An Efficient DBSCAN using Genetic Algorithm based Clustering

Lovely Sharma, Prof. K. Ramya

Abstract— Data mining is widely employed in business management and engineering. The major objective of data mining is to discover helpful and accurate information among a vast quantity of data, providing a orientation basis for decision makers. Data clustering is currently a very popular and frequently applied analytical method in data mining. DBSCAN is a traditional and widely-accepted density-based clustering method. It is used to find clusters of arbitrary shapes and sizes yet may have trouble with clusters of varying density. In this paper an efficient DBSCAN clustering using genetic algorithm is proposed. DBSCAN clustering provides some problem such as the algorithm is not efficient for noisy clusters; hence it is enhanced using genetic algorithm. The proposed technique is efficient in terms of accuracy and execution time.

.

—————————— ◆ ——————————

1 Introduction

TDue to this the large amount of text are usually uploaded into many sites and thus it need to be classified. Data mining is the process of extracting useful information from databases. Many approaches to temporal data mining have been proposed to extract useful information, such as time series analysis, temporal association rules mining, and sequential pattern discovery. Several core techniques that are used in data mining describe the type of mining and data recovery operation.

Clustering is still an important research issue in the data mining, because there is a continuous research in data mining for optimum clusters on spatial data. There are various types of partition based and hierarchal algorithms implemented for clustering and the clusters which are formed based on the density are easy to understand and it does not limit itself to certain shapes of the clusters. Density-based clustering methods try to find clusters based on the density of points in regions. Intense or dense regions that are accessible from each other are merged to produce clusters. Density-based clustering methods surpass at finding clusters of arbitrary shapes [1].

A spatial database system is a database system for the management of spatial data. Speedy growth is happening in the number and the size of spatial databases for applications such as traffic control, geo-marketing, and environmental reviews. Spatial data mining or knowledge discovery in spatial databases links to the mining from spatial databases of contained knowledge, spatial relations, or extra patterns that are not unambiguously stored [1].

Clustering schemes are classified as hierarchical partitioning, density-based, grid-based and mixed methods. Partitioning methods are the most popular clustering algorithms. The advantage of partitioning approaches is fast clustering, while the disadvantages are the instability of the clustering result, and inability to filter noise data. Hierarchical methods involve constructing a hierarchical tree structure, and adopting it to perform clustering. These methods have high clustering accuracy, but suffer from continuously repetitive merging and partitioning: each instance must compare the attribute of all objects, leading to a high calculation complexity. Density-based methods perform clustering based on density. These approaches can filter noise, and perform clustering in tangled patterns, but take a long time to execute clustering. Grid-based clustering algorithms segment data space into various grids, where each data point falls into a grid, and perform clustering with the data points inside the grids, thus significantly reducing the clustering time [2].

Spatial clustering goals to group alike objects into the same group based on considering both spatial and non-spatial attributes of the object and a regular clustering algorithm can be modified to account for the special nature of spatial data to give a spatial clustering algorithm [3].

a. DBSCAN

Density-based approaches apply a local cluster criterion and are very popular for the purpose of database mining. Clusters are regarded as regions in the data space in which the objects are impenetrable that are separated by regions of low object density (noise). A common way to find regions of high density in the data space is based on grid cell densities [4]. The basic idea for the algorithm is that the data space is partitioned into a number of non overlapping regions or cells, and cells containing a relatively large number of objects are potential cluster centers. However, the success of the method depends on the size of the cells which must be specified by the user. DBSCAN algorithm is based on center-based approach, one of definitions of density [5]. In the center-based approach, density is estimated for a particular point in the dataset by counting the number of points within a particular radius, Eps, of that point. This comprises the point itself. The center-based approach to density permits to categorize a point as a core or main point, a border point, a noise or background point. A point called core point if the number of points inside Eps, a user-specified parameter goes beyond a certain threshold; MinPts is also a user-specified parameter [6].

a. Spatial Data Mining

Clustering is still an important research issue in the data mining, because there is a continuous research in data mining for optimum clusters on spatial data. There are numerous types of partition based and hierarchal algorithms implemented for clustering and the clusters which are formed based on the density are easy to understand and it does not limit itself to certain shapes of the clusters.

