Comparative Study of Density Based Clustering Algorithms for Data Mining

[1]Deepak Jain, [2]Manoj Singh, [3]Dr. Arvind K Sharma
[1]Dept. of CSE, Gurukul Institute of Engineering and Technology (RTU) Kota, Rajasthan, India
[2]Dept. of CSE, Gurukul College of Engineering & Technology, Kota, Rajasthan, India
[3]Dept. of CSI, University of Kota, Rajasthan, India

Abstract
Now days, due to the explosive growth of huge amount of data have been uploaded into several websites. Thus it needs to be classified. Data mining is the process of extracting useful information from huge databases. Many approaches of data mining have been proposed to discover useful and accurate information among vast amount of data such as clustering, association rule mining,time series analysis and sequential pattern discovery etc. Thus, the density-based clustering algorithms have been used to find clusters based on the density of points in dense regions. Data clustering can be used in many application areas such as marketing, planning, insurance, biology, network security, earthquake, crime detections, intrusion detection systems etc. This paper presents a comparative study of various density based clustering algorithms for data miningalongwith their merits and demerits.

Keywords
Clustering Algorithms, Data Mining, Density based Algorithms, DBSCAN.

I. Introduction

Now day's it is believed that the information leads to power as well as a great success. The computers, satellites and other technologies are present today which are source of deriving information. Huge amount of data is collected from surroundings and utilized by the people for their own purposes. The data and information which can be further required is stored eventually. There are computers and other mass digital storage devices present in the technological world which provide the facility of storing the required information. There are however, varieties of devices which helps provide the facility to store different varieties of data. A structured database has been created in order to avoid all the chaos. For the purpose of its complete management, the Database Management System (DBMS) has been evolved that helps in proper arrangement of large data in an effective manner.

The DBMS also ensures that the data can be efficiently be retrieved from the huge collection as and when required by the users. The huge collection of all such data is thus possible mainly because of the expansion of the DBMS. The data from all of the fields such as the business world, the scientific data, satellite pictures, text reports, or the military intelligence is to be handled. For the purpose of decision making, the information retrieval method is not enough. For the purpose of making the management of data better, various new methods have been evolving. The activities which involve the automatic summarization of data, the extraction of important information stored, and the discovery of patterns in raw data are to be taken care of here. The analysis and interpretation of such huge amount of data is very important from the stored files and databases. This can also be required for the purpose of providing important related information which can help further in decision-making [1].

The rest of paper is organized as follows: Section 2 describes related work. Section 3 explains Data mining and KDD process and its necessary steps. Section 4 presentsa taxonomy of density based clustering algorithms. Section 5 describesa comparative study of density based clustering algorithms. Lastly, Section 6 concludes the paper.

II. Related Work

In this section, we discuss the several research workhave been carried out by many researchers in the field of density based clustering for data mining.

Qi Xianting et al. [2]proposed a paper in which a sub-part of the density-based representative algorithms is the density-based spatial clustering of applications with noise (DBSCAN) which has been utilized in certain fields because of its property of detecting the clusters which are of various shapes as well as sizes. When high dimensional data is present in certain applications that is when the algorithm stays no more stable. For the purpose of resolving this issue, an enhanced DBSCAN algorithm which is based on feature selection (FS-DBSCAN) is put forth. This algorithm is provided on various real world datasets and the various series of simulations are achieved.

Ahmad M. Bakr et al. [3]proposedan enhanced version of the incremental DBSCAN algorithm is introduced for incrementally building and updating arbitrary shaped clusters in extensive datasets. The need to organize such information in an efficient manner is more essential than any other time in recent memory. With such dynamic nature, incremental clustering algorithms are constantly preferred compared to traditional static algorithms. The proposed algorithm enhances the incremental clustering process by limiting the search space to partitions as opposed to the whole dataset which results in significant improvements in the performance compared to relevant incremental clustering algorithms. Experimental results with datasets of various sizes and dimensions demonstrate that the proposed algorithm speeds up the incremental clustering process by factor up to compared to existing incremental algorithms.

