Enhance DBSCAN Algorithm To Increase Accuracy Using PSO Algorithm

Anjali Soni[1], Mr. Gaurav Kumar Srivastav[2]
[1]M.Tech Student, [2]Assistant Professor
[12]Institute of Technology & Management Aligarh, India

Abstract- There are various data mining techniques which have been evolved from earlier times and are being still used for reliable and scalable tools which involve some older statistical methods. The division of data into groups of similar objects is known as clustering. This work is to improve the performance of incremental DBSCAN algorithm which is improved version of DBSCAN algorithm. In the incremental DBSCAN algorithm the EPS value is defined on the basis of input dataset and Euclidian distance is remain static which reduce accuracy of clustering. In this work, improvement in the incremental DBSCAN algorithm is done using the PSO algorithm which Euclidian distance will be calculated dynamically which leads to increase accuracy of clustering reduction in execution time. The PSO algorithm will take the every point and their value as input and calculate the error at each point for clustering. The point at which error is minimum is considered as the best point because at that point the accuracy of clustering is maximum. The efficient calculation of Euclidian distance also defines the accurate point for the clustering. The distance defined the similarity between the data points for the clustering.

Keywords- DBSCAN, PSO, Density Based Clustering

I. INTRODUCTION

For storing the required information there are number of devices present in today world of technologies such as computers and other mass digital storage devices. 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. 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 [1]. 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. In order to navigate the data warehouse in the case of OLAP (On-Line Analytical Processing) server an enhanced end-user business model is to be used. Very similar depiction is to be provided by the multidimensional structures as per the way in which the user wishes to see his business layout. A proper summary of the product line, region and the other necessary aspects is provided here. The data mining server is integrated with the data warehouse and OLAP server in order to embed the ROL-focused business analysis in the designing of data mining infrastructure. There are numerous issues which arise within the data mining [2]. The prospecting and promotion optimization are presented here which help in advancing the centric metadata template of the applications. With the integration amongst the data warehouses, the direct implementation and tracking of the operation decisions is made. On the basis of various future decisions present within the organization, best services are to be mined and further applied. On the basis of new decisions and results, the warehouse is evolved. 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 [3]. 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. The clustering mechanism does not have only one step it can be analyzed from the definition of clustering. There are certain partitions generated by the partitional clustering, for a given database of certain number of objects. A clustering criterion is followed by each cluster defined here. From the mean within every cluster the sum of squared distance is minimized. In this case there will be more chances that the grouping is exist in those clusters and it also try to find a optimum value which is global. Due to this the complexity will increase for these algorithms. Even for less number of present objects huge number of partition is taken place. The solution for this reason begins with an initial, basically random, partition. Further it proceeds by refining it in a proper manner [4]. For the purpose of executing it for practice, various sets of initial points are made to run by partitional algorithm. The hierarchical decomposition of objects is involved in the hierarchical abased clustering algorithms. There are two broader divisions of this type. One is the agglomerative (bottom-up) and the other is the divisive (topdown). 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. There are various unconstrained solutions which are used for real-world applications for customers. There are various problem-specific limitations within the clusters which make specific business actions [5]. There are various individual objects and parameter values on which the taxonomy of clustering constraints is involved. These values can be denoted through preprocessing or external cluster parameters. The k-means clustering algorithm is one of the measures which are used to solve the clustering problems. Out of number of unsupervised learning algorithms the k-means clustering algorithm is easiest. Certain numbers of clusters which include a fixed apriori are to be given classify provided data set using certain number of clusters which is simple and easy in this procedure [6]. For each present cluster the main objective is to introduce k centers. A very careful placement of these centers is required. As per the location variations number of results can be achieved. So, the centers are to be placed away from each other [7]. Then the every point associated to the given data set is selected and then it is related to the closet center. Partitioning, hierarchical, density, grid, model and constraint based clustering are the various types of clustering techniques present in data mining. The density based clustering algorithm is applicable on the density based parameters. The regions which are different from a thin region are formed as thick regions or areas. Until the density in the neighbors rises above certain threshold, the identified cluster is increased here. One good clustering based algorithm is DBSCAN (Density Based Spatial Clustering of Applications with Noise). With the help of arbitrary shaped clusters the noise can be separated from the large spatial databases by using this algorithm.