Spatial data mining is the branch of data mining that deals with spatial (location, or geo-referenced) data. The knowledge tasks involving spatial data include finding characteristic rules, discriminate rules, association rules; etc. A spatial characteristic rule is a general description of spatial data. A spatial discriminate rule is a common explanation of the features discriminating or contrasting a class of spatial data from other class. Spatial association rules describe the association among objects, derived from spatial neighborhood relations. It can associate spatial attributes with spatial attributes, or spatial attributes with non-spatial attributes [7].

b. Various Clustering Techniques

K-MEANS CLUSTERING is a method of cluster analysis which aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean. The algorithm is called *k-means,* where k represents the number of clusters required, since a case is allocated to the cluster for which its distance to the cluster mean is the negligible. The achievement in the algorithm centres on finding the *k-means* [8].

HIERARCHICAL CLUSTERING builds a cluster hierarchy or a tree of clusters, it is also known as a ‘dendrogram’. All cluster nodes contains child clusters; sibling clusters partition the points covered by their common parent [8].

DBSCAN finds all clusters properly, independent of the shape, size, and location of clusters to everyone, and is greater to a widely used Clarans method. DBscan is based on two main concepts: density reach ability and density connect ability. These both concepts depend on two input parameters of the DBSCAN clustering: the size of epsilon neighbourhood e and the minimum points in a cluster m. The number of point’s parameter impacts detection of outliers. Points are declared to be outliers if there are few other points in the *e*-Euclidean neighbourhood. *e* parameter controls the size of the neighbourhood, as well as the size of the clusters. The Euclidean space has an open set that can be divided into a set of its connected components. The execution of this idea for partitioning of a finite set of points requires concepts of density, connectivity and boundary [8].

OPTICS ("Ordering Points to Identify the Clustering Structure") is an algorithm for finding density-based clusters in spatial data. Its fundamental idea is comparable to DBSCAN, but it addresses one of DBSCAN's main weaknesses: the problem of detecting significant clusters in data of changeable density. Consecutively the points of the database are linearly ordered such that points that are spatially closest become neighbours in the ordering. Furthermore, a special distance is stored for each point that corresponds to the density that needs to be accepted for a cluster in order to have both points belong to the same cluster [8].

a. SOFT DBSCAN

Soft DBSCAN is a very much recent clustering techniques. This technique combines DBSCAN and fuzzy set theory [9]. The idea is to improve the clusters generated by DBSCAN by fuzzy set theory which is based on an objective function, in order to produce optimal fuzzy partitions. This new method could provide a similar result as Fuzzy C Means, but it is simple and superior in handling outlier points. Thusly, the Soft DBSCAN’s first stage runs DBSCAN which creates, many seed clusters, with a bunch of noisy points. Each noisy is consider as one cluster. These determined groups, in addition of noisy clusters, with their centers, offer a good estimate for initial degrees of membership which express proximities of data entities to the cluster centers. This update the membership values in every iteration since these last ones depend on the new cluster centers. When the cluster center stabilizes “soft DBSCAN” algorithm stops.

b. Fuzzy Set Theory

The notion of a fuzzy set provides a convenient point of departure for the construction of a conceptual framework which parallels in many respects the framework used in the case of ordinary sets, but is more general than the latter and, potentially, may prove to have a much wider scope of applicability, particularly in the fields of pattern classification and information processing. Fuzzy set theory provides a strict mathematical framework (there is nothing fuzzy about fuzzy set theory!) in which vague conceptual phenomena can be precisely and rigorously studied. It can also be considered as a modeling language, well suited for situations in which fuzzy relations, criteria, and phenomena exist. Fuzzy Logic can be applied to a Decision Tree to generate a Fuzzy Rule-Based System. This is particularly useful when at least some of the attributes that are tested in the decision nodes are numerical. In this case, the tests can be formulated using labels (e.g., “if Temperature is low”) and the children nodes will be partially activated. As a consequence, two or more (possibly conflicting) terminal nodes will be partially activated and an aggregation method should be used to generate the final output/conclusion of the FRBS [10].