Cheng-Fa Tsai et al. [4]proposed anew data clustering method which also includes the data mining to be performed in applications which involve huge databases. There are many issues which arise during the clustering with respect to many aspects. There is a need of providing proper data analysis to reflect the broader appeal of it. The main objective of clustering techniques is to partition a set of data points into classes in such a manner that properties of points within a similar class are different from the ones belonging to other classes. There are various experiments conducted and the simulation results achieved show that the newly proposed clustering algorithm is efficient as compared to the Fast SOM combined with K-means algorithm as well as the genetic K-means algorithms. The errors given by this algorithm are also very less as compared to the other methods.

Wenbin Wu et al. [5]proposed a new approachin this paper. For this purpose of enhancing the forecasting accuracy and handling the training sample dynamics, In this approach, the K-means clustering algorithm is used along with the neural network for the cases where short term WPF is involved. The samples are classified into various categories on the basis of the similarities provided from earlier concepts which involve the K-means clustering. These categories hold the information of meteorological conditions and historical power data. 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.

Vadlana Baby et al. [6] presented a paper based on 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. There is less number of iterations in the newly proposed protocol as compared to the previous ones. There is no trust required from the end of servers or users. There are certain comparisons made with respect to various techniques. 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.

K.M. Archana Patel et al. [7] given the most genuinely utilized unsupervised learning method in the case of data mining is the clustering mechanism. The similar data objects are located within same clusters based on any particular type of similarity amongst them. There are seven various groups in which the clustering algorithms are categorized. As per their conditions, various results are given by these different clustering algorithms. There are certain techniques which help in clustering data present in huge data sets. There are other techniques which provide better results for the purpose of finding cluster with arbitrary shapes. In this paper, various data mining clustering algorithms are learned and related. There are some clustering algorithms such as k-means algorithm, K-medoids, and distributed k-means clustering algorithm which are discussed. There are various factors which are kept as base for providing comparisons in between these algorithms. There are certain specifications which are enlisted after comparisons of these algorithms which define which algorithm will be beneficial at certain conditions. The clustering algorithms are to known well for providing better results. There is therefore, no such algorithm which can be applied at all different scenarios.

Yumian Yang et al. [8] provided a paper in which with the rapid development of E-commerce, how to evaluate the e-commerce sites accurately has turned into an imperative issue. Be that as it may, evaluation record of e-commerce sites has characteristics of high dimensions and uneven density, which leads to awful performance of the evaluation result. To analyze 100 e-commerce show undertakings in 2013-2014 named by the Ministry of Commerce People's Republic of China, this paper reduces dimensionality by factor analysis method firstly, then implements an improved DBSCAN algorithm to process the uneven density data, finally offers suggestions to these 100 Ecommerce endeavors based on investigation results. Since DBSCAN algorithm ignores weights while calculating the Euclidean distance, the result of the similarity measurement is not accurate, while factor analysis is a good means to deal with weights. This paper improves the clustering accuracy and reasonableness of the evaluation by combining factor analysis and DBSCAN. Notwithstanding, the data processed by factor analysis have characteristics of uneven density. The traditional DBSCAN is improved to partition the data with various densities and cluster these sites. This paper advances another processing idea on E-commerce sites evaluation: another DBSCAN algorithm combining factor analysis with various densities. Compared with the traditional DBSCAN algorithm, the results of evaluating websites are more reasonable and interpretable with the improved DBSCAN algorithm. Later on work, the scale of the evaluation object will be further expanded and more research should be finished.

XiaoqingYu et al. [9] have authored a paper that examines the basic principle of DBSCAN algorithm and its implementation process. Spatial clustering is one of the principle methods of data mining and knowledge discovery. DBSCAN algorithm can be found in space with "noise" database clustering of arbitrary shape, is a sort of good clustering algorithm. This paper introduces the basic concept and principle of DBSCAN algorithm, and applies this algorithm to perform clustering analysis distributions of web location information. The article contrast K-means algorithm and DBSCAN algorithm in order to demonstrate the effectiveness of DBSCAN algorithm. The DBSCAN algorithm will generate much noise points, and there are a couple of data points in its each cluster. Be that as it may, the K-means algorithm doesn't shape noise points. So it can be affected by a couple points. It applies this algorithm in the field of city planning to locate the hot territory in the city. What's more, it compared DBSCAN algorithm and k-means algorithm, and demonstrate its effectiveness. Later on, it can be utilized to analyze city public offices or municipal public offices to give a scientific premise and guidance for city planning.

