ISSN XXXX XXXX © 2017 IJESC

Research Article Volume 7 Issue No.6

Efficient Incremental Density-Based Algorithm for Clustering Large Datasets

Mandeep Kaur[1], Rupinderpal Singh[2]
Research Scholar[1], Assistant Professor[2]
Department of Computer Science
Global Institutes of Management &Emerging Technologies, India

Abstract:
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

Keywords: Data mining, Classification, Clustering, DBSCAN.

1. INTRODUCTION

1.1 Introduction to Data Mining

Data mining is viewed as a result of the natural evolution of information technology. The early development of data collection and database creation mechanisms proved to be important for the later development of effective mechanisms for data storage and retrieval, query and transaction processing. The database and data management industry evolved in the development of several critical functionalities: data collection and database creation, data management and advanced data analysis(involving data warehousing and data mining). One of the emerging data repository architecture is the data warehouse. It involves multiple heterogeneous data sources organized under a unified schema at a single site to manage decision making. Data cleaning, data integration, and online analytical processing (OLAP) are involved in Data warehouse technology. Data mining is the duty of discovering interesting patterns from large amounts of data where the data can be stored in databases, data warehouses and repositories of other information. It is also commonly referred to as knowledge discovery in databases (KDD). Data mining involves an integration of techniques from number of disciplines such as statistics, database technology, machine learning, neural networks, high-performance computing and pattern matching, data visualization, information recovery etc [1].

Data mining tasks have the following categories [2]:

1. **Class description**: Description of the class can be helpful to depict individual classes and concepts in summarized, concise, and yet precise terms.
2. **Association analysis:** Association analysis is the discovery of association rules showing attribute-value conditions that occur frequently together in a given set of data.
3. **Classification:** Classification is the procedure of ruling a set of models that explain and differentiate data classes, its concepts, for the reason of being able to utilize the model to estimate the class of objects whose class label is unknown.
4. **Cluster analysis**: Clustering analyzes data objects without consulting a known class. In general, the class labels are not present there in the training data because they are not recognized where to begin. The objects are grouped or clustered based on the principle of maximizing the intra-class similarity and minimizing the inter-class similarity.
5. **Outlier analysis**: Outliers are data objects that do not conform to the broad-spectrum behavior of model of the data. Outliers may be detected using statistical tests or using distance measures.
6. **Evolution analysis**: It describes and models trends for objects whose behaviors changes over time. It usually includes time-series data analysis, succession or periodicity pattern matching, and similarity-based data analysis [2].

In outline, the abundance of data, coupled with the requirement for powerful data analysis tools, has been described as a data rich yet information poor situation.

1.2 Knowledge Discovery from Data (KDD) Process

This section explains the various steps in knowledge discovery from data or KDD process. The term Knowledge Discovery from data or KDD, refers to the broad process of finding knowledge in data, and emphasizes the "high-level" application of particular data mining methods. The role of data mining (KDD) is very important in many of the field such as the analysis of market basket, classification, etc. KDD has been very interesting topic for the researchers as it leads to automatic discovery of useful patterns from the database. This is also called knowledge discovery from the large amount of database. Many techniques have been developed in data mining amongst which primarily Association rule mining is very important which results in association rules. These rules are applied on market based, banking based etc. for decision making. The unifying goal

of the KDD process is to extract knowledge from data in the context of large databases.

Steps in KDD include [4]:

1. **Data Cleaning:** in this step noise and inconsistent data is removed.
2. **Data Integration:** in this step data from multiple data sources is combined.
3. **Data Selection:** in this step data relevant to the analysis task are retrieved from the database.
4. **Data Transformation:** in this step aggregate or summary operations are performed to transform and consolidate data into forms appropriate for mining.
5. **Data Mining:** it is an essential process where data patterns are extracted by applying intelligent methods.
6. **Pattern Evaluation:** in this step the truly interesting patterns are identified for representing knowledgebase on interestingness measures.
7. **Knowledge Presentation:** in this step visualization and knowledge representation techniques are used to present mined knowledge to users.

