iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** V **Month of publication:** May 2022

DOI: https://doi.org/10.22214/ijraset.2022.43067

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue V May 2022- Available at www.ijraset.com

Techniques to Enhance the Performance of DBSCAN Clustering Algorithm in Data Mining

Shyam Lal[1], Dr. Vaishali Singh[2]
[1, 2]*Maharishi University of Information Technology, Lucknow*

Abstract: ***Clustering is a form of learning by observations. It is an unsupervised learning method and does not require training data set to generate a model. Clustering can lead to the discovery of previously unknown groups within the data. It is a common method of data mining in which similar and dissimilar type of data would be clustered into different clusters for better analysis of the data. In this paper the DBSCAN algorithm has been applied to compute the EPS value and Euclidian distance on the basis of similarity or dissimilarity of the input data. Also back propagation algorithm is applied to calculate Euclidian distance dynamically and simulation study is conducted that shows improvement to increase accuracy and reduce execution time.***
Keywords: Clustering; DBSCAN; Back-propagation; Accuracy; Execution time

I. INTRODUCTION

Data mining is viewed as a result of the natural evolution of information technology. The early development of data collection and database creation mechanism proved to be important for the later development of effective mechanisms for data storage and retrieval, query and transaction processing [1,2,3]. The information and knowledge management business evolved within the development of many crucial functionalities: knowledge assortment and information creation, knowledge management and advanced data analysis. One of the emerging data repository architecture is the data warehouse. It involves multiple heterogeneous information sources organized underneath a unified schema at one website to manage decision making. Cluster analysis is widely used in various applications, together with research, pattern recognition, information analysis, and image process. In business, bunch will facilitate marketers discover interests of their customers supported getting patterns and characterize teams of the purchasers [4][5]. In biology, it is often accustomed derive plant and animal taxonomies, categories genes with similar practicality, and gain insight into structures inherent in populations. Information bunch(or simply clustering),is associated unattended classification methodology. This methodology aims at making teams of objects or clusters, in such the simplest way that objects within the same objects in distinct normal topics within the data processing field. Most partitioning ways cluster objects supported distance between objects. Spherical formed clusters are often discovered by these ways and encounter bother in discovering clusters of discretional shapes. Thus for discretional shape cluster square measure terribly similar and numerous clusters square measure quite [6][7]. Cluster analysis is one among the new way square measure utilized referred to as density- based ways that square measure supported the notion of density. In these methods the cluster is continued to develop as long as the density in the area exceeds some threshold [8][9][10]. The fundamental thought is to bear on the growing the given cluster as long as the density in the area exceeds some threshold i.e. for every datum within a given cluster, the radius of a given cluster has to contain no but a minimum variety of points. It discovers discretionary form clusters. It likewise handles clamor within the data [11][12]. It is just the once scan. It needs density parameters additionally [14]. The rest of paper is organized as follows: Section-II Present a brief summary of Related Work. Section-III describes the DBSCAN Algorithm. Section-IV Covers Proposed Methodology. Section-V describes the Experimental Evaluation. Section-VI Concludes the paper. While references are mentioned in the last.

II. PRIOR WORK DONE

This section provides a brief summary of various research work carried out by many researchers in this domain based on past research articles. Spieth et al. [21] applied DBSCAN to spot solutions for the illation of restrictive networks. Wen et al. 20] adopted DBSCAN and progressive DBSCAN because the core algorithms of their question cluster tool. They used DBSCAN to cluster commonly asked queries and Guangchun Luo, et.al, proposed system of cluster analysis occupies a pivotal position in data mining, and the DBSCAN algorithm is a standout amongst the most broadly utilized algorithms for clustering. Nonetheless, when the existing parallel DBSCAN algorithms create information partitions, the first information is partitions; with the cacophonous and consolidation of high-dimensional area can consume lots of your time [14]

