An Inclusive Literature assessment on Multiple View Grouping in Process Mining

Akhil Gupta
School of Electronics and Electrical Engineering
Lovely Professional University
Phagwara, India

Abstract:

This paper provides a rigorous literature study about the multiple view grouping in process mining. This paper has also shown the significant work of renowned researchers in this field.

Keywords: Multiple view grouping, process mining

I. Literature Study

Hao Huang *et al* (2014) proposed in paper [16] that within the data mining methods, the mining of arbitrary shaped clusters is a great challenge. As there is higher time complexity, there are various solutions proposed to solve such major issue. There are various algorithms that try to reduce the size of the dataset which further help in minimizing the computational costs of the systems. However, the clustering performance can be affected with the help of user-defined reduction ratios provided within this method. In this paper, an effective and efficient algorithm known as CLASP is proposed which mines the arbitrary shaped clusters present within the data. This process helps in reducing the overall size of the dataset along with the preservation of the information of the shapes of the clusters. The information is saved within the dataset through the representative data examples. The positions of the representative data examples are modified for improving the internal relationships within this method. This helps in providing more clear and distinct clusters within the clustering technique. Here, a Pk metric is utilized for recognizing the clustering structures within the agglomerative clustering. This is done on the basis of k-nearest neighbours based metrics. There are various experiments performed on the synthetic as well as real datasets. The simulation results achieved show that the newly proposed technique helps in providing effective and efficient mechanisms to solve the problems.

Gunnar Carlsson *et al* (2014) proposed in this paper [17] that the asymmetry of the input data can be preserved through a hierarchical quasi clustering method. This is a generalized hierarchical clustering method which is utilized for the output structure within the asymmetric networks. There is a great similarity between the finite quasi ultra-metric space and the output of the asymmetric network whose admissibility related to the two distinct properties of the system. The only admissible quasi-clustering technique is the enhanced version of the single linkage method. There are various invariance properties that are achieved through this process and the method thus provides stability within the system.

Satoshi Takumi*et a l*(2012)explains [20] algometric algorithms of hierarchal clustering using the asymmetric similarity measures. There are linkage methods proposed into this research are of two methods, first bottom up methods and other is top down methods. The tree diagram structure used to show result of hierarchical clustering called dendrogram result of the hierarchical clustering sometimes shows reversely. This paper gives emphasis to show no reversals in the dendogram. The first method of bottom-up approach does not show reversal in output of algometric hierarchal clustering and another method top-down approach use hypothesis. Example of this is based on real data which show these methods work.

S. R. Pande *et al* (2012) provides [22] the data mining techniques of clustering. Cluster analysis divides data into the groups having similar properties. Clustering is unsupervised classification technique. Clustering is divided into two classes, first is hierarchical clustering techniques and other is partitioning technique. They also density based methods like DBSCAN, DENCLUE. In this paper they process of clustering from the point of view of the data mining.

Ming-Yi Shih *et al* (2010) proposed in this paper [23] that there are various techniques being introduced for clustering the diverse data which is to be stored by the applications within groups. There are two various ways utilized by the clustering algorithms. Either the pure numeric data or the pure categorical data can be performed here on both the mixed categorical as well as the numeric data types within this system. In this paper, a new two-step clustering method is proposed where the items present within the categorical attributes are processed such that relationships amongst them are recognized on the basis of various similar properties. The co-occurrence of various objects is proposed on the basis of similarity amongst the objects. There are various clustering algorithms within the dataset that can be applied to convert the categorical data into numeric. There are various demerits of these already existing technique and for avoiding such issues the two-step method is required that adds attributes to the clusters along with the integrated hierarchical and partitioning clustering algorithms within the system. As per the simulation results achieved it can be seen that the accuracy is improved here and enhanced results are gathered which can help to cluster the mixed numeric and categorical methods.