1.3 Data Mining

Data mining is an interdisciplinary subfield of computer science which is the process of discovering insightful, interesting and novel patterns from large-scale sets. Data mining [8] is an essential step of the "Knowledge Discovery in Databases" process, or KDD. Data mining is often treated as synonym for another popularly used term, Knowledge Discovery in Databases (KDD).The data sources can include databases, data warehouses, the web, other information repositories, or data that are streamed into the system dynamically. It can be applicable to any kind of data repository. There is different kind of algorithms and techniques are available for different types of data. Data mining is studied for different databases like object-relational databases, relational database, data ware houses and multimedia databases etc. Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks. In general, data mining tasks can be classified into two categories: descriptive and predictive. Descriptive mining tasks characterize the general properties of the data in the database. Predictive mining tasks perform inference on the current data in order to make predictions.

1.4 Clustering in Data Mining

Cluster analysis has been widely used in numerous applications, including market research, pattern recognition, data analysis, and image processing. In business, clustering can help marketers discover interests of their customers based on purchasing patterns and characterize groups of the customers. In biology, it can be used to derive plant and animal taxonomies, categorize genes with similar functionality, and gain insight into structures inherent in populations. In geology, specialist can employ clustering to identify areas of similar lands; similar houses in a city and etc. data clustering can also be helpful in classifying documents on the Web for information discovery [7]. There are numerous clustering algorithms utilized for clustering. The major fundamental clustering methods can be classified into taking after categories.

1. **Partitioning Methods:-**The general criterion for partitioning is a combination of high similarity of the samples within clusters with high dissimilarity between distinct clusters. Most partitioning methods are distance-based. Given k, the quantity of partitions to build, a partitioning strategy creates an initial partitioning and afterward utilizes an iterative relocation system that attempts to enhance the partitioning by moving objects starting with one group then onto the next. The goal is, given a k, discover a partition of k clusters that optimizes the picked partitioning criterion. Here k is an input parameter. E.g. K-mean and K-centroid[10]

2. **Hierarchical Methods:-**In this technique hierarchical decomposition of the given set of data objects is created. It can be classified as being either agglomerative or divisive based on how hierarchical decomposition is formed. Agglomerative approach is the base up approach starts with every object forming a separate group. It then merges groups close to each other until every one of the groups is merged into one. algorithms make a hierarchical decomposition of the given data set of data objects. The hierarchical decomposition is represented by a tree structure, called dendrogram. It needn't bother with clusters as inputs. In this sort of 3. **Density Based Methods:-**Most partitioning methods cluster objects based on distance between objects. Spherical shaped clusters can be discovered by these methods and encounter trouble in discovering clusters of arbitrary shapes. So for arbitrary shapes new methods are utilized known as density-based methods which are based on the notion of density. In these methods the cluster is continued to develop as long as the density in the area exceeds some threshold. This strategy is based on the notion of density.

4. **Grid Based Methods:** Grid based methods quantize the object space into a finite number of cells that frame a grid structure. It is a fast technique and is independent of the number of data objects and depends just on the number of cells in every dimension in the quantized space. In this objects meet up to shape grid. The object space is quantized into finite number of cells that shape a grid structure.

1.5 DBSCAN

Density based clustering algorithms have a wide applicability in data mining. They apply a local criterion to group objects: clusters are viewed as regions in the data space where the objects are dense, and which are separated by regions of low object density (noise) [12]. Among the density based clustering algorithms 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 process, this algorithm needs two numeric input parameters, minPts and which together characterize the desired density characteristics of the generated clusters. In particular, minPts is a positive integer determining the minimum number of objects that must exist inside a maximum distance of the data space all together for an object to have a place with a cluster. 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, the client gets a proposal on which parameter value that would be reasonable. Along these lines, minimal knowledge of the domain is required. DBSCAN can discover clusters of arbitrary shape. In any case, clusters that

lie close to each other have a tendency to have a place with a similar class [13].

