

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology



ISSN 2320-088X
IMPACT FACTOR: 6.199

IJCSMC, Vol. 8, Issue. 5, May 2019, pg.234 – 238

Analysis of Density Based Clustering Techniques in Data Mining

Anjali Soni¹; Gaurav Kumar Srivastav²

¹M.Tech, Department of Computer Science & Engineering, Institute of Technology & Management, Aligarh, India

²Assistant Professor, Department of Computer Science & Engineering, Institute of Technology & Management, Aligarh, India

¹ anjalisoni0989@gmail.com; ² gauravjau@gmail.com

Abstract: *The data mining is the approach which extracts useful information from the rough information. The clustering is the approach of data mining which cluster the similar and dissimilar type of information. The clustering techniques is of various type which hierarchal clustering, density based clustering and so on. The DBSCAN algorithm is the density based clustering algorithm. The density based clustering has the various algorithms. In this research work, clustering techniques related to density based clustering is reviewed and analyzed in terms of certain parameters.*

Keywords: *Data Mining, Clustering, Density Based Clustering.*

Introduction

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) [1]. 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 [2]. OLAP basically is analysis techniques with functionalities for example, summarization, consolidation, and aggregation. It can also be known as the ability to view information from different angles. The effective and efficient analysis of data from such different forms of data by integration of information retrieval, data mining, and information network analysis technologies has proven to be a challenging task. 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 [3]. 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. Cluster analysis has been widely used in numerous applications, including market research, pattern recognition, data analysis, and image processing [4]. 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 [5]. 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. Data clustering (or just clustering), is an unsupervised classification method [6]. This method aims at creating groups of objects or clusters, in such a way that objects in the same cluster are very similar and objects in different clusters are quite distinct. Cluster analysis is one of the traditional topics in the data mining field. It is the first step in the direction of exciting knowledge discovery. The procedure of grouping data objects into a set of disjoint classes, called clusters is known as clustering [7]. Now objects within a class have high resemblance to each other in the meantime objects in separate classes are more unlike. Clustering is a technique used to group similar documents, however it differs from classification of documents are clustered on the fly instead of using predefined topics [8]. Another advantage of clustering is that documents can appear in multiple subtopics, in this manner guaranteeing that a helpful document won't be misplaced from indexed lists. A fundamental clustering algorithm shapes a vector of topics for every document and measures the weights of how healthy the document fits into every cluster [9]. Clustering goes under unsupervised classification. Classification alludes to a technique that assigns data objects to a set of classes. Unsupervised clustering means that clustering does not rely on predefined classes and training. Unsupervised clustering is not the same as pattern reorganization in the area of statistics known as discriminate analysis and decision analysis which arrange the objects from a given set of object [10]. 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) [11]. 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 [12]. 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 [13]. 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.

Literature Review

Guangchun Luo, et.al, (2016) proposed a parallel DBSCAN algorithm (S_DBSCAN) based on Spark, which can quickly realize the partition of the original data and the mix of the clustering results. It is divided into the three strides [14]. First, partitioning the raw data based on a random sample, then computing local DBSCAN algorithms in parallel, third, merging the data partitions based on the centroid. Compared with the traditional DBSCAN algorithm, the experimental result demonstrates the proposed S_DBSCAN algorithm gives better operating efficiency and scalability. This paper evaluates the S_DBSCAN algorithm by dealing with annual outpatient data. 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, (2016) studied that DBSCAN algorithm has been extremely famous since it can identify arbitrary shaped clusters and additionally handle noisy data. Be that as it may, parallelization of DBSCAN based on MPI and OpenMP suffers from lack of fault-tolerance. Also, in order to implement parallelization with MPI or OpenMP, data scientists need to deal with implementation in detail [15]. 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, (2016) proposed an efficient approach for clustering analysis to detect embedded and nested adjacent clusters utilizing idea of density based notion of clusters and neighborhood difference. 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]. Additionally the paper has preset

method to evaluate the global density parameters utilizing sorted k-distance plot and first order derivative. Through 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.

Jianbing Shen, et.al, (2016) proposed a real-time picture superpixel segmentation method with 50fps by utilizing the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm. In order to decrease the computational costs of superpixel algorithms, the method received a quick two-stage framework [17]. A robust and straightforward distance function is defined for getting better superpixels in these two stages. The experimental results demonstrate that our real-time superpixel algorithm (50fps) by the DBSCAN clustering outperforms the state-of-the-art superpixel segmentation methods in terms of both accuracy and efficiency. 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, (2016) designed a three phase parallel version of DBSCAN [18]. The obtained experimental results are exceptionally promising and demonstrate the correctness, the scalability, and the effectiveness of the technique. In this work, a parallel version of the notable DBSCAN was presented and implemented utilizing MPI. 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, (2014) proposed an enhanced version of the incremental DBSCAN algorithm 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 significant improvements in the performance compared to relevant incremental clustering algorithms [19]. Experimental results with datasets of various sizes and dimensions demonstrate that the proposed algorithm speeds up the incremental clustering process by factor up to 3.2 compared to existing incremental algorithms. The proposed algorithm is additionally proved to perform better in expansive datasets with higher dimensions compared to related algorithms.