To unravel the matter, this paper proposes a parallel often divided into many disjoint rise in information dimension, DBSCAN rule (S_DBSCAN) supported Spark, which can quickly realize the partition of the original data and the mix of the clustering results. It is divided into the subsequent strides: 1) Partitioning the data supported a DBSCAN Merging the random sample 2) Computing native algorithms in parallel, 3) information partitions supported the center of mass. DBSCAN is an program that is density and might determine discretionary formed clusters and Dianwei Han, et.al, analyzed that impressive bunch algorithmic predicated on density and might determine discretionary formed clusters and eliminate noise data. Be that because it might, parallelization of DBSCAN could be a testing work on the grounds that supported MPI or Open MP environments, there exist the problems of lack of fault tolerance and there is no guarantee that employment is balanced. Also, programming with MPI needs information scientists to own a complicated expertise to handle communication between nodes that could be a huge challenge [15]. DBSCAN algorithmic rule has been very illustrious since it will establish whimsical formed clusters and in addition handle clamorous information Nagaraju S. et.al; introduce associate degree economical approach for agglomeration analysis to find embedded and nested adjacent clusters utilizing plan of density based mostly notion of clusters and neighbourhood distinction. the planned formula is improved version basic DBSCAN formula, planned to handle the clump downside with the employment world density parameters in basic DBSCAN formula and downside of police work nested adjacent clusters in Basically EnDBSCAN algorithm. The experimental results that suggested that proposed algorithm is more effective in detecting embedded and nested adjacent clusters compared both DBSCAN and EnDBSCAN without adding any additional computational complexity [16].Jian bing Shen, et. al; proposes a real-time picture super pixel segmentation method with 50 fps by utilizing the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm. In order to decrease the process prices of super pixel strategy received a fast two-stage the principal agglomeration stage, algorithms, the framework within the DBSCAN formula with color similarity and geometric confinements is employed to quickly cluster the pixels, and later little clusters square measure incorporate into super pixels by their neighbourhood through a distance measuring outlined by color and spatial options within the second merging stage [17]. A robust and easy distance perform is outlined for recovering super pixels in these two stages Ilias K. Savvas, et. al; propose standout amongst the foremost fascinating and productive techniques, so as to find and extract data from knowledge storehouses is cluster and DBSCAN would be a successful density based mostly algorithmic rule that clusters knowledge accordant its characteristics [18]. Be that because it could, its element aryburden is its severe process quality that proves the technique exceptionally inadequate to use on huge datasets. Despite the very fact that DBSCAN is associated in nursing exceptionally noticeably studied technique, a very operational parallel version of it, has not been accepted nevertheless by thought researchers. During this work, a three section parallel version of DBSCAN is presented. The obtained experimental results are exceptionally promising and demonstrate the correctness, the measurability, and therefore the effectiveness of the technique. Ahmad M. Bakr, et.al, the planned rule enhances the progressive agglomeration method by limiting the search house to partitions complete dataset which up in important enhancements within compared as opposition the ends the performance to relevant progressive agglomeration algorithms [19]. Experimental results with datasets of assorted sizes and dimensions demonstrate that the planned rule hastens the progressive agglomeration method by issue up to 3.2 compared to existing progressive algorithms. The rule incrementally partitions the dataset to scale back the search house to each partition as opposition filtering the complete dataset. Subsequently the rule incrementally forms and updates dense regions in each partition. Following distinctive potential dense regions in each partition, the rule utilizes associate inter- connectivity live to merge dense regions to form the ultimate range of clusters.

III. METHODOLOGY

In the DBSCAN rule the foremost dense region is calculated from the dataset. The central purpose is calculated from the foremost dense region that is referred to as EPS worth of the dataset. To calculate similarity between the info purposes of the info geometer distance is calculated from central point to any or all different points. The whether that is similar is clustered in one dataset and different are within the second dataset. In the existing work, to enhance accuracy of bunch EPS values is calculated within the dynamic manner that results in the bunch of points which are remained un-clustered. To achieve additional accuracy of cluster technique of back propagation are going to be applied that calculate geometer distance within the dynamic manner and it increases accuracy and scale back execution time of improved DBSCAN algorithm.