II. LITERATURE REVIEW

Ahmad M. Bakr, et.al (2015) proposed improvement in incremental DBSCAN algorithm is order to build and update the shaped clusters present in large sized datasets. The modification is required to be done in an incremented way for which this process is enhanced [8]. The complete dataset is divided into limited search spaces such that modifications can be made within it. This can further help in enhancing the overall performance of the system. In order to check the enhancements made within this method, it is compared with the already existing approaches. Within the comparison results achieved, it can be seen that the proposed method has been enhanced in terms of various aspects. The algorithm is applied within the various scenarios that are different in size and dimensions amongst numerous datasets. It is seen here that the speed of incremental clustering process is increased for up to 3.2 in comparison to the previous method.

Iyer Aurobind Venkatkumar, et.al (2016) studied that there are numerous techniques involved within the data mining process which include clustering, prediction, association and so on. A prediction technique is the one in which the predictions are set as per the present data [9]. It is not necessary that the prediction provided in correct for all the times. Also there is no guarantee to ensure that they are correct. According to the advantages and disadvantages of the respective algorithms, the data clustering algorithms are studied. The major four clustering algorithms which are k-means, BIRCH, DBSCAN and STING, the comparative studies are made which help in understanding their different characteristics.

Qi Xianting, et.al (2016) studied that the density-based clustering algorithm is the most popularly used algorithm for removal of noise in the applications. The clusters of different shapes and sizes can be recognized with the help of their different properties. The stability of the algorithm is no more left when there is a presence of high dimensional data within the respective applications. An improved DBSCAN algorithm is proposed in order to solve these arising issues, which is known as the feature selection based DBSCAN algorithm [10]. This algorithm is provided on various real world datasets and the various series of simulations are achieved. This helps in testing the performance of this newly proposed algorithm. The results depict that this new algorithm is more efficient as compared to the already existing ones. The high-dimensional data also is very easy to deal with through this algorithm proposed.

Kuan-Teng Liao, et.al (2016) proposed two algorithms namely centroid based clustering algorithms and UKmeans algorithm. With the help of former technique, the similarity of an application can be enhanced [11]. On the basis of the similarity factor, the time taken as well as the effectiveness of the system is affected. The similarity of calculations does not handle the time duration being spent. Its major focus is on the effectiveness of the clustering method. This helps in limiting the upper bound of the positions that are possible for the centroids. This helps in incrementing the effectiveness of clustering method. As per the simulation results it can be seen that there is an enhancement in the results achieved with the application of this method.

Wenbin Wu et.al (2016) proposed a novel approach in order to enhance the accuracy of forecasting and also to manage the training sample dynamics. Along with utilization of neural network, the k-means clustering algorithm is applied in mainly the scenarios which include small terms of WPF within them [12]. On the basis of the similarities amongst different methods that also use k-means clustering mechanism, there are various categorizations made. The information related to meteorological conditions and the other earlier present data is presented within this method. For resolving the over-fitting and instability problems involved in conventional networks, the integration of bagging-based ensemble approach is done into the back propagation neural network. There should be a proposed research related to the effective meteorological forecasting which will help in enhancing the forecasting accuracy**.** There should be a design proposed for the relative optimal method which involves the BDNN method.

