# Novel Technique for Density Based Clustering Using Neural Networks

**Asha Devi[1], Saurabh Sharma[2]**
[1]Research Scholar, Department of Computer Science and Engineering, Sri Sai University, Palampur, India
[2]Assistant Professor,Department of Computer Science and Engineering, Sri Sai University, Palampur, India

## ABSTRACT

In order to separate similar and dissimilar type of data into different clusters, the clustering mechanism is used. This helps in analyzing the data that is given as input in more efficient manner. The EPS is computed in case when DBSCAN algorithm is applied on the data that is to be processed. The Euclidean distance is computed from the central point which is mainly the EPS point calculated. This helps in further defining the similar and dissimilar type of data present. The accuracy of clustering is minimized in cases where there is a dynamic calculation of EPS and static computation of Euclidean distance. The back propagation algorithm is used in this paper that computes the Euclidean distance in a dynamic manner. The DBSCAN algorithm has less complexity as compared to the other algorithms. It can also identify any type of shaped cluster which is difficult with in other algorithms. The identification of specific types of clusters is difficult in cases where objects are circulated in heterogeneous manner. As per the various experiments conducted here, it is seen that the accuracy of the algorithm is enhanced and the execution time is minimized.

**Keywords:**Clustering, DBSCAN, Back-Propagation, Accuracy, Execution Time.

## I. INTRODUCTION

In order to store the data in a proper manner various mechanisms have been evolved within the information technology. Data mining is one to those emerging technologies in which the data is stored in proper and organized manner and can be extracted as and when required [10]. There have been numerous data extracting methods involved as well. These methods ensure proper storage and retrieval of data along with proper query and transaction processing [11]. There have been numerous functionalities involved within this method which include the collection of data, the creation and management of database and various other activities [12]. The data warehouse is one of the famous data repository architectures evolving gradually with time. In order to handle the decision making, the multiple heterogeneous data sources are organized within unified mechanism[7]. Within numerous applications such as design recognition, information examination and image processing, the cluster analysis mechanism has been utilized. Within the business applications, the interests of the customers are gathered together and characterized into various groups. This helps in getting a proper analysis of the feedback of customers in this field [17]. The plant as well as animal taxonomies can be generated with in the biological applications here. The genes that have similar functionality are assembled together here and the studies are presented on the basis of this analysis [13]. An unsupervised classification mechanism that generates groups of objects or clusters on the basis of their various properties is known as data clustering mechanism. Here, there are different clusters generated for similar and dissimilar types of objects. Within the data mining, the cluster analysis is thus of major concern. On the basis of distance between various objects, the various partitioning methods are presented. It is easy to identify the spherical shaped clusters however difficult to discover the arbitrary shaped clusters present [8]. In order to identify these arbitrary shaped clusters, the density-based clustering mechanisms are utilized that are based on the density of the clusters present [14]. When the density of an area is higher than the threshold value, the cluster is generated with in these methods [15]. The main objective here is to provide at minimum amounts of points for the radius of a cluster for Each information point inside the cluster [9]. Also the arbitrary shaped clusters are identified with the help of this mechanism.There is a need of various density parameters with in this type of algorithm [16].