2. BACKGROUND

Data mining is used everywhere and large amounts of information are gathered: in business, to analyze client behavior or optimize production and sales. This encompasses a number of technical approaches like data summarization, clustering, analyzing changes, data classification, finding dependency networks, and detecting anomalies. Data mining is the search for the relationship and global patterns that exist in large database but are hidden among vast amount of data, such as the relationship between patient data and medical diagnosis. This relationship represents valuable knowledge about the database, if the database is a realistic epitomize of the real world registered by the database.

3. RELATED WORK

Xin Wang and Howard J. Hamilton presented A Comparative Study of Two Density-Based Spatial Clustering Algorithms for Very Large Datasets [1]. They compare two spatial clustering methods. DBSCAN gives extremely good results and is efficient in many datasets. However, if a dataset has clusters of widely changeable densities, DBSCAN is unable to handle it proficiently. If non-spatial attributes play a role in determining the desired clustering result, DBSCAN is not appropriate, because it does not consider non-spatial attributes in the dataset. DBRS aims to reduce the running time for datasets with varying densities. It scales well on high-density clusters. DBRS can be deal with a property associated to non-spatial attribute(s) through a purity threshold, when finding the matching neighbourhood. One limitation of the algorithm is that it sometimes may fail to combine some small clusters [1].

Cheng-Fa Tsai and Chun-Yi Sung proposed DBSCALE: An Efficient Density-Based Clustering Algorithm for Data Mining in Large Databases [2]. They present a novel clustering algorithm that incorporates neighbour searching and expansion seed selection into a density-based clustering algorithm. Data Points have been clustered require not be input again when searching for neighbourhood data points and the algorithm redefines eight Marked Boundary Objects to add expansion seeds according to far centrifugal force that increases coverage. Investigational results point out that the proposed

DBSCALE has a lower execution time cost than KIDBSCAN, MBSCAN and DBSCAN clustering algorithms. DBSCALE has a highest divergence in clustering accuracy rate of 0.29'Yo, and a maximum deviation in noise data clustering rate of 0.14% [2].

DBSCAN Algorithm proposed by Ester et al. in 1996 [4], was the first clustering algorithm to employ density as a condition. It utilizes density clustering to place data points into the same cluster when their density within their data points is higher than a set threshold value, and sets this cluster as the seed for outward expansion. This algorithm must set two parameters, the radius (e) and the minimum number of included points (MinPts). DBSCAN can conduct clustering on disordered patterns, and has noise filtering capacity, as well as clusters that can be stabilized [4].

K. Ganga Swathi and KNVSSK Rajesh proposed Comparative analysis of clustering of spatial databases with various DBSCAN Algorithms [6]. They present the comparative analysis of the various density based clustering mechanisms. There is certain problem on existing density based algorithms because they are not capable of finding the meaningful clusters whenever the density is so much different. VDBSCAN is commenced to compensate this problem. It is same as DBSCAN (Density Based Spatial Clustering of Applications with Noise) but only the difference is VDBSCAN selects several values of parameter Eps for different densities according to k-dist plot. The difficulty is the significance of parameter k in k-dist plot is user defined. This introduces a new technique to find out the value of parameter k automatically based on the characteristics of the datasets. In this method they consider spatial distance from a point to all others points in the datasets [6].

The clustering algorithm is based on density approach and can detect global as well as embedded clusters. Investigational outputs are reported to establish the superiority of the algorithm in light of several synthetic data sets. In this they [6] considered two-dimensional objects. But, spatial databases also contain extended objects such as polygons. Due to that, there is possibility for scaling the proposed algorithm to detect clusters in such datasets with minor alterations, research is in progress. From a proper analysis of the intended technique, it can be securely concluded that the algorithm implemented is working appropriately to a great extent [6].