A. Dbscan Algorithm

Density based clustering algorithms have a wide applicability in data mining. They apply local criterion to cluster objects: clusters square measure viewed as regions within the information house wherever the objects are dense and that square measure separated by regions of low object density (noise) [12]

DBSCAN is exceptionally well known due both to its low complexity and its capacity to detect clusters of any shape, which is a desired characteristics when one doesn't have any knowledge of the possible clusters' shapes, or when the objects are circulated heterogeneously, for example, along paths of a graph or a road network. In any case, to drive the methos, this algorithmic rule desires two numeric input parameters, minPts and that along characterise the specified density characteristics of the generated clusters. especially, minPts could be a positive whole (number|number} deciding the minimum number of objects that has got to exist within a most distance of the information area all at once for associate in Nursing object to own an area with a cluster.

Since DBSCAN is extremely sensible to the setting of these input parameters they should be picked with incredible accuracy by considering both the scale of the dataset and the closeness of the objects all together not to affect an excessive amount of both the speed of the algorithm and the effectiveness of the outcomes. To settle the right values of these parameters one by and large engages an exploration phase of trials and errors in which the clustering is run several times with distinct values of the parameters. The DBSCAN algorithm can identify clusters in extensive spatial data sets by taking a gander at the local density of database components, utilizing one and only input parameter. Besides, it the client gets a proposal on which parameter value that would be reasonable. Along these lines, minimal knowledge of the domain is required. The DBSCAN can likewise figure out what data ought to be classified as noise or outliers. Regardless of this, it's working process is quick and scales extremely well with the extent of the database linearly. By utilizing the density distribution of nodes within the info, DBSCAN will categories these nodes into separate clusters that characterize the various categories. DBSCAN will discover clusters of arbitrary form. In any case, clusters that lie near one another have a bent to own an area with an identical category.

B. Proposed Algorithm

1) *Input:* Dataset for clustering, desired and output patterns
2) *Output:* Clustering of input data
   a) *M ← List of objects which will amendment their centroids*
   b) *D ← Most dense regions within the dataset*
   c) *For each purpose p(i) in P do*
   d) *C ← nearest centroid ()*
   e) *Function nearest centroid ()*
   f) *initpopulation P*
   g) *evaluate P ;*
   h) *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)
   i) *C ← P;*
   j) *M ← update centroid*
   k) *End*
   l) *For each r to M*
   m) *For each ri in M do*
   n) *c <-ri new_centroid*
   o) *Co<-ri old_centroid*
   p) *Apply incDbscanDel to remove ri from co*
   q) *Apply incDbscanAdd to insert ri to cn*
   r) *Add updated dense regions to D*
   s) *end for*
   t) *For each di in D do*
   u) *For each dj in D and i–j do*

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue V May 2022- Available at www.ijraset.com

v) *If inter_connectivty (di,dj) > a merge*
w) *merge(di,dj)*
x) *end if*
y) *end for*
z) *end for*

The complete flow of our proposed work is shown in fig.1. The proposed enhancement of the algorithm that is accomplished in the DBSCAN algorithm to improve the accuracy of clustering algorithm is referred in fig.1. In existing DBSCAN algorithm EPS price is calculated dynamically statically that improvement dynamically attributable propagation algorithm and geometer distance is calculated cut back potency of the algo. This relies on to calculate geometer distance to that technique back has applied that outline the geometer distance within the unvarying manner and distance at that error is minimum is that the final geometer distance, once the ultimate geometrician distance is taken into account similar and dissimilar sort of knowledge is clustered for complete analysis.