Jing Gaoet al.[10] proposed an improved algorithm (ICFS) to deal with the several weaknesses of it. Unlike CFS, the proposed algorithm designs a formula for the cutoff distance calculation and a method for cluster centers selection to improve its robustness. Moreover, a new non-center point's allocation strategy and the cluster merging and splitting processes were developed to adapt to the density peaks and adjust the clusters dynamically, which can improve the clustering accuracy and scalability. The ICFS method was evaluated on several datasets by comparison with the original CFS algorithm. Results demonstrate the effectiveness of the proposed method.

Kuan-Teng Liao et al. [11] presented a paper that discusses the centroid based clustering and the UKmeans algorithm. There is uncertain data clustering also present within the applications which results in causing errors. Due to these errors, the time cost as well as the effectiveness of the system is affected gradually. The time cost needs to be reduced and the effectiveness needs to be increased for providing a proper clustering algorithm. The similarity of the application is improved through the first mechanism. The time cost and effectiveness are affected through this similarity factor. For instance, the time cost is ignored by the similarity calculations along with a special focus on the effectiveness of the clustering. In contrast, the time cost issue is a major concern for the similarity calculations along with simplified approaches while the effectiveness property is ignored. So, for taking time cost and effectiveness measures as a major concern equally, an enhancement is proposed. A simplified similarity is utilized to reduce the time cost and add additional two factors which are intersection and density of clusters. These two factors will further help in increasing the effectiveness of the clustering mechanism. During the overlapping of a cluster on the object, the degree of the object belongingness is increased with the help of the intersection

factor. The range can be decreased here by providing square root boundary mechanism for limiting the upper bound of possible positions of centroids which will further help in increasing the effectiveness of clustering. It is seen through the experimental results that the mechanisms provide better results in terms of time cost and effectiveness.

III. Data Mining and KDD Process

Data mining allows extracting data from the huge information and changing that data into a reasonable and important structure for additionally utilize. Data mining is a fundamental task during the time spent learning revelation from large information[12].Data mining is an advance mechanism that is very useful to mine the comprehensible knowledge, previously unknown, information from large amount of data stored in various formats, with the objectives of improving the decision of companies, organizations where the data would be collected [13].Another name for Data Mining is the Knowledge Discovery in Databases (KDD). From the data present in the databases, the KDD helps in nontrivial extraction of implicit as well as potentially useful information. Data mining is originally a part of KDD process that is used as a synonym. There are various steps followed in the case of knowledge discovery from databases. The steps begin from identifying the raw material and gathering it to form new important information.

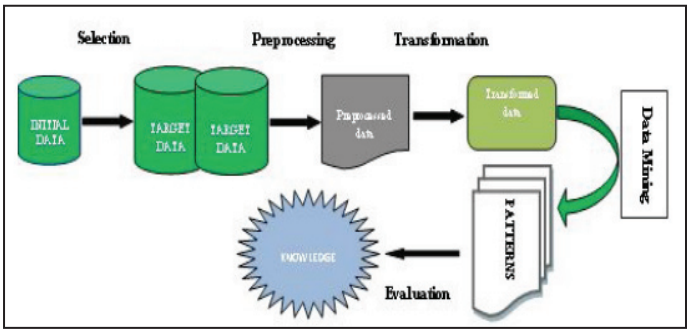

Fig. 1: Data Mining as Core of KDD Process [14]

The following steps are involved in the data miningprocess as under:

A. Data Cleaning

The removal of noise data or the irrelevant data from the whole collection is known as data cleaning.

B. Data Integration

The combination of multiple data sources which are heterogeneous, into a common source is known as data integration.

C. Data Selection

The data which is relevant to data analysis to be performed is identified and retrieved from the collection in this step.