DBSCAN algorithm is based on center-based approach, one of definitions of density. In the center-based approach, density is estimated for a particular point in the dataset by counting the number of points within a particular radius, *Eps*, of that point. This contains the point itself. The center-based approach to density tolerates to classify a point as a core point, a noise, a border point and background point. A point called core point if the numerous points inside *Eps*, a user-specified parameter, surpass a certain threshold, *MinPts*, which is a user-specified parameter [6].

Pragati Shrivastava and Hitesh Gupta present a review of Density-Based clustering in Spatial Data [7]. Spatial data mining is the branch of data mining that deals with spatial (location, or geo-referenced) data. The knowledge tasks involving spatial data include finding characteristic rules, discriminate rules, association rules; etc. A spatial characteristic rule is a general description of spatial data. A spatial distinguish rule is a universal description of the features discriminating or contrasting a class of spatial data from other class. Spatial association rules describe the association between objects, based on spatial neighbourhood relations. They can associate spatial attributes with spatial attributes, or spatial attributes with non-spatial attributes. They represent the density based clustering. That is uses to reduced core points, outliers and noise. When reduces this points than increase the efficiency of clustering. Core points are basically related to the centres at any single tone problem and noise is the combination of outlier and core point [7].

Manish Verma et al [8] proposed A Comparative Study of Various Clustering Algorithms in Data Mining. They provide a comparative study among various clustering. They compared six types of clustering techniques- k-Means Clustering, Optics, DBScan clustering, Hierarchical Clustering, Density Based Clustering and EM Algorithm. These clustering techniques are implemented and analyzed using a clustering tool

WEKA. Performances of the 6 techniques are presented and compared. Running the clustering algorithm using any software produces almost the same result even when changing any of the factors because most of the clustering software uses the same procedure in implementing any algorithm [8].

Abir and Eloudi presented Soft DBSCAN: Improving DBSCAN Clustering method using fuzzy set theory [9]. They propose a novel clustering algorithm called “Soft DBSCAN” which is inspired by FCM algorithm. Much of the strength of this approach comes from FCM’s ideas. The plan of “soft DBSCAN” is to make the DBSCAN’s clusters robust, extending them with the fuzzy set theory. DBSCAN is run in the first phase to produce a set of clusters with diverse shapes and sizes, in the company of noisy data discrimination. In the second phase, it computes the degrees of fuzzy membership which express proximities of data entities to the cluster centers. They suggested method does not only outperform FCM clustering by detecting points expected to be noises and handling the arbitrary shape, but also by generating more dense clusters. Evaluations demonstrate that our solution generates more accurate groups for input dataset and objective function is improved better than FCM [9].

In year 2012, Xiaojun LOU, Junying LI, and Haitao LIU proposed a technique to Improved Fuzzy C-means Clustering Algorithm Based on Cluster Density. They study on the distribution of the data set, and introduce a definition of cluster density as the representation of the inherent character of the data set. A regulatory factor based on cluster density is proposed to correct the distance measure in the conventional FCM. It differs from other approaches in that the regulator uses both the shape of the data set and the middle result of iteration operation. And the distance measure function is dynamically corrected by the regulatory factor until the objective criterion is achieved. Two sets of experiments using artificial data and UCI data are operated. Comparing with some existing methods, the proposed algorithm shows the better performance. The experiment results reveal that FCM-CD has a good tolerance to different densities and various cluster shapes. And FCM-CD shows a higher performance in clustering accuracy [11].

Andrew McCallum et al [12] offered Efficient Clustering of High Dimensional Data Sets with Application to Reference Matching. have focused on *reference matching*, a particular class of problems that arise when one has many different descriptions for each of many di_erent objects, and wishes to know (1) which descriptions refer to the same object, and (2) what the best description of that object is. They present experimental results for the domain of bibliographic reference matching. Another significant illustration of this class is the *merge-purge problem*. Companies often purchase and merge multiple mailing lists. The resulting list then has multiple entries for each household. Even for a single person, the name and address in each version on the list may diverge slightly, with middle initials absent or present, words shortened or expanded, zip codes present or absent. This problem of merging large mailing lists and eliminating duplicates becomes even more complex for *householding*, where one wishes to collapse the records of multiple people who live in the same household.