Vadlana Baby, et.al (2016) proposed an effective distributed threshold privacy-preserving k-means clustering algorithm. In this method, the code based threshold secret sharing is utilized as a privacy-preserving method. There is a code based approach involved here which allows the division of data into various shares which is further processed at various servers [13]. There is less number of iterations in the newly proposed protocol as compared to the previous ones. There are certain comparisons made with respect to various techniques. The security analysis of this proposed method is also given here. The code based threshold secret sharing scheme along with secure addition and comparison protocols is utilized for privacy preserving k-means clustering algorithm. The clustering mechanism is performed in a collaborative manner and the third party's trust is avoided. A perfect preservation of the user data is provided through this newly proposed method.

III. RESEARCH METHODOLOGY

The density based clustering is the type of clustering in which clusters are formed based on the data density. The I-DBSCAN algorithm uses the two values which are EPS and Euclidian distance. The EPS value defines the radius of the cluster. In the previous research work, the EPS value defines the radius of the data statically. The PSO algorithm is defined in this work which calculates the EPS value dynamically. The PSO includes a dynamic definition of the objective function. Based on the value of swarm, comparisons are done against the current iteration and previous iteration. The swarm value that has the highest iteration is considered for identifying the objective function. The description of objective function which is dynamic is shown in the below equation. The value is changed once each iteration is executed.

$$v_{i+1} = v_i + c * rand * (p_{best} - x_i) + c * rand * (g_{best} - x_i)$$

Here, the velocity of element is represented by V_i, and the best value among available options is denoted by p_{best}. The random number is represented as rand. x is the value provided for each attribute of the website and c variable is used to define it. This process chooses the best value identified from every population and represents it as p_{best}. g_{best} is the best value that is chosen from every iteration. The value that is achieved is added with the traversing value of each attribute to finalize the objective function as shown in the equation below.

$$x_{i+1} = x_i + v_{i+1}$$

The position vector is denoted here by $x_{(i+1)}$. PSO algorithms that are dynamic with respect to the calculated best value are used to solve such multi-objective optimization problems. The data used for encryption is given as input by PSO. The key that is used for encryption helps in generating the optimized value.

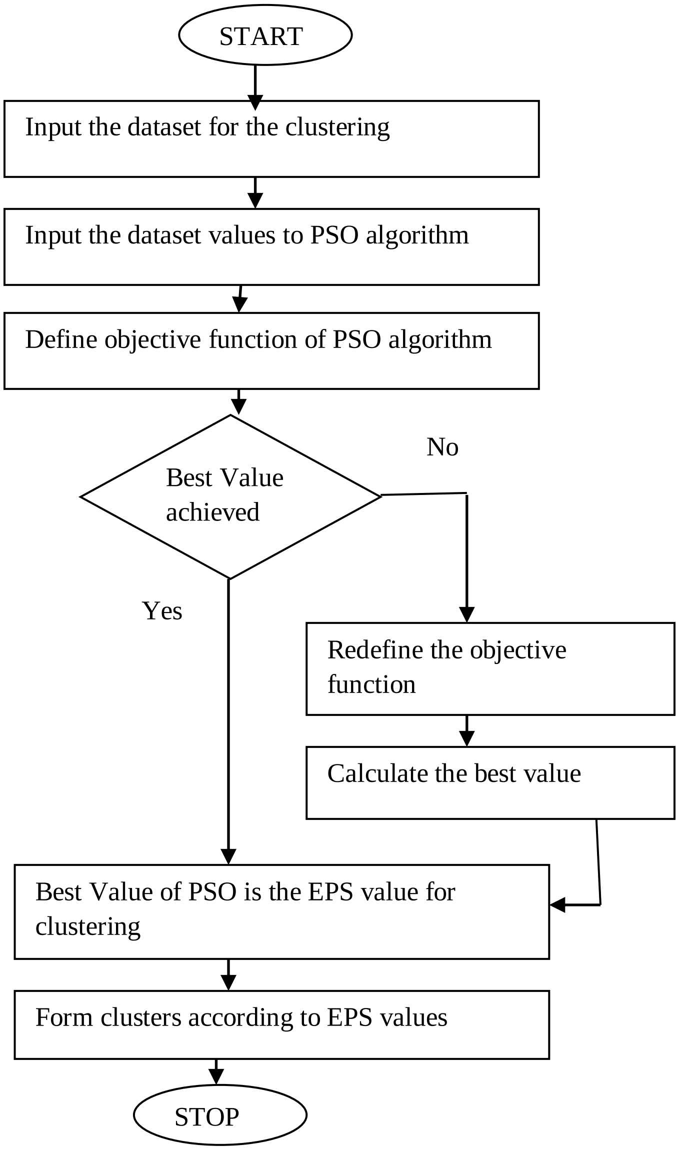
START
Input the dataset for the clustering
Input the dataset values to PSO algorithm
Define objective function of PSO algorithm
Best Value achieved
No
Yes
Redefine the objective function
Calculate the best value
Best Value of PSO is the EPS value for clustering
Form clusters according to EPS values
STOP

