Enhancement of DBSCAN Algorithm and Transparency Clustering of Large Datasets

Kumari Silky[1], Nitin Sharma[2]
[1]*Research Scholar,* [2]*Assistant Professor*
Institute of Engineering Technology, Alwar, Rajasthan, India

Abstract: The data mining is the technique which can extract useful information from the raw data. The clustering is the technique of data mining which can group similar and dissimilar type of information. The density based clustering is the type of clustering which can cluster data according to the density. The DBSCAN is the algorithm of density based clustering in which EPS value is calculated which define radius of the cluster. The Euclidian distance will be calculated using neural networks which calculate similarity in the more effective manner. The proposed algorithm is implemented in MATLAB and results are analyzed in terms of accuracy, execution time.

Keywords: Clustering, Density, DBSCAN, Neural Networks

I. INTRODUCTION

To store different kind of data varieties of devices are present. In the purpose to avoid the chaos a structured database has been created. For the proper arrangement of huge data in effective manner a Database Management System (DBMS) has been evolved which help in achieving the objective. When the data is required by the users than that data can be retrieved efficiently by the use of DBMS. Because of the feature of proliferation available in the DBMS the huge collection of data is possible [1]. There is need to handle all the data from different field such as from business world, the scientific data, satellite pictures, text reports, or the military intelligence. Only using the information retrieval method is not enough in case of making decisions. Number of methods has been evolving for making the management of data better than the previous. There are number of activities which have to be taken care such as the activities which automatic summarize the data, extraction of the important stored and for the purpose of discovery of patterns in raw data. So, the analysis and interpretation of such huge stored files and database is very important. In further decision making process the above mention method is required for providing the important related information. The data mining technique has been largely utilized in order to provide huge research and developments within various scenarios. When the business data is initially stored within the computers, the initialization of such technique begins. This method needs various modifications along with the data accessing facility being utilized in the real time. Appropriate data access and navigation methods are needed in the data mining method in order to provide prospective and proactive information [2]. Clustering is the process of dividing the data into similar objects groups. A level of simplification is achieved in case of less number of clusters involved. But because of less number of clusters some of the fine details have been lost. With the use or help of clusters the data is modeled. According to the machine learning view, the clusters search in a unsupervised manner and it is also as the hidden patterns. The system that comes as an outcome defines a data concept [3]. The clustering mechanism does not have only one step it can be analyzed from the definition of clustering. Apart from partitional and hierarchical clustering algorithms number of new techniques has been evolved for the purpose of clustering of data. Then the different clustering techniques are implemented on the basis of various data sets present. The density based clustering algorithms group the objects on the basis of density objective functions of the systems. On the basis of number of objects that are present within the neighboring data object, the density of a specific object can be defined. As the number of objects increase with respect to certain parameters, there is growth of the cluster. In comparison to the paritional algorithms the methodology is different in case of the algorithms applied here. This is mainly due to the fact that the iterative relocation of points is utilized here in order to handle the fixed number of clusters of the networks. There is the partitioning of Euclidean space of open set in the set of connected components. The density, connectivity and boundary are required in order to partition the finite sets of points. There is a close relation between the nearest neighbors of the point [4]. A dense component can be considered as the cluster which has higher growth and can move to any direction for leading the density. The arbitrary shapes of the clusters are identified here with the utilization of density-based algorithms. With the help of this, the authentication against the outliers can be provided. There are numerous types of clustering techniques which can be utilized within data mining. They are partitioning, hierarchical, density, grid, model as well as constraint based clustering methods. On the basis of density based parameters, the density based clustering algorithm can be proposed. The thick regions can be generated by considering the regions that are apart from the thin regions. There is an increment in the identified number of clusters with the increase in density of the neighbors which is above threshold. The DBSCAN (Density Based Spatial Clustering of Applications with Noise) is one of the best performing clustering based algorithms [5]. The noise can be extracted from huge spatial databases with the help of arbitrary

shaped clusters. The two parameters utilized by this algorithm are Eps (radius) and MinPts (minimum points-a threshold). The number of points present within a specific radius Eps can be counted in order to estimate the density of data set at particular point. This approach is called the center-based approach. There are various categories on the basis of which the points are classified. Providing minimum number of points (MinPts) for the neighborhood of a certain radius (Eps) is a very important step in this algorithm. The DBSCAN algorithm, visits every point of the database. Numerous region query invocations are to be provided here in order to compute the time complexity on the basis of various practical algorithms [6]. One query is executed for each point when the DBSCAN algorithm is used. The indexing structure is used while the neighborhood query is being executed. The overall complexity of O(nlogn) is obtained with the help of this method. At times when the accelerating index structure is not used in proper manner, the worst run time complexity arises which is represented as $O(n^2)$.