## II. Literature Review

DBSCAN algorithm is one of the most prominently utilized clustering algorithms which provide an efficient cluster analysis. There is a division of normally divided database into various disjoint patterns here. A lot of time is taken for splitting and consolidating the high-dimensional space. In this paper, an parallel DBSCAN algorithm known as S_DBSCAN is present for resolving all such issues. This algorithm involves Spark in it and can identify the original data. It is seen through various simulations that the proposed algorithm out performs all the existing algorithms [1]. The arbitrary shaped clusters can be identified with the help of DBSCAN algorithm and also the noise of data can be eliminated. On the basis of MPI or OpenMP environments, the testing of parallelization of DBSCAN can be done. There are numerous problems arising within these systems such as the fault tolerance is less here along with imbalance in the workload. The communication amongst the nodes is also to be handled with the help of programming in MPI. The performance has been improved with the utilization of this proposed algorithm in this paper [2]. An efficient method that identifies embedding and nested adjacent clusters. The clusters are generated here on the basis of density here. In the proposed method, the global density parameters are utilized for highlighting the major issues arising within the clustering mechanism along with the DBSCAN algorithm. Also, with in the EnDBSCAN algorithm, the identification of nested adjacent clusters is done. It is seen through various experimental results that the proposed algorithm is more efficient for identifying the embedded and nested adjacent clusters. The computational complexity is also reduced here [3].DBSCAN algorithm is a real-time image super pixel segmentation method that involves 50fps. A quick two-stage method is proposed in this paper that helps in minimizing the computational costs of the super pixel algorithms. The pixels are clustered here with the help of DBSCAN algorithm that involves color similarity and geometric confinements. Within the secondary merging stage, the smaller clusters are merged into super pixels with the help of neighbor pixels. Here, the distance measurement is done with the help of color and spatial features. All such changes have been proposed in this new algorithm and the results have been improved as per the simulations performed [4]. A new clustering method that will help in identifying and extracting the information from the warehouses. DBSCAN has been utilized as a famous clustering algorithm however there have been many issues arising within it that might result in increasing the complexity of the algorithm. This algorithm thus cannot be applied to complex and large datasets.In this paper a three phase parallel version of DBSCAN has been proposed that attempts to eliminate all the issues arising in it. The simulation results ensure that the proposed mechanism has improved the outcomes achieved by providing scalable and correct results. This has also helped in increasing the efficiency of the mechanism [5]. An improvement in the incremental clustering method. The search space that is to be partitioned is limited here which helps in enhancing the performance of the algorithm as compared to the other incremental clustering algorithms available. This algorithm has been experimented on various sizes and dimensions of datasets. It is seen through the various results achieved that the proposed algorithm helps in incrementing the speed of incremental clustering process. This is higher as compared to the existing approaches. The search space for each partition is minimized here by partitioning the dataset .There is no filtering performed on the complete dataset at the same instant. The process is followed as per the steps and the results are provided that are better in comparison to the other previous results achieved [6].

## III. Research Methodology

In order to calculate the dense region from the dataset, the DBSCAN algorithm is utilized. The EPS value of the dataset is calculated which helps in determining the central point from the dense region. The Euclidean distance is calculated from central point to all the other points within the area. This helps in computing the similarity amongst the various data points of Euclidean distance. The clustering of different elements is done in different datasets. In order to enhance the accuracy, the dynamic calculation of accuracy of EPS values is done. The points that remain unclustered can also be clustered with the help of this method. The Euclidean distance will be calculated with the help of back propagation method which will help in enhancing the accuracy of clustering technique. This helps in minimizing the execution time of enhanced DBSCAN algorithm.

### DBSCAN Algorithm

Input:Dataset for clustering , desired and output patterns

Output: Clustering of input data

```
M← List of objects that may change their
centroids
D← Most dense regions in the dataset
For each point p(i) in P do
C← nearest centroid ()
Function nearest centroid ()
initpopulation P
      evaluate P ;
Network ConstructNetworkLayers()
InitializeWeights(Network, test cases)
For ( i=0;i=P ;i++)
      SelectInputPattern(Input fault values)
    ForwardPropagate(p)
    BackwardPropagateError(P)
    UpdateWeights(P )
End
Return (P)
C←P;
M←update centroid
End
For each r to M
For each ri in M do
 c <-ri new_centroid
 Co<-ri old_centroid
Apply incDbscanDel to remove ri from co
Apply incDbscanAdd to insert ri to cn
Add updated dense regions to D
end for
For each di in D do
For each dj in D and i – j do
If inter _connectivity (di,dj) > a merge
merge(di,dj)
end if
end for
end for
```