Fig.1: Proposed Algorithm

IV. EXPERIMENTAL RESULTS

The proposed approach is implemented in MATLAB and the results are evaluated by comparing performances of existing and proposed algorithms.

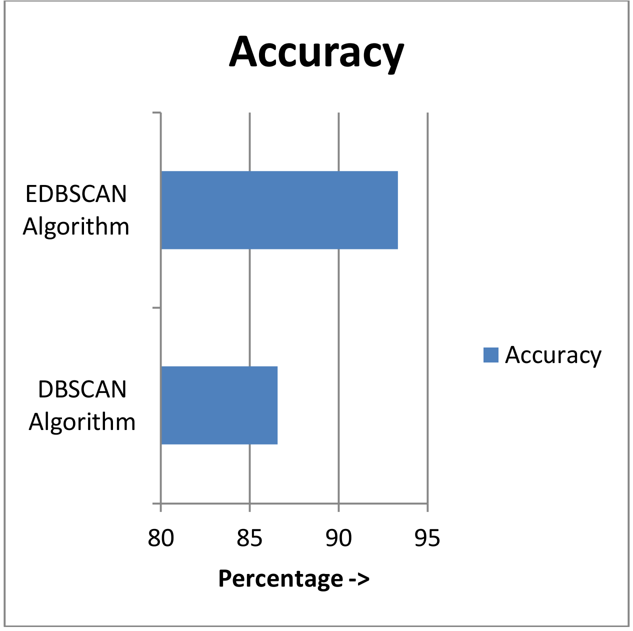

Fig.2: Accuracy of clustering

As shown in figure 2, the accuracy of proposed and existing algorithm is compared to check reliability of the algorithms and it is been analyzed that accuracy of proposed algorithm is more as compared to existing algorithm.

