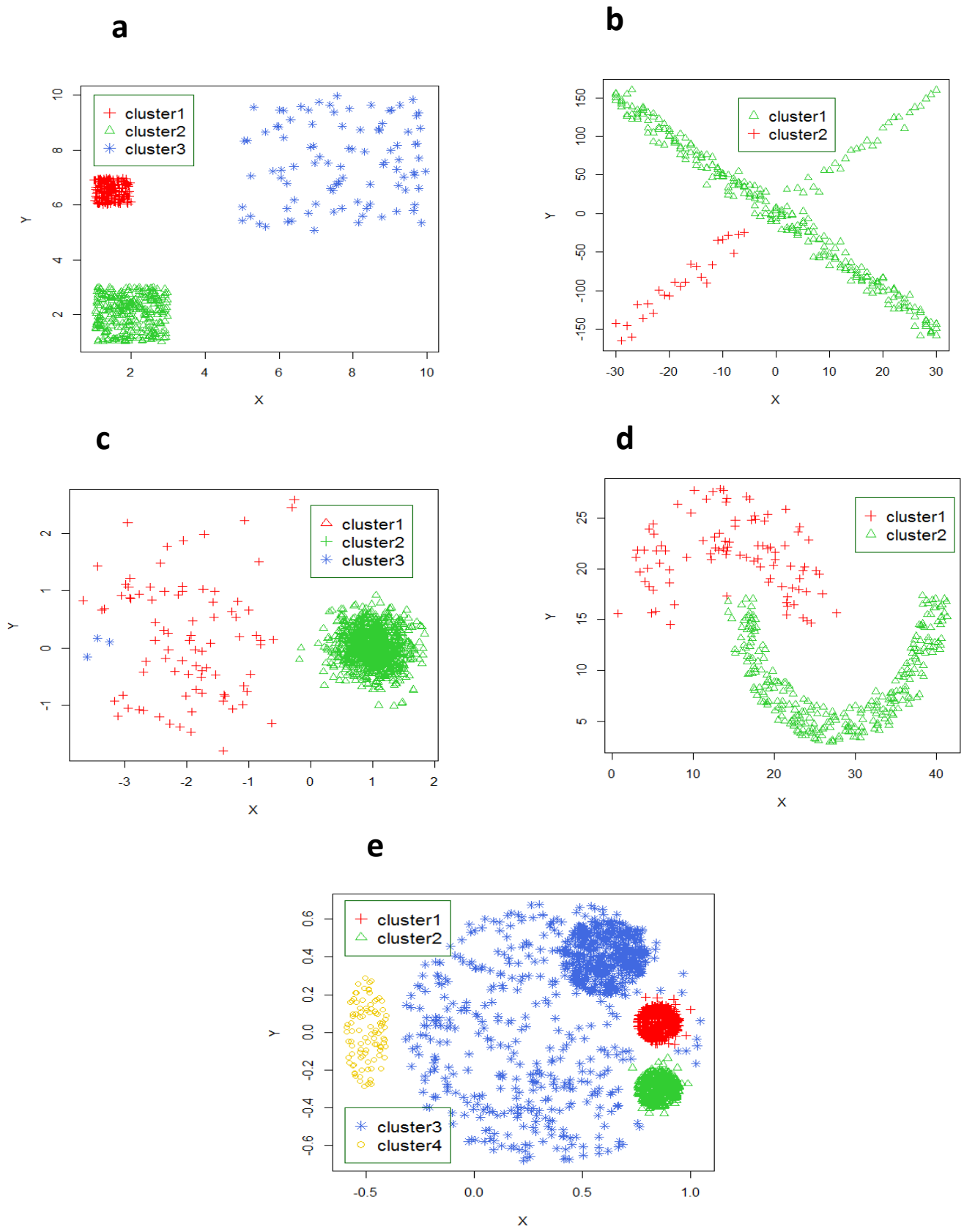

Figure (2) : The clustering results of the proposed algorithm on (a)squares , (b)X data , (c) Gaussian , (d) jain , (e) fish