Jiawei Han J and Kamber M (2012) proposed in this paper [25] that there is a need of execution of huge parallel computer programs to achieve the simulation of complex scientific systems. Across the spatio-temporal space, there are large-scale datasets present which provide simulation programs. In this paper, a simple however effective multivariate clustering algorithm is presented which provides simulations for huge datasets. A linking algorithm is also utilized within this system for connecting the clusters to their appropriate nodes within the dataset of the topology tree. According to the simulation results achieved, the value of our multivariate clustering and linking algorithm determined from the two huge simulation datasets given.

Hui Xiong et.al (2013) proposed in this paper [26] that within the data cleaning process, the removal of noises present within the data is an important task. The analysis of data is interrupted due to the presence of noise within it. So, it is important to remove the noise present here. There are various noises that are present due to the low-level data errors present within the data for the removal of which various algorithms already exist. However, there are various data objects that are irrelevant or weakly relevant to each other which can also degrade the analysis process. If the analysis is to be done at a very higher level, the objects present within the data should also be considered as noise as per the underlying analysis. It is seen through the experimental results that the proposed methods provide the performance of clustering to be enhanced. There are many quality association patterns proposed here which have higher quality. This is due to the removal of noise in higher amount with the help of this technique. There are various other techniques that involve the binary data but have less efficient results as compared to the technique proposed within this paper.

Yu Qian and Kang Zhang (2005) proposed in this paper [27] that it is important to use the visualization techniques for assisting the conventional data mining tasks. A major issue within the visualization process is to select appropriate parameters for spatial data cleaning methods. To resolve this method the performance of visualization technique is enhanced in this paper. Further, the characteristics and properties of the methods and various features of the data are to be presented so that the user can get a feedback regarding its own data. Waterfall, a 3-D visualization model is proposed in this paper for assisting the spatial data cleaning with the four important measures.

Sumit Garg and Arvind K. Sharma (2013) proposed in this paper [28] that the there have been many recent advancements made in the data mining techniques for growing its efficiency. Various new patterns are to be discovered within these huge datasets through this process. There are various algorithms being proposed by the researchers in the recent times for providing enhancements in the techniques. There are variety of data types available and so each of them cannot utilize one single algorithm. There have been varieties of algorithms proposed for various types of datasets present. Therefore the compatibility of the dataset is an equally important factor to be considered such as the end goal to be achieved by the application for choosing an appropriate data mining algorithm. The main objective here is to provide an appropriate algorithm on the educational dataset provided by an application. A comparative analysis has been done and various data mining algorithms have been compared with each other to provide results that would be beneficial during the selection of algorithms.

K. Krishna and Raghu (2001) proposed in this paper [29] that there are various subsets of dataset which satisfy the various criterions for identifying the data from the datasets. This can be done with much ease through the relations that exist among the data present within datasets. The relation identified should be symmetric in nature. For instance, the inclusion relation which is mainly the block of text within the meaning of another block is an asymmetric relation present within the text analysis. The asymmetric data is related to each other through the newly proposed algorithm such that it can be clustered easily. There are two applications which utilize such technique. First is the summarization of short documents and the second is the creation of a hierarchy from the set of documents present within the dataset. The simulation results achieved determine that the performance of this algorithm is efficient than the already existing techniques.

Guangchun Luo, et.al, (2016) proposed in this paper [31] that there has been advancement in the big data which is mainly due to the increased growth of data in all the fields. There are various parallelization methods used for the processing or extracting of data as per the requirements of the user. The cluster analysis is an important task being implemented within the data mining and among all the techniques derived within it, the DBSAN algorithm is the most prominently utilized algorithm. There are various divisions of the database generated into disjoint partitions with the help of the already existing parallel DBSAN algorithm. The data dimensions are also increased here as the splitting and consolidating the high-dimensional space will take more duration. Hus, to solve all such issues arising in the existing algorithm, a parallel DBSAN algorithm known as the S_DBSAAN which utilizes Spark in proposed. The partitions of the original data can be done in a very easy manner through this process. Further, the clustering results are mixed. There has been some data provided on the annual basis on which this proposed algorithm has been applied. It has been seen through the experimental results that there can be an effective and efficient generation of the clusters and the noise data present within the data set can also be recognized.