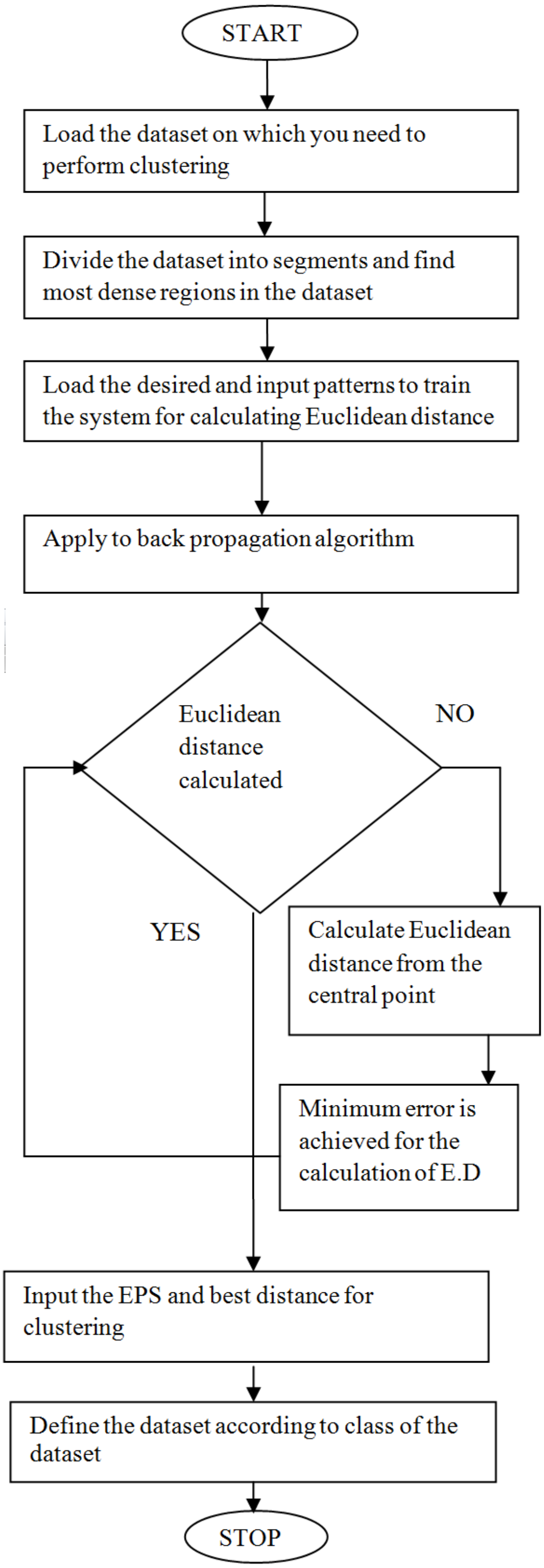
START
Load the dataset on which you need to perform clustering
Divide the dataset into segments and find most dense regions in the dataset
Load the desired and input patterns to train the system for calculating Euclidean distance
Apply to back propagation algorithm
Euclidean distance calculated
NO
YES
Calculate Euclidean distance from the central point
Minimum error is achieved for the calculation of E.D
Input the EPS and best distance for clustering
Define the dataset according to class of the dataset
STOP

**Figure 1**: Flow chart of Proposed Work

As illustrated in figure 1, the flowchart of the proposed improvement which is done in the DBSCAN algorithm to improve accuracy of clustering.In the existing DBSCAN algorithm EPS value is calculated dynamically and Euclidean distance is calculated statically which reduce efficiency of the algorithm. This work is based on to calculate Euclidean distance dynamically due to which technique back propagation algorithm is applied which define the Euclidean distance in the iterative manner and distance at which error is minimum is the final Euclidean distance . When the final Euclidean distance is considered similar and dissimilar type of data is clustered for analysis.

Within the data mining technology, Density-based clustering algorithm is the most popularly utilized method. The objects are grouped here with the help of a local criterion. With in the data space where the objects are dense, the clusters are considered to be the regions. The regions that have low object density are used for separating the various regions that exist [18]. The DBSCAN algorithm has less complexity as compared to the other algorithms. It can also identify any type of shaped cluster which is difficult within other algorithms. The identification of specific types of clusters is difficult in cases where objects are circulated in heterogeneous manner. There are two numeric input parameters known as minPts that are required for executing this algorithm. The required density based properties are presented with the help of these parameters. The least number of objects that might be present within the maximum distance of the data space is represented by the positive integer known as minPts. This helps in defining a place for an object within the cluster. The selection of these parameters must be done very carefully as the performance of DBSCAN relies largely on these input parameters. The scale of the dataset as well as the closeness of objects collectively provides the algorithm with huge changes on the basis of speed and effectiveness of the results being achieved. An exploration phase of numerous trial experiments is conducted here where the clustering is performed with some selected different values of different parameters. The parameters involved within this algorithm are related to the local density of the database components which can further help in recognizing the clusters that are present within the extensive spatial datasets. There is only one input parameter required here. There is very less knowledge of the domain required here and the reasonable parameter value is provided by the client as per its requirement. The classification of data can be done either as noise or outliers which can be identified by DBSCAN. The DBSCAN algorithm is very quick and can provide linear scaling in very efficient manner. The nodes can be categorized into different clusters on the basis of their properties with the help of density distribution of nodes within the database. The clusters that are of arbitrary shape can be identified with the help of DBSCAN algorithm. The clusters with similar properties can be put with in similar classes and result in forming larger clusters.