D. Data Transformation

In this step, the gathered data is transformed into forms which are useful for the mining process. It is also known as data consolidation.

E. Data Mining

In this step, the patterns which are extremely important are extracted with the help of clever techniques.

F. Pattern Evaluation

On the basis of provided measures, the interesting patterns which represent the knowledge are recognized within this step.

G. Knowledge Representation

In this last step, the user can view the discovered knowledge. Visualization techniques are used to help the users to understand and interpret the results achieved from data mining [15].

There are different types of clustering methods have been developed namely partitioning, hierarchical, density, grid, model, and constraint based. Among these, the density based method works based on the notion of density. Here, the clusters are formed as thick regions which are apart from thin regions. The fundamental idea is that increasing the identified cluster until the density.

IV. Density Based Clustering Algorithms

At the outset, clustering is a fundamentaltechniqueunder many circumstances including data mining, patternrecognition, image processing and other industrial applications.During the past decades, many clustering algorithms have been developedsuch as DBSCAN, GDBSCANUDBSCAN, P-DBSCAN VDBSCAN etc.The taxonomy of various density based clustering algorithms is shown in fig. 2.

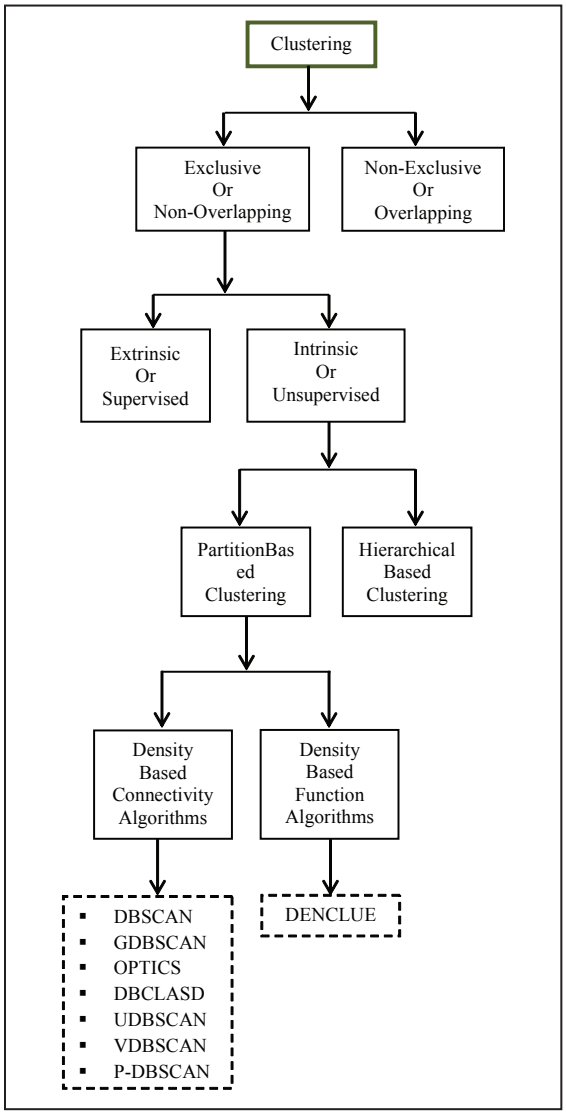

Fig. 2: Taxonomy of Clustering Algorithms

(number of objects) in the neighborhood goes beyond some threshold.

V. Comparative Study of Density Based Clustering Algorithms

In this section, we summarize the various density based clustering algorithms along with their merits and demerits and different-different parameters as shown in Table 1.