2. LITERATURE REVIEW

AUTHOR & TITLE	YEAR	DESCRIPTION	OUTCOME
Guangchun Luo, et.al," A Parallel DBSCAN Algorithm Based On Spark".	2016	With the explosive growth of data, the method has entered the period of big data. In order to sift through masses of information, numerous data mining algorithms utilizing parallelization are being implemented. Cluster analysis occupies a pivotal position in data mining, and the DBSCAN algorithm is a standout amongst the most broadly utilized algorithms for clustering.	The experimental result demonstrates the proposed S_DBSCAN algorithm can effectively; and efficiently; generates clusters and identify noise data. In short, the S_DBSCAN algorithm has superior performance when dealing with massive data, as compared to existing parallel DBSCAN algorithms.
Dianwei Han, et.al," A novel scalable DBSCAN algorithm with Spark".	2016	DBSCAN is an outstanding clustering algorithm which is based on density and can identify arbitrary shaped clusters and eliminate noise data.	This paper presented another Parallel DBSCAN algorithm with Spark. It maintains a strategic distance from the communication amongst executors and in this way prompts to a better scalable performance. The results of these analyses demonstrate that the new DBSCAN algorithm with Spark is scalable and outperforms the implementation based on MapReduce by a factor of more than 10 in terms of efficiency.
Nagaraju S, et.al," A Variant of DBSCAN Algorithm to Find Embedded and Nested Adjacent Clusters",	2016	This paper introduce an efficient approach for clustering analysis to detect embedded and nested adjacent clusters utilizing idea of density based notion of clusters and neighborhood difference.	This paper the notion of density based approaches for data clustering and thought of neighborhood difference is utilized effectively detect embedded and nested adjacent clusters. Our experimental results suggested that proposed algorithm effective in detecting nested adjacent clusters compared to DBSCAN and EnDBSCAN algorithm with computational complexity as same as DBSCAN algorithm.
JianbingShen, et.al," Real-time Superpixel Segmentation by DBSCAN Clustering Algorithm".	2016	This paper proposes a real-time picture superpixel segmentation method with 50fps by utilizing the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm.	This algorithm achieves the state-of-the-art performance at a considerably smaller calculation cost, and significantly outperforms the algorithms that require more computational costs even for the pictures including complex objects or complex texture regions.
Ilias K. Savvas, et.al," Parallelizing DBSCAN Algorithm Using MPI".	2016	The most recent years, immense bundles of information are extracted by computational frameworks and electronic gadgets. To exploit the derived amount of data, new innovative algorithms must be employed or the set up ones might be changed.	The results obtained from various concrete examples proved that were identical with the results delivered by the application of the original sequential technique. The time complexity reduced dramatically and the experimental results demonstrated that the algorithm scales in an exceptionally efficient manner.
Ahmad M. Bakr, et.al," Efficient incremental density-based algorithm for clustering large datasets".	2014	In this paper, an enhanced version of the incremental DBSCAN algorithm is introduced for incrementally building and updating arbitrary shaped clusters in extensive datasets. The proposed algorithm enhances the incremental clustering process by limiting the search space to partitions as opposed to the whole dataset which results in	Experimental results demonstrate that the proposed algorithm has a comparable accuracy compared to related incremental clustering algorithms. In any case, the proposed algorithm has significant improvements on the runtime with a speedup factor of 3.2.

		significant improvements in the performance compared to relevant incremental clustering algorithms.	
SaefiaBeri, et.al," Hybrid Framework for DBSCAN Algorithm Using Fuzzy Logic".	2015	Data mining process is to get information from a data set and after that convert it into understandable and meaningful information for further utilize. DBSCAN, a density based clustering algorithm, identifies clusters of shifting shape and outliers.	With this improved hybrid DBSCAN algorithm, certain parameters, for example, accuracy, geometric accuracy, bit error rate, specification, and sensitivity and error rate will be evaluated and the results will be compared over the DBSCAN algorithm. The hybridization will allow DBSCAN to choose the cluster in more efficient manner.
KarlinaKhiyarinNisa, et.al," Hotspot Clustering Using DBSCAN Algorithm and Shiny Web Framework".	2014	Clustering is performed on a dataset of hotspots on Kalimantan Island and South Sumatra Province in 2002-2003. The spread example of hotspots resulted by this clustering can be utilized as a predictive model of forest fires event and can be accessed through the internet browser.	This research developed a web-based application clustering with DBSCAN algorithm utilizing the R programming language with Shiny framework. DBSCAN needs minPts and Eps parameter. The bigger values of minPts will create less, however more the number of noises. While the bigger value of Eps will result in less clusters. MinPts parameter determination is finished by taking a gander at the dimensions of the data and plot the graph of minPts and the number of clusters and noise. While Eps parameter determination is obtained from k-dist graph observation and the slope difference calculations.
NegarRiazifar, et.al," Retinal Vessel Segmentation Using System Fuzzy and DBSCAN Algorithm".	2015	The point of this paper is to analyze the retinal vessel segmentation based on the clustering algorithm DBSCAN relying on a density-based notion of clusters which is designed to find clusters of arbitrary shape. DBSCAN requires one and only input parameter and a value for this parameter is suggested to the client.	The DBSCAN algorithm solves every one of the problems when utilizing clustering methods, finds the right input parameters, localizes clusters of arbitrary shapes and does the whole processes in a reasonable time. The performance of the algorithm in this paper is better in general than the previous ones. The new plan reduces the idle time and enhances the speed and accuracy of segmentation.