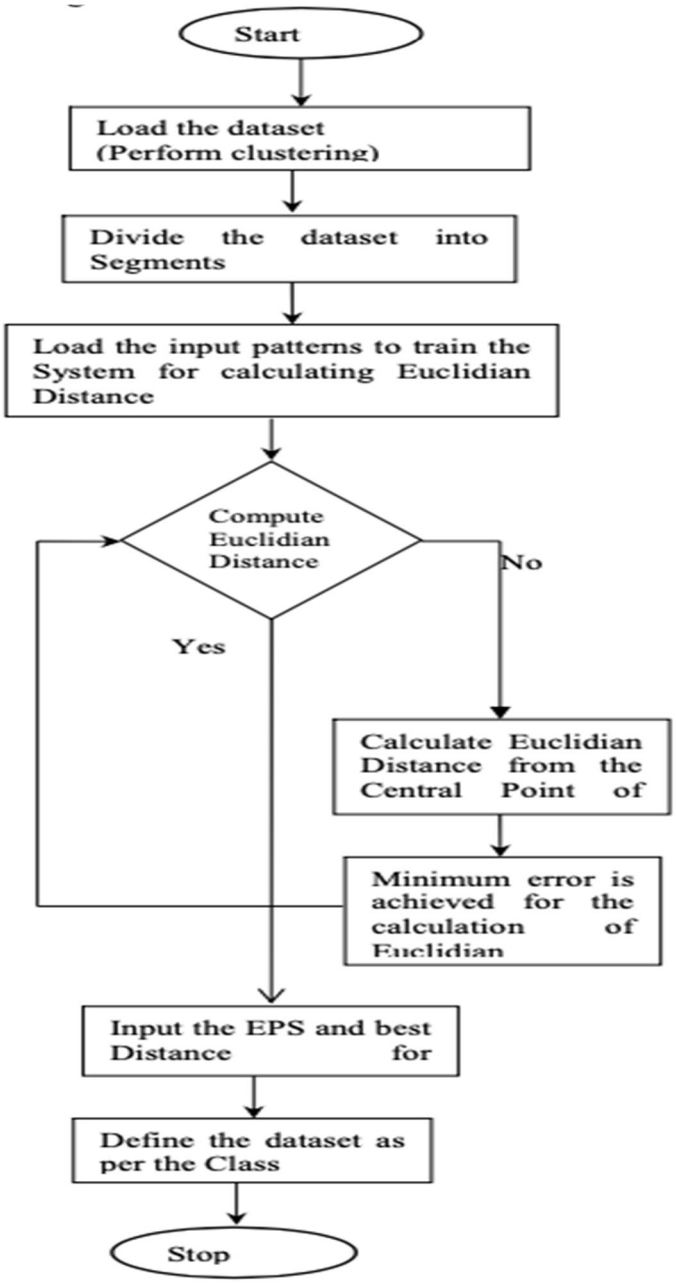


Fig.1. Flow of Proposed Algorithm

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue V May 2022- Available at www.ijraset.com

IV. EXPERIMENTAL STUDY

The proposed and existing algorithm has been implemented by using MATLAB to test on the desired dataset

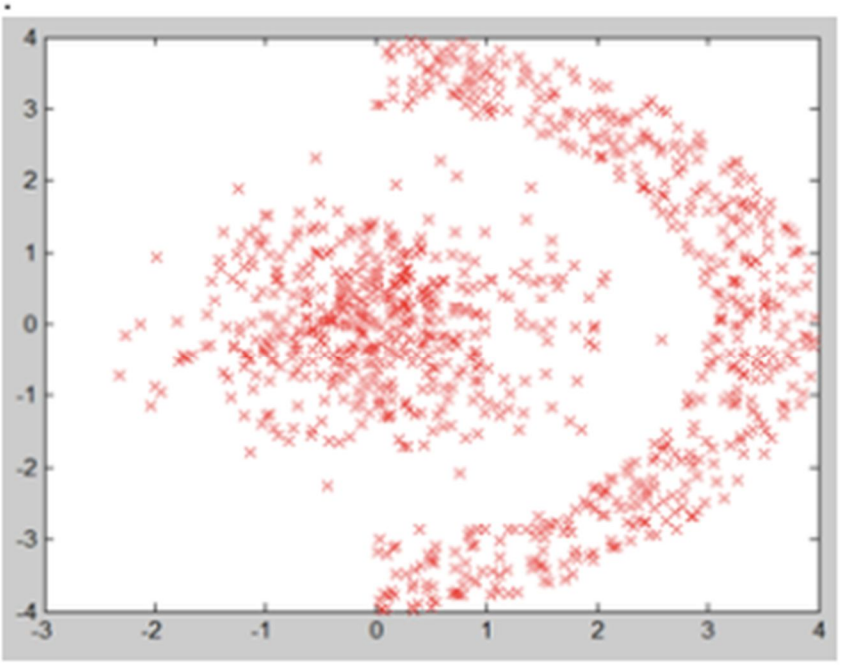


Fig.2. Incremental DBSCAN Algorithm

The rule of DBSCAN algorithm is applied that could cluster the similar and dissimilar style of knowledge from the foremost dense regions within the input dataset as shown in the figure 2:

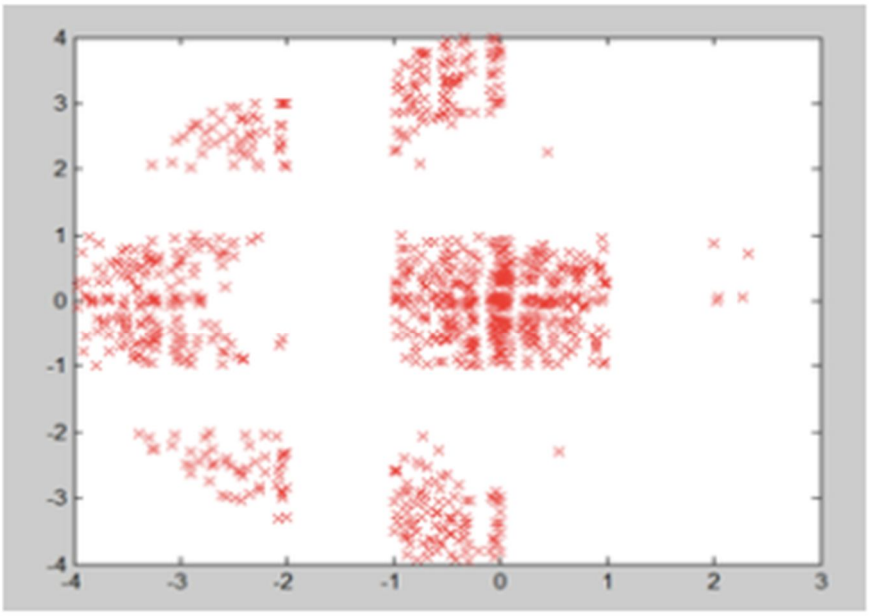


Fig.3. Enhancement DBSCAN Algorithm

The advance within the existing DBSCAN algorithmic rule has been projected within which back propagation algorithmic rule has been applied to calculate geometrician distance in reiterative manner. The obtained results provides enhancement in the accuracy of clustering as shown in figure 3.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue V May 2022- Available at www.ijraset.com