## IV. RESULTS AND DISCUSSION

The proposed and existing algorithm is implemented in MATLAB to test on the desired dataset.

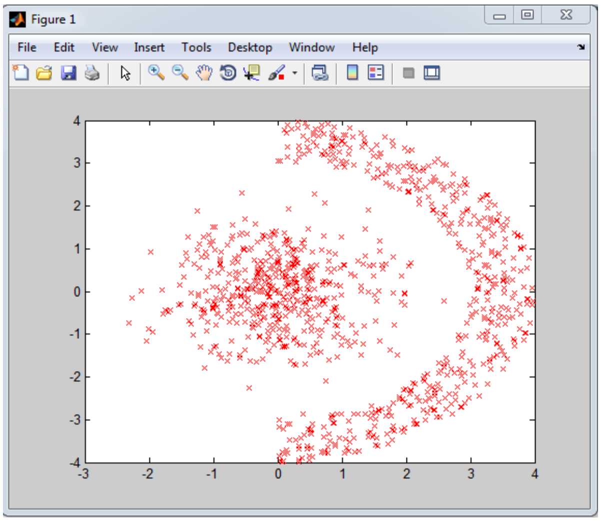

**Figure 2**: Incremental DBSCAN algorithm

As shown in figure 2, the algorithm of DBSCAN is applied which will cluster the similar and dissimilar type of data from the most dense region in the input dataset.

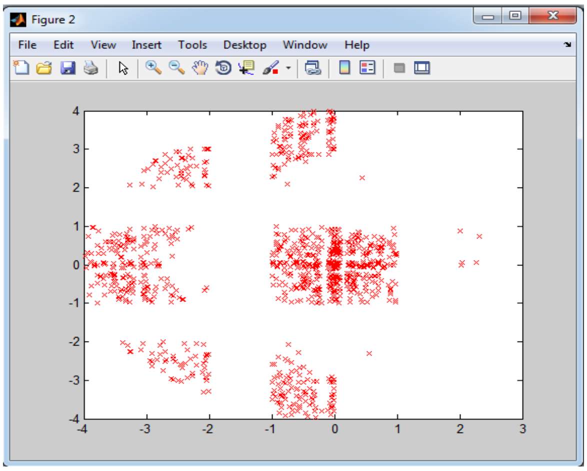

**Figure 3**: Competent DBSCAN Algorithm

As shown in figure 3, the improvement in the existing DBSCAN algorithm is been proposed in which back propagation algorithm is been applied to calculate Euclidean distance in dynamic manner this leads to increase accuracy of clustering.

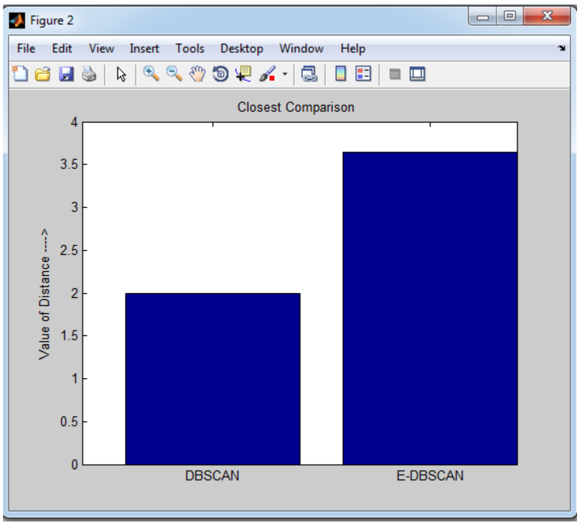

**Figure 4**: Distance Comparison

As shown in figure 4, the distance of the proposed DBSCAN algorithm and Incremental DBSCAN algorithm is compared and it is been analyzed that distance of Competent DBSCAN algorithm is more accuracy than existing DBSCAN Algorithm.

**Table 1:** Table of comparison

| Parameter | Incremental DBSCAN | Competent DBSCAN |
|---|---|---|
| Accuracy | 86 percent | 92 percent |
| Time | 5.5 second | 3.41 seconds |
| Distance | 2 | 3.7 |
| EPS | 1.33 | 0.9 |

As illustrated in table 1, the Performance of Incremental DBSCAN algorithm and Competent DBSCAN algorithm is compared in terms of accuracy, time, distance and EPS.

