Density Based Clustering Method Using Bio-Inspired Method

Mr. Malik Tajamul Hussain
Research Student, Dept. of CSE, Swami Vivekananda Institute of Engineering and Technology - Affiliated with Maharaja Ranjit Singh Punjab Technical University, Punjab, India.

Abstract: Since data mining is applied in many applications, improvements are being made with time. Clustering is known as the method using which the data is classified into groups based on its similarity patterns. The type of clustering that defines the clusters based on the density of data is known as density based clustering. The density based algorithms help in discovering the arbitrary shaped clusters which results in providing an authentication against the outliers entering the systems. Initiating from the output, backward propagation of error values is calculated. Improving the performance of incremental DBSCAN is the objective of this research. the Euclidian distance is calculated in dynamic manner and due to which the execution time of clustering is reduced thus, increasing its accuracy. Each point and their value will be taken by the PSO as input and at every clustering point the error will be calculated. The specific point at which accuracy of clustering is the highest is considered as the best point since at this point, the error is the least. The accurate point for clustering is defined by the efficient calculation of Euclidian distance. The similarity among data points for clustering is defined by distance.

IndexTerms **- DBSCAN, PSO, Accuracy, EPS.**

I. INTRODUCTION

Nowadays, the information is the perfect source to power and success in several applications. It is possible to collect the information from different sources. The people use large amount of information generated from various sources for their own profits. The information is stored such that it can be used in future as per the needs. In today's technology based world several devices like the huge digital storage devices and computers have been designed to store the required information. There are several devices available for storing the variety of data being generated [1]. A structured database is designed for avoiding chaos. To achieve the objective, a Database Management System (DBMS) has been designed through which the data can be efficiently arranged. By using DBMS, it is possible to efficiently retrieve the data as per the requirements of users. Data mining is also known commonly as KDD (knowledge Discovery in Database). KDD is used to extract the new and potentially useful information. This extraction is a nontrivial extraction of the implicit data. KDD and data mining can be used as synonyms since they both mean the same. It is important to follow certain steps to discover the knowledge from databases. Collecting raw material once it has been identified is the initial step for generating new information. The method with which the data is categorized among similar object groups is known as clustering. During the involvement of less numbers of clusters, simplification level is achieved. Few finer details are lost however, due to the presence of less numbers of clusters. Data modeling is performed using the clusters [2]. For a given database of particular number of objects the partition based clustering generates few partitions. Each of the defined clusters follows a clustering criterion. It is possible to reduce the sum of squared distance from the mean of every cluster. The possibility of a group to exist in those clusters is higher in this case. So, an optimum global value is identified. For these algorithms, the complexity increases. Several partitions are generated even when less numbers of objects are available. In an initial but random partition, the solution for reason is initiated. These algorithms are responsible for the hierarchical decomposition of objects [3]. Depending on the density objective functions, it is possible to group the objects by applying density-based clustering. The total number of objects existing in the neighbor of data object defines the density of a specific object. Based on the increase in number of objects of certain parameters, a given cluster is expanded. For fixed number of clusters, there is variation in the methodology applied in partition-based algorithms as compared to the algorithms applied in density-based conditions. It is possible to divide the Euclidean space of an open set in sets of connected components. The connectivity, boundary and density of partitioning are required for finite sets of points in a cluster. There is a close relation among the nearest neighbor of a point. The dense component that moves towards a direction leading to the density is responsible for the growth of a cluster. It is possible to discover the arbitrary shapes of clusters by applying density-based algorithms [4]. The data is searched for better approximations of model parameters using this algorithm. For refining the model such that the partitioning can be recognized as hierarchical or partitioning based, the data set and method's structure or models can be proposed. Due to their closeness towards density-based algorithms the clusters are designed such that the preconceived model can be improved. For solving the clustering related issues, one of the measures to be taken is the k-means clustering algorithm. It is the easiest way to apply k-means clustering algorithm in comparison to the other unsupervised algorithms. For the classification of provided dataset with the help of certain number of clusters, a fixed apriori is provided. Introducing k-centers is the major objective for every cluster. It is important to place these centers very carefully [5]. The results are achieved based on the variations in location. Thus, the placement of centers is done at a distance. Further, the selection of given dataset is done at each associated point and then the closest center is presented in relation to it. In the absence of any point, the initial step is completed. The clusters processed are achieved at the final step to calculate the new centroids. The similar data set points and the closest new centers are bound together after achieving the new centroids. A loop is generated to provide results. The location of k-center is modified due to this loop and as per the outcomes of each step. Till no more changes are left, the process keeps running [6]. The process stops when there is no possibility of moving to the center. To analyze clustering, a simple learning algorithm known as k-means algorithm is used. The major objective here is to identify the best division of n entities in k groups. Different kinds of clustering techniques are presented in data mining among which few are grid, model, density, hierarchical and partitioning based clustering [7]. On the density based parameters, the