Hrishav Bakul Barua et al [13] offered A Density Based Clustering Technique for Large Spatial Data Using Polygon Approach. The technique of data clustering has been inspected, which is a particular type of data mining problem. The procedure of grouping a set of physical or abstract objects into classes of similar objects is called clustering. The objective of this paper is to present a Triangle-density based clustering technique, which named as TDCT, for efficient clustering of spatial data. This algorithm is accomplished of recognizing embedded clusters of arbitrary shapes as well as multi-density clusters over large spatial datasets. The Polygon approach is being accessed to execute the clustering where the number of points inside a triangle (triangle density) of a polygon is calculated using barycentre formula. This is because of the information that partitioning of the data set can be performed more efficiently in triangular shape than in any other polygonal shape due to its smaller space dimension. The ratio of numerous points among two triangles can be found out which forms the basis of nested clustering [13].

Chaudhari Chaitali G. “Optimizing Clustering Technique based on Partitioning DBSCAN and Ant Clustering Algorithm [14]. Clustering is the process of organizing similar objects into the same clusters and dissimilar objects in to dissimilar cluster. Correspondences between objects are estimated by using the attribute value of object; a distance metric is used for evaluating difference. DBSCAN algorithm is striking because it can find arbitrary shaped clusters with noisy outlier and require only two input parameters. DBSCAN algorithm is very successful for analyzing huge and complex spatial databases. DBSCAN necessitate bulky volume of memory support and has complexity with high dimensional data. Partitioning-based DBSCAN was suggesting overcoming these problems. But DBSCAN and PDBSCAN algorithms are responsive to the initial parameters [14].

They [14] present a new algorithm based on partitioning-based DBSCAN and Ant-clustering. This algorithm can partition database in to N partitions according to the density of data. New PACA-DBSCAN algorithm reduces the sensitivity to the initial parameters and also can deal with data of uneven density. This algorithm does not need to discuss the distribution of data on each dimension for multidimensional data. PACA-DBSCAN algorithm can cluster data of very special shape. To evaluate the performance of proposed algorithm they use three dataset to compare with other algorithms [14].

Glory H. Shah et al [15] proposed An Empirical Evaluation of Density-Based Clustering Techniques. Conventional database querying methods are inadequate to extract useful information from massive data banks. Cluster investigation is one of the most important data analysis methods. It is the ability of detecting groups of comparable objects in bulky data sets without having specified groups by means of unambiguous features. The difficulty of detecting clusters of points is challenging when the clusters are of unusual size, density and shape. The expansion of clustering algorithms has received a lot of attention in the last few years and many new clustering algorithms have been proposed [15].

Santosh Kumar Rai and Nishchol Mishra "DBCSVM: Density Based Clustering Using Support Vector Machines [16]. They present an improved DBSCAN clustering algorithm named DBCSVM: Density Based Clustering Using Support Vector Machines. In the process of feature extraction generator, huge amount of matrix for the calculation of description of feature for the purpose of clustering, for this purpose previous density based clustering take more time and does not give better result. From this method the separation of farer and nearer points are very efficient. The farer points jumps into the next step of clustering. This method gives better result and takes less time comparison to previous DBSCAN clustering methods [16].

4. RELATED WORK

The proposed methodology is a combinatorial method of DBSCAN clustering and genetic algorithm. First of all DBSCAN clustering is applied on the noisy dataset and then gentic algorithm is applied on these noisy clustered data so that the clustering gets efficient.

DBSCAN Clustering Algorithm

1. Arbitrary select a point *p*
2. Retrieve all points density-reachable from *p* wrt *Eps* and *MinPts*.
3. If *p* is a core point, a cluster is formed.
4. If *p* is a border point, no points are density-reachable from *p* and DBSCAN visits the next point of the database.
5. Continue the process until all of the points have been processed.