Dianwei Han, et.al, (2016) proposed in this paper [32] that for the purpose of identifying the arbitrary shaped clusters and eliminating the noise data the DBSCAN clustering algorithm is utilized. On the basis of the MPI or OpenMP environments, the parallelization of DBSCAN algorithm is utilized. There is an absence of fault tolerance within this method. The workload is balanced within this algorithm and the process is enhanced in such prominent manners. There is a need of much experience for providing enhancements within such algorithms for handling the communication amongst the nodes. There have been a lot of applications that have utilized the DBSCAN algorithm for their own needs. So, this algorithm has proven to be more efficient in terms of performances and the experience is also huge. Also, this algorithm has been utilized for detecting the arbitrary shaped clusters and so it is very helpful in providing such efficient results which remove the noise within the data easily. The Spark is utilized within the DBSCAN algorithm for providing enhancement in the DBSCAN algorithm. As per the simulation results achieved, the proposed algorithm has been more efficient in providing required results.

Nagaraju S, et.al, (2016) proposed in this paper [33] that for the detection of embedded and nested adjacent clusters an efficient algorithm has been utilized within the cluster analysis method. This is done on the basis of the density based notion of the clusters as well as the difference between the neighborhood clusters. There has been enhancement made within the already existing DBSCAN algorithm with the help of the global density parameters. This also provides the identification of nested adjacent clusters within the EnDBSCAN algorithm. there are various parameters to be utilized within this proposed method for enhancing the performance of the algorithms. The detection of embedded and nested adjacent clusters is done with the help of density based notion parameters and the difference between the neighbors. It is seen through the experimental results that the enhanced DBSCAN algorithm has provided better results as compared to the earlier provided algorithm. The nested adjacent clusters have provided comparisons between the both of the algorithms. The processes included here do not add any computational complexity within the algorithms and the procedure has provided enhanced results. The global density parameters are also achieved within this paper with the help of sorted k-distance plot and the first order derivative processes.

Jianbing Shen, et.al, (2016) proposed in this paper [34] with the help of DBSCAN algorithm a real-time image super-pixel segmentation method. A faster two-stage framework is proposed in this paper for decreasing the computational costs within the super-pixel algorithms. For the purpose of clustering the pixels at higher rate, the various factors such as color similarity and geometric confinements are used at the principal clustering stage. Further, the neighborhood clusters help in merging the smaller clusters into super-pixels on the basis of the similarity between color and spatial features. For the purpose of achieving better super-pixels within the two mentioned stages, a robust and straightforward distance function is proposed here. As per the experimental results achieved, the proposed algorithm has outperformed the existing algorithm in terms of accuracy and efficiency. The calculation cost within this method is also reduced and the results are also enhanced as compared to the algorithms that have more computational costs. There are also algorithms that have complex objects or texture areas. These algorithms have also been easily calculated and the computational costs have also been less as compared to when the existing DBSCAN algorithms are utilized.

Ilias K. Savvas, et.al, (2016) proposed in this paper [35] that through the computational frameworks and the electronic devices, a large amount of information is being extracted every day. There is a need of new algorithms for managing and extracting all such data from the datasets. For the purpose of allocating and extracting the required data from the data warehouses, various algorithms have been proposed. The most prominently utilized methods here is the clustering process. The clustering of the data according to its characteristics is done with the help of DBSCAN algorithm. The computational complexity within these processes is higher due to which the applications of such algorithms within the large datasets is not possible. There have been numerous enhancements made within the DBSCAN algorithm. However, the changes made have not yet been up to the satisfaction of the researchers and there has been no fixed algorithm that has met all the needs of the researchers. A three phase parallel version of DBSCAN is proposed in this paper. The accuracy, scalability as well as the effectiveness of the results has however been achieved with the help of the algorithm proposed in this paper. The parallel version of the DBSCAN is proposed here and the implementation is done with the help of MPI. Here were similar results achieved within the original sequential technique within this process. However, there has been a reduction in the time complexity within this paper. The performance provided has been improved to huge extent within this paper.