II. LITERATURE REVIEW

Vaibhav Kumar, et.al, (2016) presented in this paper [7] the k-means clustering based method in order to enhance the performance enhancement of cooperative spectrum sensing. The k-means clustering method is unsupervised learning method in which the generalized k-μ fading channels are utilized in this paper. In order to characterize the receiver operating characteristics, extensive simulation has been performed in this paper for various system parameters trade-offs. In comparison to the classical energy detection based CSS, the learning based method provides enhancement in the performance. On the basis of simulation results achieved it is determined that the proposed method provides best results amongst the other existing approaches. **R. Kumari, Sheetanshu, et.al, (2016)** studied in this paper [8], on detecting the intrusions using machine learning based k-means clustering approach. They have also utilized different data analytical techniques and used that experiment results in avoiding all those attacks. The cluster quality has been characterized using sillhoute coefficient that helps in determining the closeness of data points to one and different clusters. In the last after performing all the tasks authors have substituted a Gaussian mixture model or DBSCAN model along with k means iterative model. The financial data, behavior of customers and market crate analysis can be studied through identified abnormalities through clustering. **Kaustubh S. Chaturbhuj, et.al (2016)** presented in this paper [9], the utilization of K-means to discover better clusters and initial centriods are discovered using PSO. The large datasets are processed quickly and parallel using Hadoop as the big data is generated rapidly in huge amount which is not easy to be handling using traditional techniques of data mining. There are some issues that make traditional K-means clustering ineffective to be used in most of the situations. These issues can be resolved using parallel k-means clustering with Particle Swarm Optimization (PSO) by generating optimal clusters globally. The required time for performing has been reduced by much extent by handling large data using Hadoop and MapReduce. The scalability of method has been increased by much extent using Multi hubs for parallel processing. **Daniele Casagrande, et.al, (2012)** presented in this paper [10] that Hamiltonian system trajectories have been used for determining the level lines and its function is interpreted as integrated Hamiltonian systems. In discrete and continuous time cases, clustering can be characterized using dynamic clustering are utilized to exploit essential static algorithm. There is need to characterized different problem solutions that occurs due to different way of the two time scales. It is straightforward to augment a method into n-dimensional data points clustering that initially comprise the algorithm two dimensional versions iterative applications. Then finally the results of every iteration are intersected and portrayed different applications that's shows the effectiveness of method. **Manish Kumar Sharma, et.al, (2015)** proposed in this paper [11], an unconventional approach that helps in detecting fatigue in vehicular drivers. The loss in lives and vehicular accidents has been reduced to much extent using Oximetry Pulse (OP) signal that distinguish between cognitive fatigue of the driver. The fatigue conditions of driver has been determined using traditional k-means and modified k-means and implemented it. The datasets has been trained and tested using K-means classifiers and each extracted features has been treated as single decision making parameter. The results show that the proposed method is able to fetch classification accuracy of 100% using modified k means along with wavelet for feature extraction. **Aimin Yang, et.al, (2009)** introduced in this paper [12], a k-means clustering algorithm along with fuzzy classifier constructing method and divided into three phases. The kernel k-means clustering algorithm has been utilized to clusters some of the training samples in feature space. The proper membership function has been utilized to characterize a fuzzy rule for each made cluster point. The authors have also introduced a method for fuzzy classifier along with KKMC algorithm and a new classification rule has been made for each made cluster. They have also used Gas for modifying some parameters such as CK and ∂. The simulation results show that the proposed method requires very less training time and prove to be more accurate as compared to existing methods.

III. RESEARCH METHODOLOGY

In this work, further improvement in the incremental DBSCAN algorithm is done which calculates the Euclidian distance dynamically. The back propagation algorithm is one of the most utilized Neural Network algorithms. This method is used for training the artificial neural networks and also utilizes the two phase cycle which involves the propagation and weight updates. When an input network enters the network, it is propagated forward through the network across each layer until it reaches the output layer. The comparisons are made using the output achieved as well as the desired output. This is done utilizing a loss function. For every neuron in the output layer, an error value is calculated. The propagation of the error values is then done in backward manner which starts from the output.