Saefia Beri, et.al, (2015) proposed a framework of methodology of DBSCAN algorithm with the integration of fuzzy logic. The degree to which an object belongs to a particular cluster will be resolved utilizing membership values [20]. The improved version of DBSCAN algorithm will be the hybridization of DBSCAN algorithm with fuzzy if-then rules. To enhance the degree of membership, multivalent logic will mull over in which the membership values are to be utilized. This algorithm will be consolidated with fuzzy if then rules for breast cancer detection. 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.

Karlina Khiyarin Nisa, et.al, (2014) designed a web-based application clustering with DBSCAN algorithm utilizing the R programming language with Shiny framework. DBSCAN needs minPts and Eps parameter [21]. 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.

Table of Comparison

Authors Names	Year	Description	Outcomes
Guangchun Luo, Xiaoyu Luo, Thomas Fairley	2016	A parallel DBSCAN algorithm (S_DBSCAN) based on Spark was proposed, which can quickly realize the partition of the original	In short, the S_DBSCAN algorithm has superior performance when dealing with massive data, as compared to existing parallel
Gooch, Ling Tian, Ke Qin		data and the mix of the clustering results.	DBSCAN algorithms.
Dianwei Han, Ankit Agrawal, Wei-keng Liao, Alok Choudhary	2016	A Parallel DBSCAN algorithm with Spark was designed. 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, Manish Kashyap, Mahua Bhattacharya	2016	An efficient approach was proposed for clustering analysis to detect embedded and nested adjacent clusters utilizing idea of density based notion of clusters and neighborhood difference.	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.
Jianbing Shen, Xiaopeng Hao, Zhiyuan Liang, Yu Liu, Wenguan Wang, and Ling Shao	2016	A real-time picture superpixel segmentation method was proposed 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, and Dimitrios Tselios	2016	A three phase parallel version of DBSCAN was designed.	The time complexity reduced dramatically and the experimental results demonstrated that the algorithm scales in an exceptionally efficient manner.
Ahmad M. Bakr, Nagia M. Ghanem, Mohamed A. Ismail	2014	An enhanced version of the incremental DBSCAN algorithm was proposed for incrementally building and updating arbitrary shaped clusters in extensive datasets.	The proposed algorithm is additionally proved to perform better in expansive datasets with higher dimensions compared to related algorithms.
Saefia Beri, Kamaljit Kaur	2015	A framework of methodology of DBSCAN algorithm was proposed with the integration of fuzzy logic.	The hybridization will allow DBSCAN to choose the cluster in more efficient manner.
Karlina Khiyarin Nisa, Nisa, Hari Agung Andrianto, Rahmah Mardhiyyah	2014	A web-based application clustering was proposed with DBSCAN algorithm utilizing the R programming language with Shiny framework.	While Eps parameter determination is obtained from k-dist graph observation and the slope difference calculations.

Conclusion

In this work, it is concluded that clustering is the approach of data mining which is applied to cluster similar and dissimilar information. The density based clustering is the approach which cluster information based on the density of the data. The DBSCAN is the algorithm of density based clustering which cluster data based on the cluster radius and distance. In future, approach will be proposed which improve accuracy of DBSCAN algorithm.

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 ISSN: 2277 128X.
- [5] Tapas Kanung, 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, Adebisi EF, Oyelade JO and Doumbia S "Reducing the Time Requirement of K-Means Algorithm" PLoS ONE, Volume 7, Issue 12, 2012.
- [10] Azhar Rauf, Sheeba, Saeed Mahfooz, Shah Khuro 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.
- [11] 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.
- [12] M. N. Vrahatis, B. Boutsinas, P. Alevizos and G. Pavlides, "The New k-Window 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, Xiaoyu Luo, Thomas Fairley Gooch, Ling Tian, Ke Qin, "A Parallel DBSCAN Algorithm Based On Spark", 2016, IEEE, 978-1-5090-3936-4.
- [15] Dianwei Han, Ankit Agrawal, Wei-keng Liao, Alok Choudhary, "A novel scalable DBSCAN algorithm with Spark", 2016, IEEE, 97879-897-99-4.
- [16] 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.
- [17] 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.
- [18] Ilias K. Savvas, & 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] Saefia Beri, Kamaljit Kaur, "Hybrid Framework for DBSCAN Algorithm Using Fuzzy Logic", 2015, IEEE, 978-1-4799-8433-6.
- [21] Karlina Khiyarin Nisa, Hari Agung Andrianto, Rahmah Mardhiyyah, "Hotspot Clustering Using DBSCAN Algorithm and Shiny Web Framework", 2014, IEEE, 978-1-4799-8075-8.