Saefia Beri, et.al, (2015) proposed in this paper [37] that the process of acquiring the required data from the dataset is known as the data mining technique. The important part within this process is also to convert the data achieved into an understandable and meaningful manner for utilizing it further as well. The arbitrary shapes as well as the outliers are recognized with the help of DBSAN algorithm which is based on the bivalent logic. With the help of this algorithm it can be seen that the objects belong to a specific cluster or not. Within this paper, a new DBSAN algorithm is proposed which uses the fuzzy logic method within it. With the help of the membership values present within this algorithm, the degree to which the object belongs to a specific cluster can be determined. There are fuzzy if-then rules utilized within the DBSAN algorithm for hybridizing it. The multivalent logic is utilized for improving the membership values to certain degree. He simulation results achieved have shown improvement as per certain aspects such as bit error rate, specification, sensitivity as well as accuracy he results are also compared with the results achieved through previous algorithms. This technique has been helpful in selecting the cluster in a more appropriate manner.

II. **References**

[1] Rajkumar Buyya, James Broberg, Andrzej Goscinski, “Cloud Computing Principles and [1] Rajkumar Buyya, James Broberg, Andrzej Goscinski, “Cloud Computing Principles and Paradigms”, 2011, John Wiley & Sons, Inc publications

[2] Batagelj,V., Mrvar, A.,andZaversnik,M., “,Partitioning approaches toclusteringin graphs, Pr Drawing’1999, LNCS, 2000, pp. 90-97

[3] Ertoz, L., Steinbach, M., and Kumar, V., “Finding clusters of different sizes, shapes, and densitie dimensional data”, In Proc. of SIAM DM’03.

[4] Ester, M., Kriegel, H.P., Sander,J., and Xu, X., “ A density-based algorithm for discovering clusters databases with noise”, in Proc. of 2nd Int. Conf. on Knowledge Discovery and DataMining(KDD-96),AAAI Press, 1996, pp. 226-231.

[5] Fayyad, U., Piatetsky-Shapiro,G.,Smyth,P., and Uthurusamy,R. (eds.), “A and Data Mining, AAAI/MIT press, 1996.

[6] Fayyad, U. and Grinstein,G.,Information Visualization in Data Mining and Knowledge Discovery, M 2001, pp. 182-190.

[7] Fayyad,U. and Uthurusamy,R., “Evolving data mining intosolutions for insight pp. 28-31.

[8] Han, J., Kamber, M., and Tung, A. K. H., “Spatial clustering methods in (eds.), Geographic Data Mining and Knowledge Discovery, TaylorandFrancis, 2001.

[9] Harel, D.andKoren, Y., “Clustering spatial data using random walks”, In Proc. 7th and Data Mining(KDD-2001),ACM Press, New York, pp. 281-286

[10] K.Rajkumar “Dynamic Web Page Segmentation Based on Detecting Reappearance and Layout of Tag Patterns for Small Screen Devices”,IJSET,2011

[11] Shuang Lin, Jie Chen, Zhendong Niu, “Combining a Segmentation-Like Approach And A Density-Based Approach In Content Extraction” TSINGHUA SCIENCE AND Technologyissnll1007-0214ll05/18llpp256-264 Volume 17, 2012

[12] Yan Gu, “ECON: An Approach to Extract Content from Web News Page” 12th International Asia-Pacific Web Conference,2010

[13] Chaw Su Win, Mie Mie Su Thwin, “Informative Content Extraction By Using Eifce” International Journal Of Scientific & Technology Research Volume 2, Issue 6, 2013