Table 2 :Comparison between DBSCAN and the proposed algorithm

Dataset	Accuracy of DBSCAN	Accuracy of proposed algorithm
Squares	95.49%	100.00%
Fish	52.63%	77.51%
X	89.77%	79.54%
Gaussian	93.08%	99.72%
Jain	91.97%	99.39%

6.Conclusions

In this paper ,a new technique was presented, in which the data is separated into groups based on the density that determined by applying some constraints .In the clustering process, the data was separated perfectly, then new sampling technique was applied in order to reduce the density of dense data and obtain data with only one density distribution (sparse data), the results of sampling were very effective .The experiments performed on synthetic data show that the proposed algorithm enhances the performance of DBSCAN on data with different densities as the accuracy measure prove that .

7. References

Amini, A., Saboohi, H., Herawan, T., & Wah, T. Y. (2014). MuDi-Stream: A multi density clustering algorithm for evolving data stream. *Journal of Network and Computer Applications*, 1–16. http://doi.org/10.1016/j.jnca.2014.11.007

Darong, H., & Peng, W. (2012). Grid-based DBSCAN Algorithm with Referential Parameters. *Physics Procedia*, *24*, 1166–1170. http://doi.org/10.1016/j.phpro.2012.02.174

Deng, M., Liu, Q., Cheng, T., & Shi, Y. (2011). An adaptive spatial clustering algorithm based on delaunay triangulation. *Computers, Environment and Urban Systems*, *35*(4), 320–332. http://doi.org/10.1016/j.compenvurbsys.2011.02.003

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *Second International Conference on Knowledge Discovery and Data Mining*, 226–231. http://doi.org/10.1.1.71.1980

Fix, E., & Hodges, J. (1951). Discriminatory analysis, nonparametric discrimination: Consistency properties. *US Air Force School of Aviation Medicine, Technical* (3).

Levent Ertöz, M. S. V. K. (n.d.). Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data.

Liu, Q., Deng, M., Shi, Y., & Wang, J. (2012). A density-based spatial clustering algorithm considering both spatial proximity and attribute similarity. *Computers and Geosciences*, *46*, 296–309. http://doi.org/10.1016/j.cageo.2011.12.017

Miller, H. (2008). Geographic data mining and knowledge discovery. *The Handbook of Geographic Information Science*, 651.

Ren, Y., Liu, X., & Liu, W. (2012). DBCAMM: A novel density based clustering algorithm via using the Mahalanobis metric. *Applied Soft Computing Journal*, *12*(5), 1542–1554. http://doi.org/10.1016/j.asoc.2011.12.015

Ruiz, C., Spiliopoulou, M., & Menasalvas, E. (2010). Density-based semi-supervised clustering. *Data Mining and Knowledge Discovery*, *21*(3), 345–370. http://doi.org/10.1007/s10618-009-0157-y

Xiaoyun, C., Yufang, M., Yan, Z., & Ping, W. (2008). GMDBSCAN: Multi-density DBSCAN cluster based on grid. *IEEE International Conference on E-Business Engineering, ICEBE'08 - Workshops: AiR'08, EM2I'08, SOAIC'08, SOKM'08, BIMA'08, DKEEE'08*, 780–783. http://doi.org/10.1109/ICEBE.2008.54

Xiong, Z., & Chen, R. (2012). Multi-density DBSCAN Algorithm Based on Density Levels Partitioning DBSCAN Algorithm and Related Work Introduction to DBSCAN Algorithm, *10*, 2739–2749.

Zhang, L., Xu, Z., & Si, F. (2013). GCMDDBSCAN: Multi-density DBSCAN Based on Grid and Contribution. *2013 IEEE 11th International Conference on Dependable, Autonomic and Secure Computing*, 502–507. http://doi.org/10.1109/DASC.2013.115

Zhong, C., Miao, D., & Wang, R. (2010). A graph-theoretical clustering method based on two rounds of minimum spanning trees. *Pattern Recognition*, *43*(3), 752–766. http://doi.org/10.1016/j.patcog.2009.07.010