Genetic Algorithm

(1) Initialization: The first process decides initial genotype, namely value and genetic length. Fig.1. shows the basic steps taken by the genetic algorithm.
(2) Evaluation: The second process calculates the fitness for each individual with the target function. The evaluation depends on each problem.
(3) Termination Judgment: If the process satisfies the termination condition, the operation finishes and output the individual with the best fitness as the optimized solution.
(4) Selection: To generate the children, this process chooses parents from individuals. For example, if we assume parents the first generation, children become the second generation. The children generate the next children again. The children inherited the characteristic of the parents are generated in this way.
(5) Crossover: This process crosses individuals chosen by selection operation and generates the individuals of the next generation.
(6) Mutation: This process mutates the chromosome of new generation. The mutation is effective to escape from a local optimum solution.

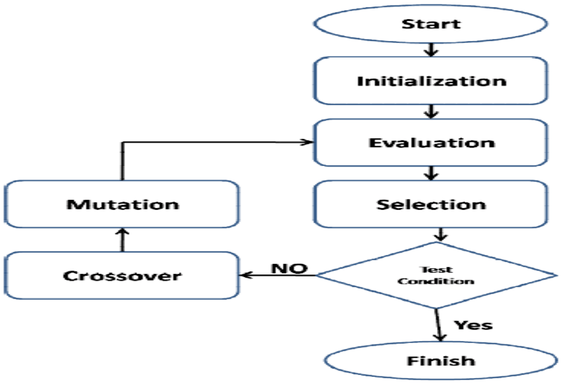

Figure 1.Flow Diagram of GA

```
Genetic Algorithm()
{
Initialize population;
Evaluate the initial population;
For all population
{
If( Test Condition=True)
{
Search Element Found
}
Else
{
Apply CrossOver();
And Mutation();
}
}
```

5. RESULT ANALYSIS

The experimental results will be analysed on the basis of following dataset.

Dataset	Instances	attributes
Iris	150	4
Ecoli	336	8
Ionosphere	351	34
Breast-W	698	9
Breast-T	748	5
Indian	768	9

Figure 2. Dataset used

The figure shown below is the result analysis of our proposed work on different dataset. The experimental shows the performance of the proposed work. Here in the result analysis PCC (classification accuracy), OBJ (Objective function) , FCM (Fuzzy Performance Index), PC (Partition Coefficient) will be shown.

Dataset	PCC		OBJ		FPI		PC		PE	
	FCM	Our work	FCM	Our work	FCM	Our work	FCM	Our work	FCM	Our work
Iris	**0.66**	**0.57**	**18.12**	**20.23**	**0.93**	**0.95**	**0.34**	**0.32**	**0.55**	**0.57**
Ecoli	**0.86**	**0.69**	**29.32**	**32.13**	**0.43**	**0.47**	**0.28**	**0.25**	**0.63**	**0.65**
Ionosphere	**0.88**	**0.67**	**133.97**	**135.65**	**0.72**	**0.74**	**0.65**	**0.63**	**0.52**	**0.57**
Breast-W	**0.90**	**0.78**	**29.43**	**31.65**	**0.92**	**0.95**	**0.37**	**0.35**	**0.50**	**0.53**
Breast-T	**0.85**	**0.68**	**70.58**	**74.52**	**0.91**	**0.93**	**0.41**	**0.38**	**0.46**	**0.49**
Indian	**0.82**	**0.72**	**46.25**	**51.75**	**0.48**	**0.51**	**0.25**	**0.23**	**0.67**	**0.69**

Figure 3. Result Analysis

6. CONCLUSION

Clustering is a technique of grouping the similar objects and dissimilar objects. Clustering is a technique that can be applied for a variety of application. Although there are various clustering techniques implemented for various applications but some of the clustering techniques have various advantages and limitations. The clustering using combinatorial method o DBSCAN and genetic algorithm performs better as compared to other techniuqes used for clustering.