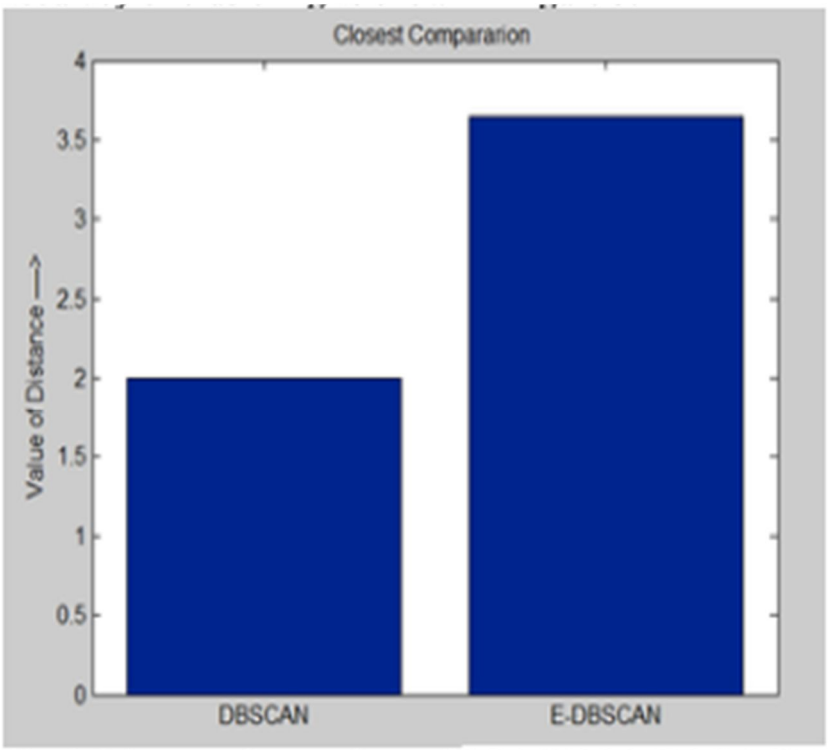


Fig.4. Distance Comparison of DBSCAN & E-DBSCAN

The gap of the proposed and progressive DBSCAN algorithm compared and it is analyzed that distance of Enhancement in DBSCAN (E-DBSCAN) algorithm is additional accuracy than existing DBSCAN algorithm as shown in figure 4. The performance enhancement of DBSCAN density based clustering algorithm in datamining is summarized in table-1.

Table-1. Performance Enhancement of DBSCAN & E-DBSCAN Algorithm

Parameter	**DBSCAN**	**E-DBSCAN**
Accuracy	86%	92%
Time	5.5 sec	4.71 sec
Distance	2	3.7
EPS	1.33	0.9

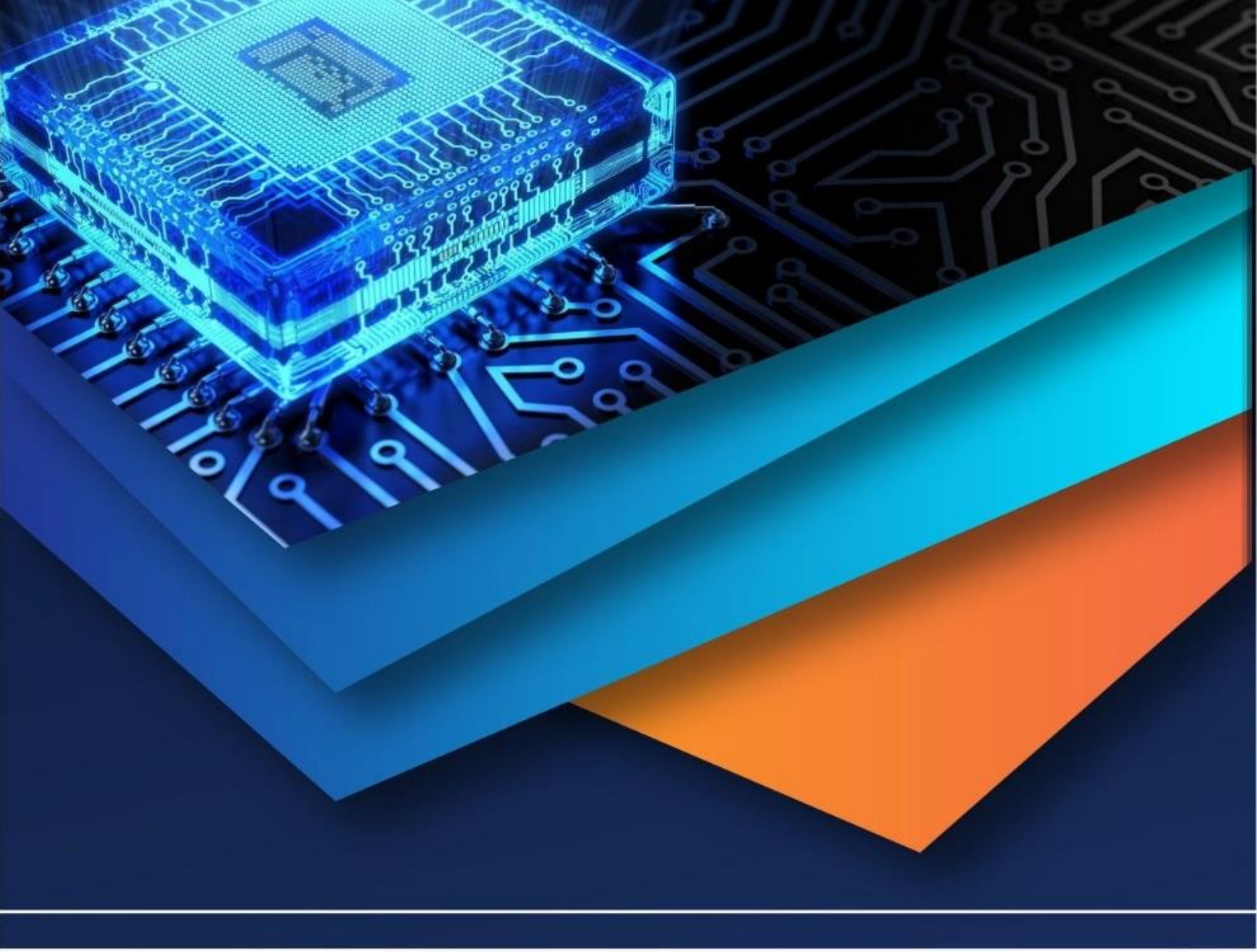
In this section, the performance enhancement of DBSCAN algorithm and Enhancement DBSCAN(E- DBSCAN) algorithm has been compared together in terms of accuracy, time, distance and EPS value which are shown in table 1.