Here, each neuron has its own error value which also shows its contribution to the originally achieved output. There are mainly four steps in which this algorithm can be executed. The required corrections are to be computed only once the weights of the network are selected randomly. The steps through which the algorithm is decomposed are: i) Feed-forward computation; ii) Back propagation to the output layer; iii) Back propagation to the hidden layer; iv) Weight updates. At the time when the values of error function become small, the algorithm is stopped. This is just an overview of the basic BP algorithm. However, various changes are proposed by researchers with time. The algorithm for back propagation is mentioned below:

Actual Output: $\sum_{\substack{x=0 \\ w=0}}^{\substack{x=n \\ w=n}} x_n w_n + bias$

$$Error = Desired\ Output - Actual\ Output$$

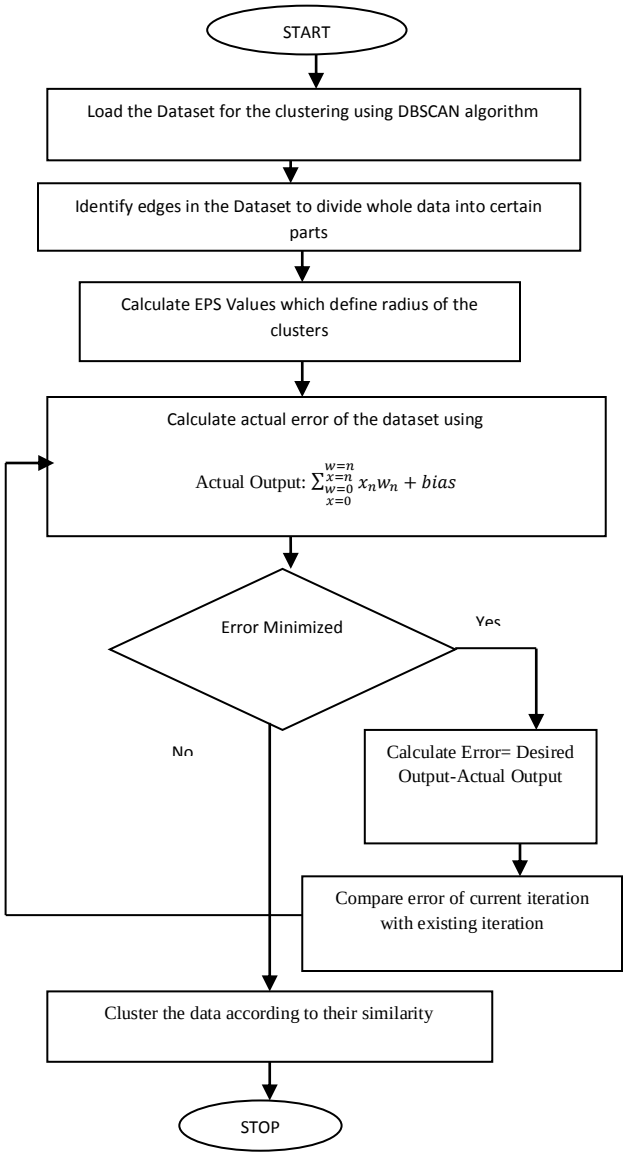

Fig. 1: Proposed Flowchart

IV. RESULTS AND DISCUSSION

The proposed and existing algorithms are implemented in MATLAB and results are analyzed in terms of accuracy and execution time

Parameters	Values
No of Attributes	24
No of instances	634
Missing values	YES
Primary attribute	Yes

Table 1: Dataset Description

As shown in table 1, the dataset description is given which is used in the proposed work

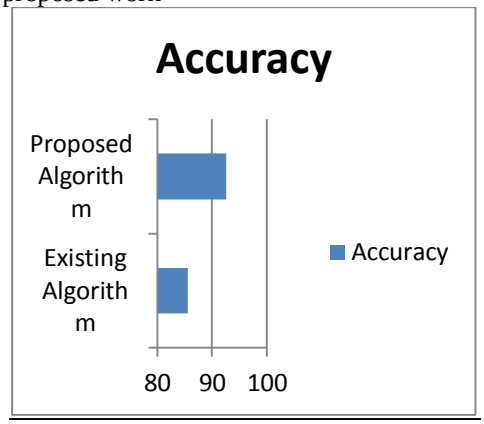

Fig. 2: Accuracy Comparison

As shown in figure 2, the accuracy of the proposed algorithm is compared with the existing algorithm. It has been analyzed that accuracy is increases at steady rate.