## V. CONCLUSION

In this work, it is been concluded that density based clustering is the efficient type of clustering in which clusters are defined on the density of the input data. The DBSCAN is the algorithm in which EPS value is calculated which will be central point and Euclidean distance is calculated from the central point which define similarity and dis-similarity of the data. In the existing work, euclidean distance is calculated in the static manner which is made dynamic in proposed work using boltzman learning algorithm.The proposed technique can be tested on the other datasets to check it accuracy and execution time.The proposed technique can be compared with the other techniques of density based clustering. The proposed improvement leads to increase accuracy of the clustering and reduction in execution time.

## VI. REFERENCES

[1]. Guangchun Luo, Xiaoyu Luo, Thomas Fairley Gooch, Ling Tian, Ke Qin," A Parallel DBSCAN Algorithm Based On Spark", 2016, IEEE.

[2]. Dianwei Han, Ankit Agrawal, Wei−keng Liao, Alok Choudhary," A novel scalable DBSCAN algorithm with Spark", 2016, IEEE.

[3]. Nagaraju S,Manish Kashyap, Mahua Bhattacharya," A Variant of DBSCAN Algorithm to Find Embedded and Nested Adjacent Clusters", 2016, IEEE.

[4]. Jianbing Shen, Xiaopeng Hao, Zhiyuan Liang, Yu Liu, Wenguan Wang, and Ling Shao," Real-time Superpixel Segmentation by DBSCAN Clustering Algorithm", 2016, IEEE.

[5]. Ilias K. Savvas, and Dimitrios Tselios," Parallelizing DBSCAN Algorithm Using MPI", 2016.

[6]. Ahmad M. Bakr , Nagia M. Ghanem, Mohamed A. Ismail," Efficient incremental density-based algorithm for clustering large datasets", 2014, Elsevier Pvt. Ltd.

[7]. Manpreet Kaur and Usvir Kaur, "Comparison Between K-Mean and Hierarchical Algorithm Using Query Redirection", International Journal of Advanced Research in Computer Science and Social , Volume 3, Issue 7, July 2013.

[8]. Harpreet Kaur and Jaspreet Kaur Sahiwal, "Image Compression with Improved K-Means Algorithm for Performance Enhancement," International Journal of Computer Science and Management Research, Volume 2, Issue 6, June 2013.

[9]. Kajal C. Agrawal and Meghana Nagori, "Clusters of Ayurvedic Medicines Using Improved K-means Algorithm," International Conf. on Advances in Computer Science and Electronics Engineering, 2013.

[10]. Anand M. Baswade, Kalpana D. Joshi and Prakash S. Nalwade, "A Comparative Study Of K-Means and Weighted K-Means for Clustering," International Journal of Engineering Research & Technology, Volume 1, Issue 10, December-2012.

[11]. Neha Aggarwal, Kirti Aggarwal and Kanika Gupta, "Comparative Analysis of k-means and Enhanced K-means clustering algorithm for data mining," International Journal of Scientific & Engineering Research, Volume 3, Issue 3, August-2012.

[12]. Ahamed Shafeeq B M and Hareesha K S, "Dynamic Clustering of Data with Modified Means Algorithm," International Conference on Information and Computer Networks, Volume 27, 2012.

[13]. Amar Singh and Navot Kaur, "To Improve the Convergence Rate of K-Means Clustering Over K-Means with Weighted Page Rank Algorithm," International journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 8, August 2012.

[14]. Osamor VC, Adebiyi EF, Oyelade JO and Doumbia S "Reducing the Time Requirement of K-Means Algorithm" PLoS ONE, Volume 7, Issue 12, 2012.

[15]. Azhar Rauf, Sheeba, Saeed Mahfooz, Shah Khusro and Huma Javed, "Enhanced K-Mean Clustering Algorithm to Reduce Number of Iterations and Time Complexity," Middle-East Journal of Scientific Research, pages 959-963, 2012.

[16]. Chieh-Yuan Tsai and Chuang-Cheng Chiu, "Developing a feature weight self-adjustment mechanism for a K-means clustering algorithm," Computational Statistics and Data Analysis, pages 4658-4672, Volume 52, 2008.

[17]. Tapas Kanungo , David M. Mount , Nathan S. Netanyahu Christine, D. Piatko , Ruth Silverman and Angela Y. Wu, "An Efficient K-Means Clustering Algorithm: Analysis and Implementation," IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume 24, July 2002.

[18]. M. N. Vrahatis, B. Boutsinas, P. Alevizos and G. Pavlides, "The New k-Windows Algorithm for Improving the k-Means Clustering Algorithm," Journal of Complexity 18, pages 375-391, 2002.