density based clustering algorithm is applied. Thick regions or areas are generated which are different from the thin regions. The identified cluster size is increased until the density of neighbors is higher than certain threshold value. DBSCAN (Density Based Spatial Clustering of Applications with Noise) is an efficient clustering based algorithm that is based on density. This algorithm helps in separating the noise from large spatial databases by applying arbitrary shaped clusters.

Literature Review

KM Archana Patel et.al (2016) recognized clustering as one of the most authentically used unsupervised learning technique of data mining. On the basis of similarity, the similar data objects were placed within similar clusters. The clustering algorithms were divided into several different categories. A variety of data mining clustering algorithms were learned and correlated in this work [8]. A discussion was made on different clustering algorithms in this work. These algorithms were compared on the basis of different factors. Some specifications were enlisted after comparing these algorithms. These specifications gave description of different algorithm advantageous in some environments. The clustering algorithms provided improved outcomes.

ZHANG Ke, et.al (2016) used density based clustering for the classification of large volume of data. In this work, density-based clustering algorithm was employed. The major objective of this work was to compute the distance of adaptive dense areas. A new density-based clustering algorithm was presented in this work after determining density threshold [9]. This process included the conversion of a density threshold selection problem into a fortitude issue associated with sperality radius. The radius threshold was identified as partial cluster. The outcomes of clustering were provided according to the available datasets. A discussion was also made in this work to decide the future tasks. The proposed approach gave superior performance in comparison to earlier approaches. The distance computations were utilized for improving the actual clustering algorithms.

Guangchun Luo, et.al (2016) proposed Spark based parallel DBSCAN algorithm in this work. This algorithm was called S_DBSCAN [10]. The proposed algorithm could divide the real data effortlessly. This algorithm combined different clustering results. In this work, a comparison of proposed algorithm and other existing algorithms was performed as well. The degree of obtained improvement was measured according to the comparative results. The tested results depicted that the proposed algorithm performed superiorly terms of different metrics than the earlier algorithms. These metrics included effectiveness and scalability. The created clusters were competent and the noise data available in the system was identified as well. The proposed algorithm performed better as compared to various other existing approaches.

Dianwei Han, et.al (2016) proposed a density based competent clustering algorithm called DBSCAN. This proposed algorithm helped to identify the random shaped clusters available in the data. In addition, the proposed algorithm was a well-organized algorithm. This algorithm was utilized for eliminating the noise occurring within the data. A testing task was presented as per the MPI or OpenMP scenarios. The involvement of different factors made it a main concern [11]. DBSCAN algorithm found its applicability in different areas due to its nose handling efficiency and the presence of arbitrary shaped clusters. The performance of this algorithm could be scaled. This algorithm maintained distance from the information sharing systems. The tested results demonstrated that the proposed algorithm was scalable and the tasks implemented on MapReduce provided effectual outcomes.