REFERENCES

[1] Xin Wang and Howard J. Hamilton " A Comparative Study of Two Density-Based Spatial Clustering Algorithms for Very Large Datasets", Proceedings of the 18th Canadian Society conference on Advances in Artificial Intelligence, pp. 120-132, 2005.

[2] Cheng-Fa Tsai and Chun-Yi Sung "DBSCALE: An Efficient Density-Based Clustering Algorithm for Data Mining in Large Databases", 2010 Second Pacific-Asia Conference on Circuits, Communications and System (PACCS-2010), pp. 98 – 101, 2010.

[3] Shekhar, S. and Chawla, S.: Spatial Databases: A Tour, Prentice Hall (2003) .

[4] Martin Ester,Han-peter Kriegel,Jorg Sander, Xiaowei Xu,"A Density-Based Algorithm for Discovering Clusters in Large

Spatial Databases with Noise”, 2nd International conference on Knowledge Discovery and Data Mining (KDD-96), pp. 226 – 231, 1996.

[5] Shashi Shekar & Sanjay Chawla, “Spatial Databases a Tour”, (ISBN 013-017480-7), Prentice Hall, 2003.

[6] K. Ganga Swathi , KNVSSK Rajesh “Comparative analysis of clustering of spatial databases with various DBSCAN Algorithms”, International Journal of Research in Computer and Communication technology (IJRCCT), ISSN 2278-5841, Vol. 1, Issue 6, pp. 340 -344, November 2012.

[7] Pragati Shrivastava and Hitesh Gupta “A Review of Density-Based clustering in Spatial Data”, International Journal of Advanced Computer Research, ISSN (online): 2277-7970, Volume-2, Number-3, Issue-5, pp. 200 – 202, September-2012.

[8] Manish Verma, Mauly Srivastava, Neha Chack, Atul Kumar Diswar, Nidhi Gupta “ A Comparative Study of Various Clustering Algorithms in Data Mining”, International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622, Vol. 2, pp.1379-1384, Issue 3, May-Jun 2012.

[9] Smiti, Abir, and Zied Eloudi. "Soft DBSCAN: Improving DBSCAN Clustering method using fuzzy set theory." In the IEEE 6th International Conference on Human System Interaction (HIS- 2013), pp. 380-385, 2013.

[10] Zimmermann, H-J. "Fuzzy set theory." Wiley Interdisciplinary Reviews: Computational Statistics, vol. 2, no. 3, pp. 317-332, 2010.

[11] Xiaojun LOU, Junying LI, and Haitao LIU “Improved Fuzzy C-means Clustering Algorithm Based on Cluster Density”, Journal of Computational Information Systems, vol. 8, issue 2, pp. 727-737, 2012.

[12] Andrew McCallum, Kamal Nigam and Lyle H. Ungar “Efficient Clustering of High Dimensional Data Sets with Application to Reference Matching”, Proceedings of the sixth ACM SIGKDD international conference on knowledge discovery and data mining, pp. 169-178, 2000.

[13] Hrishav Bakul Barua, Dhiraj Kumar Das & Sauravjyoti Sarmah “A Density Based Clustering Technique For Large Spatial Data Using Polygon Approach”, IOSR Journal of Computer Engineering (IOSRJCE) ISSN: 2278-0661 Volume 3, Issue 6, PP 01-09, 2012.

[14] Chaudhari Chaitali G. “Optimizing Clustering Technique based on Partitioning DBSCAN and Ant Clustering Algorithm”, International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-2, Issue-2, pp. 212 – 215, December 2012.

[15] Glory H. Shah, C. K. Bhensdadia, Amit P. Ganatra “An Empirical Evaluation of Density-Based Clustering Techniques”, International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-1, pp. 216 – 223, March 2012.

[16] Santosh Kumar Rai, Nishchol Mishra “DBCSVM: Density Based Clustering Using Support Vector Machines”, IJCSI International Journal of Computer Science Issues, ISSN (Online): 1694-0814, Vol. 9, Issue 4, No 2, pp. 223 – 230, July 2012.