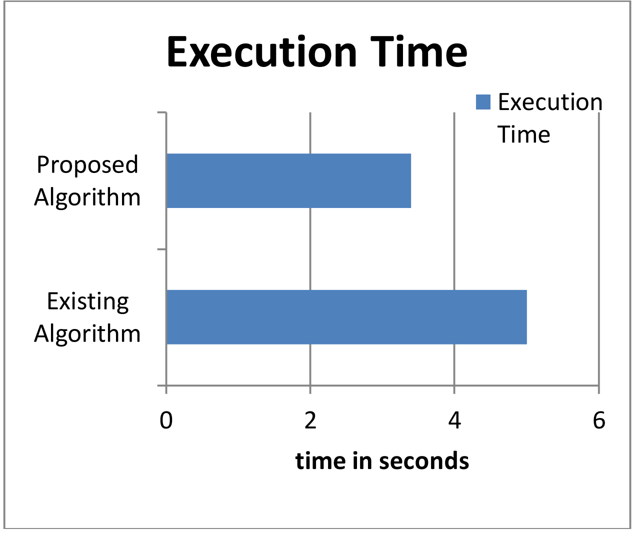

Fig.3: Execution time Comparison

As shown in figure 3, the execution time of proposed and existing algorithm is compared and it is been analyzed that due to dynamic calculation of Euclidian distance execution time is reduced in the DBSCAN algorithm. Execution time of an algorithm can be defined as run time that is utilized to achieve the main solutions to the main inputs and the run time required by the applications of data mining algorithms to the main input and main output, respectively. The time that is utilized by the running machines is known to be the run time. The sum of execution time of the inputs is calculated here along with the sums of outputs achieved within the corresponding technique. The comparisons are made amongst the two techniques on the basis of the execution times of both the separate techniques. The technique with less execution time is considered to be better.

V. CONCLUSION

The clustering is the technique in which similar and dissimilar type of data can be clustered together to analyze complex data. The technique of density based clustering is applied which can cluster the similar and dissimilar type of data according to the data density in the input dataset. In the density based clustering the densest region is calculated from which similar and dissimilar type of data is calculated using similarity technique. In the DBSCAN algorithm which is applied in this work, the EPS value is calculated which will be the central of the dataset. The EPS value is calculated dynamically to achieve maximum accuracy. The technique of Euclidian distance is applied to calculate similarity between the data points. To increase accuracy of clustering PSO technique will be applied in future which calculate Euclidian distance in dynamic manner.

VI. REFERENCES

[1]. C. Bahm, K. Haegler, N.S Maller, C. Plant," CoCo: coding cost for parameter-free outlier detection", 2009, 15th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 149–158

[2]. D. Wang, S. Zhu, T. Li, Y. Chi, Y. Gong," Integrating clustering and multi document summarization to improve document understanding," 2008, 17th ACM CIKM Conference on Information and Knowledge Management

[3]. H.-P. Kriegel, M. Pfeifle," Effective and efficient distributed model-based clustering," 2005, 5th International Conference on Data Mining (ICDM'05), pp. 285, 265

[4]. K.M. Hammouda, M.S. Kamel," Efficient phrase-based document indexing for web document clustering," 2004, IEEE TransKnowledge and Data Eng., vol. 16, no. 10, pp. 1279–1296

[5]. V. Gayathri, M.Chanda Mona and S.Banu Chitra, "A survey of data mining techniques on medical diagnosis and research," 2014, International Journal of Data Engineering (OOE) Singapore Journal of Scientific Research (SJSR), vol. 6, pp. 301-310

[6]. M.Akhil Jabbar, Priti Chandra and B.L Deekshatulu, "Heart disease prediction system using associative classification and genetic algorithm," 2012, International Conference on Emerging Trends in Electrical, Electronics and Communication Technologies-ICECIT

[7]. R. Chitra and V.Seenivasagam, "Review of heart disease prediction system using data mining and hybrid intelligent," 2013, ICTACT Journal on Soft Computing, vol. 03, no. 04

[8]. Ahmad M. Bakr, Nagia M. Ghanem, Mohamed A. Ismail," Efficient incremental density-based algorithm for clustering large datasets", 2015, Elsevier B.V.

[9]. Iyer Aurobind Venkatkumar, Sanatkumar Jayantibhai, Kondhol Shardaben," Comparative study of Data Mining Clustering algorithms", 2016, IEEE

[10]. Qi Xianting, Wang Pan," A density-based clustering algorithm for high-dimensional data with feature selection", 2016, IEEE

[11]. Kuan-Teng Liao, Chuan-Ming Liu," An Effective Clustering Mechanism for Uncertain Data Mining Using Centroid Boundary in UKmeans", 2016, IEEE

[12]. Wenbin Wu and Mugen Peng," A Data Mining Approach Combining K-Means Clustering with Bagging Neural Network for Short-term Wind Power Forecasting", 2016, IEEE

[13]. Vadlana Baby, Dr. N. Subhash Chandra," Distributed threshold k-means clustering for privacy preserving data mining", 2016, IEEE