Nagaraju S, et.al (2016) presented a novel clustering algorithm for identifying the embedded and nested neighboring clusters. The main objective of this study was to partition the clusters. In the DBSCAN algorithm, the nested neighboring clusters were detected. In this work, the proposed modified clustering algorithm and other existing algorithms were compared [12]. The comprehensive density metrics were also assessed in this work using sorted k-distance plot and the first order derivative. The proposed algorithm was used to determine test results. The nested neighboring clusters were detected in this algorithm. The comparisons were based on the complications involved in computation task.

Jianbing Shen, et.al (2016) proposed a novel image superpixel segmentation algorithm based on DBSCAN clustering algorithm. A standard shaped superpixel was produced in this work. This phenomenon utilized DBSCAN superpixel segmentation technique [13]. The application of public Berkeley Segmentation database was used for evaluation purpose. In this work, different algorithms were compared on the basis of their performances. The comparative results depicted the superiority of proposed algorithm terms of different metrics. The proposed algorithm reduced computational costs. These costs were quite high in different existing algorithms. Moreover, the complications involved in the computation task were also reduced to a huge extent. These modifications made the utilization of this algorithm easy in different areas.

Research Methodology

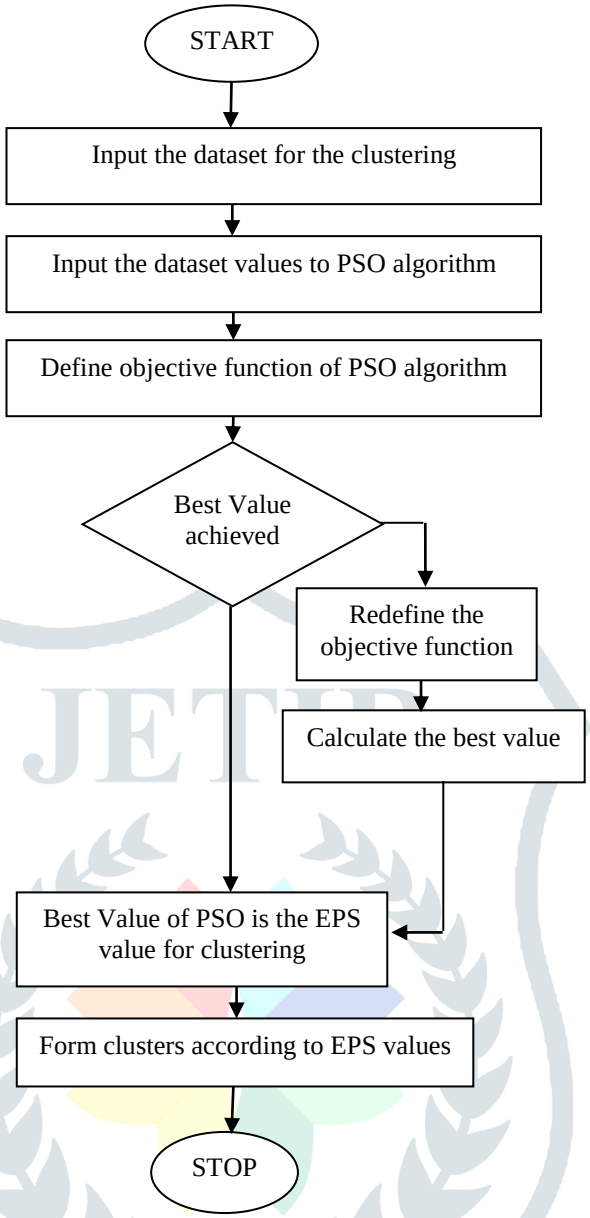
The data density is used to create clusters in density based clustering approach. The two values called EPS and Euclidian distance are used by the I-DBSCAN algorithm. The cluster radius is defined by the EPS value. The radius of the data is defined by EPS value in the earlier study in static manner. In this study, the PSO algorithm is used. This algorithm measures EPS value in dynamic manner. The objective function is defined dynamically in the PSO algorithm. The present iteration and earlier iterations are compared on the basis of swarm value. The objective function is identified using the swarm value having maximum iteration. The following expression describes the dynamic objective function. Execution of every iteration changes the value.