3. CONCLUSION AND FUTURE SCOPE:

This work is based on the density based clustering which is applied to calculate similarity from the densest region which can define clusters on the basis of similar and dissimilar type of data. The Euclidian distance is calculated in the static manner which leads to reduction in accuracy of clustering. The proposed improvement will calculate Euclidian distance is dynamic manner which increase accuracy of clustering. When the Euclidian distance is calculated in dynamic manner, it leads to increase accuracy and also reduce execution time of the algorithm

4. 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]. NehaAggarwal, KirtiAggarwal and Kanika Gupta, "Comparative Analysis of k-means and Enhanced K-means clustering algorithm for data mining," International Journal of Scientific & EngineeringResearch, Volume 3, Issue 3, August-2012

[3]. AhamedShafeeq B M and Hareesha K S, "Dynamic Clustering of Data with Modified Means Algorithm," International Conference onInformation and Computer Networks, Volume 27, 2012

[4]. ManpreetKaur and UsvirKaur, "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 ISSN: 2277 128X

[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 NavotKaur, "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 NavotKaur, "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]. HarpreetKaur and Jaspreet KaurSahiwal, "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

[10]. AzharRauf, Sheeba, SaeedMahfooz, Shah Khusroand HumaJaved, "Enhanced K-Mean Clustering Algorithm to Reduce Number of Iterations and Time Complexity," Middle-East Journal of ScientificResearch, pages 959-963, 2012

[11]. Kajal C. Agrawal and MeghanaNagori, "Clusters of Ayurvedic Medicines Using Improved K-means Algorithm," International Conf. on Advances in Computer Science and Electronics Engineering, 2013.

[12]. 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.

[13]. Chieh-Yuan Tsai and Chuang-Cheng Chiu, "Developing a feature weight self-adjustment mechanism for a K-means clustering algorithm," Computational Statistics and Data Analysis, pages 4658-4672, Volume 52, 2008

[14]. Guangchun Luo, XiaoyuLuo, Thomas Fairley Gooch, Ling Tian, Ke Qin," A Parallel DBSCAN Algorithm Based On Spark", 2016, IEEE, 978-1-5090-3936-4

[15]. Dianwei Han, AnkitAgrawal, Wei−keng Liao, AlokChoudhary," A novel scalable DBSCAN algorithm with Spark", 2016, IEEE, 97879-897-99-4

[16]. NagarajuS, ManishKashyap, Mahua Bhattacharya," A Variant of DBSCAN Algorithm to Find Embedded and Nested Adjacent Clusters", 2016, IEEE, 978-1-4673-9197-9

[17]. JianbingShen, XiaopengHao, Zhiyuan Liang, Yu Liu, Wenguan Wang, and Ling Shao," Real-time Superpixel Segmentation by DBSCAN Clustering Algorithm", 2016, IEEE, 1057-7149

[18]. Ilias K. Savvas, and Dimitrios Tselios," Parallelizing DBSCAN Algorithm Using MPI", 2016, IEEE, 978-1-5090-1663-1

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

[20]. SaefiaBeri, KamaljitKaur," Hybrid Framework for DBSCAN Algorithm Using Fuzzy Logic", 2015, IEEE, 978-1-4799-8433-6

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

[22]. NegarRiazifar, EhsanSaghapour," Retinal Vessel Segmentation Using System Fuzzy and DBSCAN Algorithm", 2015, IEEE, 978-1-4799-8445-9

[23]. 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

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