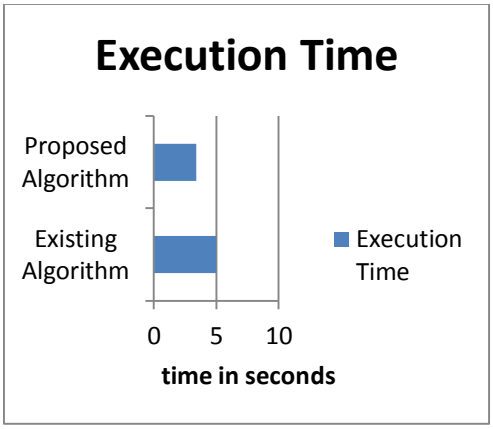

Fig. 3: Execution Time Comparison

As shown in figure 3, the execution time of proposed algorithm is compared with existing algorithm and it has been analyzed that execution time is reduced which dynamic Euclidian distance.

V. CONCLUSION

In this work, it has been concluded that clustering is the efficient approach which can group similar and dissimilar type of information. The DBSCAN is the density based clustering in which density of the data is calculated and region which has maximum dense will be clustered in one and other in the second. In this research, neural networks algorithm will be applied which will calculate Euclidian distance dynamically. The proposed improved leads to improve accuracy of clustering.

VI. REFERENCES

[1]. Karlina Khiyarin Nisa, Hari Agung Andrianto, Rahmah Mardhiyyah," Hotspot Clustering Using DBSCAN Algorithm and Shiny Web Framework", 2014, IEEE, 978-1-4799-8075-8

[2]. Negar Riazifar, Ehsan Saghapour," Retinal Vessel Segmentation Using System Fuzzy and DBSCAN Algorithm", 2015, IEEE, 978-1-4799-8445-9

[3]. Yumian Yang, Jianhua Jiang," Application of E-commerce Sites Evaluation based on Factor Analysis and Improved DBSCAN Algorithm", 2014, IEEE, 978-1-4799-6543-4

[4]. XiaoqingYu, Yupu Ding, Wanggen Wan, Etienne Thuillier," Explore Hot Spots of City Based on DBSCAN Algorithm", 2014, IEEE, 978-1-4799-3903-9

[5]. Marjia Sultana, Afrin Haider and Mohammad Shorif Uddin," Analysis of Data Mining Techniques for Heart Disease Prediction", 2016, IEEE

[6]. Kamaljit Kaur and Kuljit Kaur," Analyzing the Effect of Difficulty Level of a Course on Students Performance Prediction using Data Mining", 2015 1st International Conference on Next Generation Computing Technologies (NGCT)

[7]. Vaibhav Kumar, Deep Chandra Kandpal, Monika Jain," K-mean Clustering based Cooperative Spectrum Sensing in Generalized k-μ Fading Channels", 2016, IEEE, 978-1-5090-2361-5

[8]. R. Kumari, Sheetanshu, M. K. Singh, R. Jha, N.K. Singh," Anomaly Detection in Network Traffic using K-mean clustering", 2016, IEEE, 978-1-4799-8579-1

[9]. Kaustubh S. Chaturbhuj, Mrs. Gauri Chaudhary," Parallel Clustering of large data set on Hadoop using Data mining techniques", 2016, IEEE, 978-1-4673-9214-3

[10]. Daniele Casagrande, Mario Sassano, and Alessandro Astolfi," Hamiltonian-Based Clustering Algorithms for static and dynamic clustering in data mining and image processing", 2012, IEEE, IEEE CONTROL SYSTEMS MAGAZINE, 1066-033X

[11]. Manish Kumar Sharma, Mahesh M. Bundele," Design & Analysis of K-means Algorithm for Cognitive Fatigue Detection in Vehicular Driver using Oximetry Pulse Signal", 2015, IEEE International Conference on Computer, Communication and Control IC4

[12]. Aimin Yang, Qing Li, Xinguang Li," A Constructing Method of Fuzzy Classifier Using Kernel K-means Clustering Algorithm", 2009, IEEE, 978-0-7695-3888-4