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

In the above equation, V_i represents the element velocity. The variable p_{best} represents the optimum value among accessible options. The variable "randx" represents random number. This is the value given to every feature of the website. The "c" variable defines this value. This procedure selects the optimum value recognized from overall population and demonstrates it as p_{best}. The best value selected after each iteration is represented by "g_{best}". The obtained value is added with the traverse value of every attribute for concluding the objective function. This phenomenon is given as:

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

The "$x_{(i+1)}$" denotes position vector. These multi-objective optimization issues are solved by using dynamic PSO algorithms regarding the best computed value. The PSO algorithm gives the data utilized for encryption as input. The key utilized for encoding provides support to generate enhanced value.



Experimental Results

The proposed research is implemented in MATLAB and the results are evaluated by comparing proposed and existing techniques in terms of various parameters.

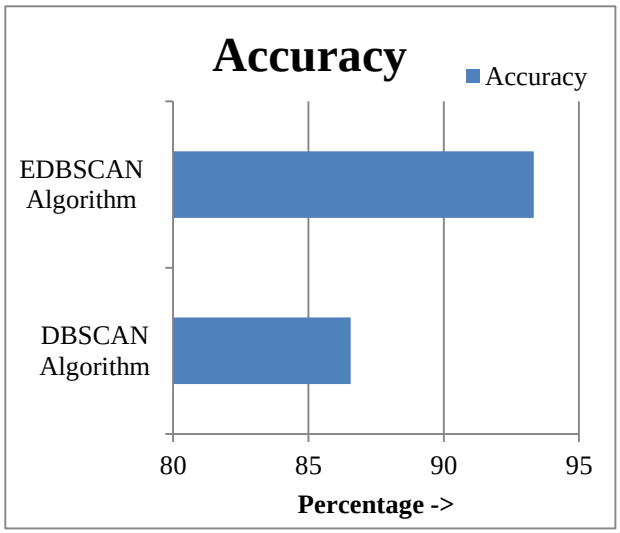


Fig 6.7: Accuracy of clustering

Figure6.7 shows the comparison of proposed and existing algorithm in terms of accuracy. The consistency of the algorithms is validated through the comparison. The comparative results prove the supremacy of proposed algorithm than accessible algorithm.