Table 1: Comparative Study of Density Based Clustering Algorithms

Name of the Algorithm	Density Based Clustering	Density Based Spatial Clustering of Applications with Noise	Distributed-Based Clustering Algorithm for Mining Large Spatial Databases	Varied Density Based Spatial Clustering of Applications with Noise	Density Based Algorithm for discovering Density Varied Clusters in Large Spatial Databases
Type of Data	Large number of data	Spatial Data with Noise	Spatial Data with uniformly Distributed points	Spatial Data with Varied Density	Spatial Data with Varied Density
Type of Density	Yes	No	Yes	YES	YES
Input Parameters	Two input Parameters	Radius and Minimum Size	Automatically Generated	Automatically Generated	Two input Parameters
Complexity	O (log D)	O (n^2)	O ($3n^2$)	Same as DBSCAN	Higher than DBSCAN
Objectives	Can discover other clustering algorithms like hierarchical clustering, partition based clustering etc.	Discover clusters with arbitrary shape	Design good cluster for spatial database	Find out meaningful cluster in database w.r.t widely varied density	Find out the density variations that exit within the cluster
Merits	Good clustering properties in data sets with large amount of noise	DBSCAN doesn't require no. of cluster in the data at prior stage	DBCLASD requires no user input	Automatically select several input parameter and detect cluster with varied density	Handles local density variation within the cluster
Demerits	Data points are assigned by hill climbing, it make unnecessary small steps	Does not respond data with varied density	Slower than DBSCAN	If parameter selection goes wrong then it has problem	High time complexity

VI. Conclusion and Future Scope

The clustering is one of the most popular datamining algorithms in which the similar and dissimilar type of data could be clustered together to analyze complex data. Moreover, the algorithm of density based clustering is applied which could cluster the similar and dissimilar type of data according to the data density in the input dataset. In the density based clustering the most dense region is calculated from which similar and dissimilar type of data computed using similarity techniques.

In future, the DBSCAN (Density Based Spatial Clustering of Applications with Noise) algorithm will be applied to compute the EPS value which will be the central of the datasets. The EPS value will be calculated dynamically to achieve maximum accuracy.

References

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

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

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

[4] Cheng-Fa Tsai, Han-Chang Wu, Chun-Wei Tsai,"A New Data Clustering Approach for Data Mining in Large Databases", 2002, IEEE.

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

[6] Vadlana Baby, Dr. N. Subhash Chandra,"Distributed threshold K-means Clustering for Privacy preserving data mining", 2016, IEEE

[7] KM Archana Patel and PrateekThakral, "The Best Clustering Algorithms in Data Mining", 2016, IEEE.

[8] Yumian Yang, Jianhua Jiang,"Application of E-commerce Sites Evaluation based on Factor Analysis and Improved DBSCAN Algorithm", 2014, IEEE.
[9] XiaoqingYu, Yupu Ding, Wanggen Wan, Etienne Thuillier, "Explore Hot Spots of City Based on DBSCAN Algorithm", 2014, IEEE.
[10] Jing Gao, et al.,"ICFS: An Improved Fast Search and Find of Density Peaks Clustering Algorithm", 2016 IEEE 14th Intl Conf on Dependable,
[11] Kuan-Teng Liao, Chuan-Ming Liu,"An Effective Clustering Mechanism for Uncertain Data Mining Using Centroid Boundary in UKmeans", 2016, IEEE
[12] K. Kameshwaran, K. Malarvizhi,"Survey on Clustering Techniques in Data Mining", International Journal of Computer Science and Information Technologies, Vol. 5(2), 2014.
[13] Ankit Bhardwaj, et.al,"Data Mining Techniques and Their Implementation in Blood Bank Sector –A Review", International Journal of Engineering Research and Applications (IJERA), Vol. 2, Issue 4, pp.1303-1309, 2012.
[14] Arvind Sharma, P.C. Gupta,"Predicting the Number of Blood Donors through their Age andBlood Group by using Data Mining Tool", Vol. 01, No. 6, Issue, 2012.
[15] D. Wang, S. Zhu, T. Li, Y. Chi, Y. Gong,"Integrating clustering and multi document summarization to improve document understanding", 17th ACM CIKM Conference on Information and Knowledge Management, 2008.

Mr. Deepak Jain received his B.Tech degree in Computer Science and Engineering from Modi Institute of Technology (RTU) Kota, Rajasthan, India in 2014 and received MBA degree from Vardhman Mahaveer Open University Kota Rajasthan India in 2017 and pursuing M.Tech in Computer Science and Engineering from Gurukul Institute of Engineering and Technology (RTU) Kota, Rajasthan, India. His research interest lies in the area of data mining and educational data mining.