[14] Jan Zeleny, “Web Page Segmentation and Classification” Journal of Data and Knowledge Engineering, 2010

[15] K. S. Kuppusamy, “A Model for Web Page Usage Mining Based on Segmentation” International Journal of Computer Science and Information Technologies, Vol. 2 issue 3, 2011

[16] Hao Huang, Yunjun Gao,_ Kevin Chiew, Lei Chen, Qinming He, “Towards Effective and Efficient Mining of Arbitrary Shaped Clusters” Department of Computer Science and Engineering, Hong Kong University of Science and Technology, China, ICDE Conference 2014

[17] Gunnar Carlsson, et.al, “Hierarchical Quasi-Clustering Methods for Asymmetric Networks”, Proceedings of the 31st International Conference on Machine Learning, Beijing, China, 2014. JMLR:W&CP volume 32, 2014

[18] R.Jensi and Dr.G.Wiselin Jiji, “A Survey On Optimization Approaches To Text Document Clustering”, International Journal on Computational Sciences & Applications (IJCSA) Vol.3, No.6, December 2013

[19] Mahendra Pratap Yadav, Mhd Feeroz and Vinod Kumar Yadav (2012) “Mining the customer behavior using web usage mining In e-commerce” Coimbatore, India. IEEE-201S0

[20] Satoshi Takumi and Sadaaki Miyamoto,”Top-down vs Bottom-up methods of Linkage for Asymmetric Agglomerative Hierarchical Clustering”, 2012 International Conference on granular Computing

[21] Neelamadhab Padhy , Dr. Pragnyaban Mishra and and Rasmita Panigrahi “The Survey of Data Mining Applications And Feature Scope”International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.3, June 2012

[22] S.R.Pande, Ms..S.S.Sambare, V.M.Thakre,"Data Clustering Using Data Mining Techniqes", IJARCCE Vol. 1, issue 8, October 2012

[23] Ming-Yi Shih, Jar-Wen Jheng and Lien-Fu Lai, "A Two-Step Method for Clustering Mixed Categroical and Numeric Data", 2010, Tamkang Journal of Science and Engineering, Vol. 13, No. 1, pp. 11-19

[24] Wilhelmiina Hamalainen, Matti Nykanen (2008) "Efficient discovery of statistically significant association rules", Eighth IEEE International Conference on Data Mining

[25] Jiawei Han J and Kamber M, Data Mining: Concepts and Techniques (3rd ed.). Morgan Kaufmann, San Francisco, CA, 2012.

[26] Hui Xiong, Gaurav Pandey, Michel and Vipun, "Enhancing Data Analysis with Noise Removal", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, Vol. 13, 2013

[27] Yu Qian and Kang Zhang, "The Role of Visualization in Effective DataCleaning", SAC'05,March 13-17,2005,Santa Fe,New Mexico,USA

[28] Sumit Garg and Arvind K. Sharma, "Comparative Analysis of Data Mining Techniques on Educational Dataset", International Journal of Computer Applications (0975 –8887) Volume 74–No.5 , July 2013

[29] K.Krishna and Raghu, "A clustering algorithm for asymmetrically related data with applications to text mining", ACM, New York, USA, 2001

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

[31] 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

[32] Dianwei Han, Ankit Agrawal, Wei−keng Liao, Alok Choudhary," A novel scalable DBSCAN algorithm with Spark", 2016, IEEE, 97879-897-99-4

[33] 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

[34] 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

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

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

[37] Saefia Beri, Kamaljit Kaur," Hybrid Framework for DBSCAN Algorithm Using Fuzzy Logic", 2015, IEEE, 978-1-4799-8433-6

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

[39] Negar Riazifar, Ehsan Saghapour," Retinal Vessel Segmentation Using System Fuzzy and DBSCAN Algorithm", 2015, IEEE, 978-1-4799-8445-9

[40] 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

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