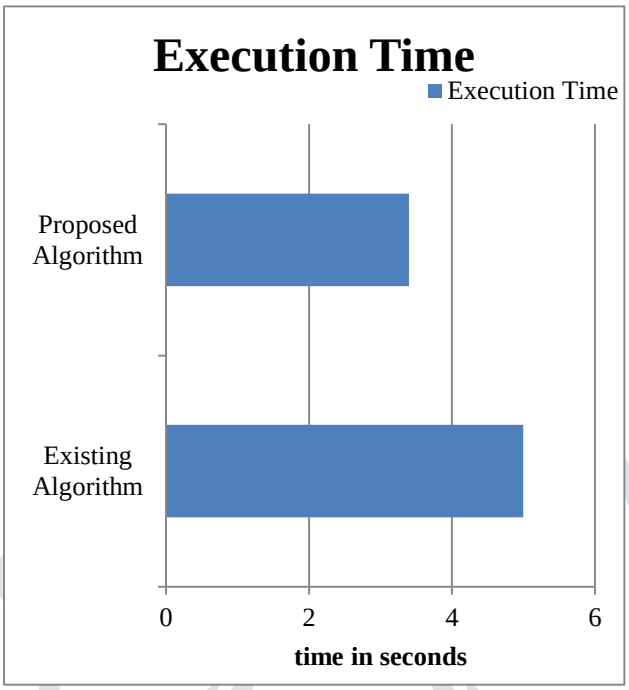


Fig 6.8: Execution time Comparison

Figure 6.8shows the comparison of proposed and existing algorithm in terms of execution time. In DBSCAN algorithm, the dynamic computation of Euclidian distance decreases the execution time as per the analysis. The execution time of an algorithm is defined as run time. The execution time is used to get major answers to the main inputs and the run time needed by the applications of data mining algorithms to the main input and main output, correspondingly. The run time is defined as the time used by running machines. Here, the sum of execution time of the inputs is computed with the sums of outputs obtained within the equivalent methodology. The two methodologies are compared on the basis of their execution time. The methodology having less execution time is regarded as superior.

Conclusion

The technique used to cluster similar and dissimilar type of data for analyzing complex data is called clustering or cluster analysis. In this work, the density based clustering algorithm is implemented to cluster similar and dissimilar data on the basis of density of data in the input dataset. The density based clustering algorithm calculates densest area. The similarity method is used to compute similar and dissimilar data from this region. The EPS value is computed by implementing DBSCAN algorithm. The EPS value will be the center of dataset. In order to obtain maximal accuracy, the EPS value is computed in dynamic manner. The similarity amid the data points is computed by applying Euclidian distance method. In future, PSO algorithm will be implemented for increasing clustering accuracy. This algorithm will compute Euclidian distance dynamically.

References

[1] D. Widyantoro, T. Ioerger, J. Yen," An incremental approach tobuilding a cluster hierarchy", 2002, ICDM Proceedings IEEEInternational Conference on DataMining,pp. 705–708

[2] S.A.L. Mary, K.R.S. Kumar," A density based dynamic dataclustering algorithm based on incremental dataset", 2012, J. ComputerSci. 8 (5) 656–664

[3] K.M. Hammouda, M.S. Kamel," Incremental documentclustering using cluster similarity histograms", 2003, IEEE/WICProceedings International Conference on Web Intelligence, pp. 597–601

[4] M. Ester, H.P. Kriegel, J. Sander, M. Wimmer, X. Xu," Incremental clustering for mining in a data warehousingenvironment",1998, Proceedings of the 24th VLDB Conference,Institute for Computer Science, University of Munich,Germany, New York, USA

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

[6] KarlinaKhiyarinNisa, Hari AgungAndrianto, RahmahMardhiyyah," Hotspot Clustering Using DBSCAN Algorithm and Shiny Web Framework", 2014, IEEE, 978-1-4799-8075-8

[7] Negar Riazifar, Ehsan Saghapour," Retinal Vessel Segmentation Using System Fuzzy and DBSCAN Algorithm", 2015, IEEE, 978-1-4799-8445-9
[8] KM Archana Patel and PrateekThakral," The Best Clustering Algorithms in Data Mining", 2016, IEEE
[9] ZHANG Ke, HUANG Lei, CHAI Yi," An Algorithm to Adaptive Determination of Density Threshold for Density-based Clustering", 2016, IEEE
[10] Guangchun Luo, Xiaoyu Luo, Thomas Fairley Gooch, Ling Tian, Ke Qin," A Parallel DBSCAN Algorithm Based On Spark", 2016, IEEE, 978-1-5090-3936-4
[11] Dianwei Han, Ankit Agrawal, Wei−keng Liao, AlokChoudhary," A novel scalable DBSCAN algorithm with Spark", 2016, IEEE, 97879-897-99-4
[12] NagarajuS,ManishKashyap, Mahua Bhattacharya," A Variant of DBSCAN Algorithm to Find Embedded and Nested Adjacent Clusters", 2016, IEEE, 978-1-4673-9197-9
[13] Jianbing Shen, XiaopengHao, Zhiyuan Liang, Yu Liu, Wenguan Wang, and Ling Shao," Real-time Superpixel Segmentation by DBSCAN Clustering Algorithm", 2016, IEEE, 1057-7149