V. CONCLUSION

This paper presents a comprehensive study of DBSCAN algorithm and the enhanced version of DBSCAN algorithm with its implementation using matlab. It is concluded that the density based cluster is that the economical kind of cluster throughout that clusters unit printed on the density of the input dataset. The DBSCAN is associate in Nursing rule that EPS worth is computed that may be central purpose and geometric distance is calculated from the central purpose that outlines similarity and unsimilarity of the knowledge. In this paper the ecludien distance is measured in dynimic manner whenever in previous work it was made in the static manner using Boltzmann learning algorithm. The proposed enhancement ends up in increase accuracy of the cluster and reducing the execution time.

REFERENCES

[1] 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.
[2] 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.
[3] 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.
[4] 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.
[5] 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.
[6] 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.
[7] 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.
[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] Osamor VC, Adebiyi EF, Oyelade JO and Doumbia S "Reducing the Time Requirement of K-Means Algorithm" PLoS ONE, Volume 7, Issue 12, 2012. 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.
[10] 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.
[11] 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. Chieh-YuanTsaiandChuang-ChengChiu,"Developingafeature weight self-adjustment mechanism for a K-means clustering algorithm," Computational Statistics and Data Analysis, pages 4658-4672, Volume 52, 2008.
[12] 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. DianweiHan,AnkitAgrawal,WeikengLiao,AlokChoudhary," A novel scalable DBSCAN algorithm with Spark", 2016, IEEE, 97879-897-99-4.
[13] Nagaraju S,Manish Kashyap, Mahua Bhattacharya," A Variant of DBSCAN Algorithm to Find Embedded and Nested Adjacent Clusters", 2016, IEEE, 978-1-4673-9197-9.
[14] Jianbing Shen, Xiaopeng Hao, Zhiyuan Liang, Yu Liu, Wenguan Wang, and Ling Shao," Real-time Superpixel Segmentation by DBSCAN Clustering Algorithm", 2016, IEEE, 1057-7149.
[15] Ilias K. Savvas, and Dimitrios Tselios," Parallelizing DBSCAN Algorithm Using MPI", 2016, IEEE, 978-1-5090-1663-1.
[16] Ahmad M. Bakr , Nagia M. Ghanem, Mohamed A. Ismail," Efficient incremental density-based algorithm for clustering large datasets", 2014, Elsevier Pvt. Ltd.
[17] J.R. Wen, J.-Y. Nie, H.-J. Zhang, Query clustering using user logs, ACM Transactions on Information Systems 20 (1) (2002) 59–81.
[18] C. Spieth, F. Streichert, N. Speer, A. Zell, Clustering based approach to identify solutions for the inference of regulatory networks, in: Proceedings of the IEEE Congress on Evolutionary Computation, Edinburgh, UK, 2005.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY