IJFANS INTERNATIONAL JOURNAL OF FOOD AND NUTRITIONAL SCIENCES

ISSN PRINT 2319 1775 Online 2320 7876

Research paper **Volume 11, Iss 4 , 2022**

Pragmatic analysis of heterogeneous high-dimensional data clustering Techniques

N. SreeRam

Department of CSE, Koneru Lakshmaiah Education Foundation (KLEF), Vaddeswaram, Green fields, Guntur, Andhra Pradesh, India -522302

sriramnimmagadda@gmail.com

Dr.M.H.M.Krishna Prasad

Professor, department of CSE, University college of Engineering, JNTUK, Kakainada

krishnaprasad.mhm@gmail.com

DOI : 10.48047/IJFANS/11/ISS4/132

Abstract. Clustering in Data Mining (DM) plays a substantial role in solving problems of data analysis in business and scientific applications. However, it has been a challenging factor in clustering Big-Data (BD) [5], as and then there is a rapid growth in size of datasets in scale of extra in the real world. The efficient way of solving the BD problem is to use a Map-Reduced with a desirable parallel data analysis with a widely used field of data processing. Hadoop provides an environment of cloud and usually used analysis utensil for big data. K-Means and Fuzzy K Means a parallelized Big Data Analysis (BDA) tool in cloud environment. The de-merits of K-Mean parallelized algorithm is very much data sensitive to noisy, sensitive to basic condition and also relate to certain fixed shape, whereas Fuzzy K-Mean clustering has a basic issue related to computing and processing time, and also is much complex than K-Means, our work presents an empirical study on present clustering methods which are used in BDA [4]. Empirical analyses of the present techniques are carried out with the appropriate methods of clustering and have been studied

Keywords: clustering, bigdata, map reduce, K-Means and FUZZY-K means.

1. Introduction

In recent day collection of data in various form is been increasing rapidly in various sizes and forms based on variability and variety. BD sets are defined as an ability of which its size is beyond the size of the database irrelative to its structure which uses certain tool to capture, manage. Store and analyses. The most common way of defining BD is via 3 Vs as Velocity of data, volumes of data and variety or variability of data. As and then the volumes of data are growing day by day irrespective of size in terms of Peta or tera bytes based on storage, table transaction and files movement. Velocity of data states at what frequency and a rapid rate the application data is streamed based on real time application and domain. Structured (RDBMS), semi-structured (RSS feeds, XML), and unstructured (text, human languages) are among the various data kinds. The representation of sources, entities, and data types in BD is astoundingly varied. Over time, data analytics (DA) has evolved significantly, with big data analytics (BDA) being the most recent development. BDA is the process of analysing huge amount of data having different variety, velocity and volume to uncover hidden patterns, correlation and other useful information. Such information results in business advantages, such as growth revenue and effective marketing. BDA helps enterprises to make business decisions better and other users to handle huge amount of transaction data in a better way which is left by Business Intelligence (BI) programs. BD can be analysts with advanced analytic tools such as DM and predictive analytics. But big data collected from unstructured data sources is not fit in the traditional data warehouses which lead to new big data technology. These technologies

ISSN PRINT 2319 1775 Online 2320 7876
Research paper **Volume 11, Iss 4 , 2022**

associated with BDA includes NoSQL, Hadoop and MapReduce databases. Across the clustered systems with a large dataset can be processed from these technologies.

Clustering analysis, also known simply as clustering, is a fundamental technique in the field of data mining and machine learning. It involves the process of grouping similar data points together into clusters while keeping dissimilar data points in separate clusters. The primary goal of clustering is to uncover the underlying structure or patterns in a dataset without prior knowledge of the groupings or labels. Both little and large datasets can be clustered using conventional approaches like K-means, FUZZY-K MEANS, and so forth, cannot be used directly to analyze large amounts of data in a cloud environment [10] since large data is gathered from many sources and is in a variety of Since these categories overlap and may contain characteristics from other categories, clustering approaches are challenging to classify precisely. Density-based, partition-based, and hierarchical clustering methods are the three main categories of clustering algorithms. There are two subcategories of the hierarchical clustering method: agglomerative and divisive approaches. Figure 1 shows the different clustering algorithms used now days for analysing BD forms. Thus, the extension and parallel version of traditional algorithms are required to be designed for analysing BD in cloud environment.

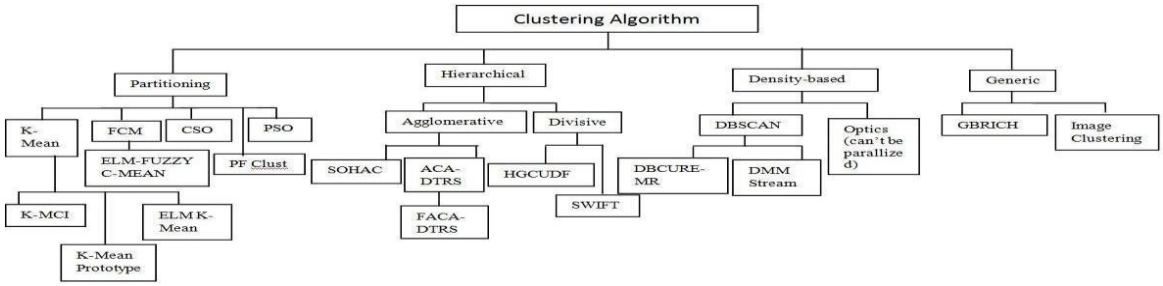

Figure 1 Evolution of Clustering Techniques

A. Partitioning Clustering Techniques

For a single hidden feed forward neural network, Extreme Learning Machine (ELM) K-means and ELM NMF (Non-negative Matrix Factorization) [15]. They used ELM techniques for clustering problem as it yields better result than MercerKernel based method and traditional algorithms. They used three different datasets; two from university machine monument and third was from document corpus. ELM NMF and ELM K-means both are stable for huge datasets. They are effective and simple to implement. The author mentions that the numerical data are clustered by incremental k-means algorithm. The categorical data are clustered by modified k-modes, while mixed data are clustered by k prototype is used. These three algorithms reduce the cost of clustering function and they are efficient. As the iteration converges the complexity is decreased and un similarity index is also low-slung. Ganesh, Krishnasamy et al. stated a Modified Cohort Intelligence (MCI) by implementing mutation manipulator into Cohort Intelligence (CI) to resolve the issue of convergence speed and quality. The hybrid K-MCI algorithm is a combined approach of K-Means and MCI for data clustering. This algorithm is dependable, well-organized, offers good quality of results It outperforms other clustering techniques in terms of convergence speed. They have taken six data sets from UCI Machine Learning monument for their performance testing. Ishak Saida, Nadjet and Omar proposed a Cuckoo Search Optimization (CSO) which is a new meta-heuristic approach to

ISSN PRINT 2319 1775 Online 2320 7876
Research paper **Volume 11, Iss 4 , 2022**

avoid difficult of k-means for data clustering. The significant feature of this optimization search is that it has good performance and it is easy to carry out. The new algorithm increases the new method's ability to find the optimal values. Four data sets from UCI Machine Learning Monument were used in the experiment. XueFeng Jiang has given a global optimization algorithm for large scale computational problems. This proposed algorithm by author is a type of swarm optimization which is based on annealing parallel clustering method. It is new method based on algorithm of group and is much effective for variable of continuous problems; it also has the ability to grain parallelism. SOA method has a parallel practical which has very less time of computation and will provide high quality in clustering and is very much improved compared to other methods. The higher the effectiveness of the method evaluated on huge datasets, will provide higher the flexibility and accurate on clustering. Musaveva. K proposed a method of clustering name PF cluster which is used to find the number of optimal clusters without having the knowledge prior to the number of clusters. PF clustering will depend on matrix similarity, which immune the issues generated by bigger dimensional datasets. In this case the time of execution of PF cluster is much better in gaining results of better quality. its performances are also measure on high dimensional data and datasets.

B. Hierarchical Clustering Techniques

Zhanguo Hong, Liu presented an automatic effective method in dealing the problem in determine the cluster number and types of clusters. This work is an extension of theoretic decision of rough set method and has been designed based on risk calculation on possibility and loss functions. Various authors has presented an hierarchical algorithm of clustering call ACADTRS [14], Which automatically used to find the number of perfect clusters? They also proposed DTRS-FACA is faster in response time of DTRS-ACA in terms of time complexity, but both are effective in terms of time cost. These two algorithms performance is evaluated on real world synthetic datasets. Nanopolous and Buza proposed a reduced tick storage data method which will increase the performance rapidly. The tick data has been decomposed into data of smaller matrix using cluster attributes using new algorithm of cluster namely SOHAC (Storage-Optimized Hierarchical Agglomerative Clustering). Through this technique, the queries can be executed effectively. To speed up the runtime they also presented a Quick SOHAC approach. This algorithm is applied to high dimensional tick data because of the lower bounding technique. Three real world data sets given by investment bank are used to evaluate the performance of these algorithm. Wang Shuliang et al. have proposed a new clustering named as Hierarchical Grid Clustering Using Data Field (HGCUDF). In this method, the grids divide large data sets into their hierarchical subsets, search scope is limited to clustering centers and data field generated from the data space is minimized. It is stable, rapid and improves clustering performance on vast computerized datasets. Naim et al. have given Scalable Weighted Iterative Flow-clustering Technique (SWIFT) a model-based clustering method for high-dimensional large-sized datasets [8]. It works in three phases comprising of multimodality splitting, iterative weighted sampling and uni-modality preserving merging. This algorithm is basically to find rare cell population and for flow cytometry. When tested over synthetic datasets, this approach resolves small and scales large datasets effectively as compared to traditional approaches. It scales well for very large FC datasets and it is beneficial for task typical in immune.

C. Density-based Clustering Techniques

Emini et al. proposed DMM stream clustering method [9] in solving real time data stream. It is a density-based clustering method for evaluating data streams. This clustering method using micro and mini cluster which is very much like micro cluster with a small radius of introducing minute and smaller units. A method of distance mahaleb is proposed in determining the number of clusters based of certain cluster purity and quality, which is defined on time complexity for

ISSN PRINT 2319 1775 Online 2320 7876
Research paper **Volume 11, Iss 4 , 2022**

evaluation. They also propose a strategy for noise filter reduction on real-data. Various types of experiments are conducted and performed on the real time synthetic datasets. Kim et al has proposed a parallelized algorithm for clustering as DBCURE [11] for clustering BD, which is based on density of cluster. The proposed method is too robust in finding the clusters with various type of varying densities and will reduce the content using Map Reduced parallelism. The above author suggested MRDBCURE method which is much suitable for framework of map reduced in finding various types of parallel clusters. Both the proposed algorithms are much efficient in identifying the number of clusters and the accuracy in identification of datasets. These methods may not be sensitive for different densities of clusters, which may not perform well on Map-reduced framework. Experimental study is don on synthetic datasets such as Window, Butterfly and Clover and on other live and real time datasets.

D. D. Generic Clustering Techniques

Tsai & Chun-Wei proposed a method for solving the high-performance issues on search technique related to DM. This method proposed is an optimized spiral method which segregates the population datasets into small units for increasing the diversity process in searching of clusters or clustering. This method is distributed, and its performance is extended using K-means clustering with oscillation technique. Optimized Spiral Novel method is quite more promising and is very much based on swirl phenomena and low pressure. The output gain shows the similarity between spiral optimization and k-means genetic clustering, Murthy & Suri proposed a rank-based clustering algorithm [3] in detection of categorical outliers in the data. It is a two phased method, here the cluster data is first exposed, next the rank is provided to identify the outliers set similarity. This method is implemented in two different types of ranks, which is based on frequency value of cluster inherent and also enhance the performance of various types of outliers. In proving the algorithm effectiveness, various types of experiment are conducted using categorical public domain set and its performance. Another type of benefit of this method is for computing the complexity and its effective number of outlier detection. Jiang. W et.al proposed and suggested a Binary Matrix Factorization [2] which reduces the high dimensional dataset into its binary equivalent. It is presented in two variants of CBMF and is related to other dimensional reduction model. In solving the Binary Linear Programming problem, an alternate update method is processed. An Effective approximation-2 method is proposed in exploiting the relationship among BLP sub-program and cluster. He also suggested and proposed randomized algorithms in obtaining the accuracy in solution, using this method various types of experiments are conducted on high dimensional datasets. Heish ch Liang et al. proposed a method in improving the response time in image search process by using Map-reduced image-based clustering [13]. The clustering methods on images are used in dealing with efficiency and scalable problems.

2. Comparative Analysis of Existing Clustering Techniques

We come across various types of techniques which are recently implemented for analysis on big data. These developed algorithms of clustering are verified on the metric parameters of scalability, speedup, accuracy, quality of clusters. A comparison related to clustering method based on merits and demerits are listed below in a tabular form.

s.no	Method	Addressing of Issue	Clustering Type	Dataset used category	Time of execution	Cluster Quality	Advantage	Disadvantage

ISSN PRINT 2319 1775 Online 2320 7876

Research paper **Volume 11, Iss 4 , 2022**

1	ELM-K-Means and NMFELM	It will solve the basic problems of clusters using ELM feature of k-means and fuzzy c means	It belongs to Partitioning	It uses UCI dataset s which is related to Machine learning Repository and corpus of cluster document	Very good	good	Features of ELM Are very easy for implementation and K-mean and ELM will give better results of output than Kernel based Mercer method	Required node count should be > than 350 and should be produce optimal performance
2	K-prototype and k-modes method	K-Means algorithm is an extension version of K-prototype and kmean	It is of Partitioning stage of clustering	It is a combination of categorical and numerical data	Good	Good	K-mean and K-prototype Combined method will give better comparative results than k-mean and functional cost reduction method	It does not provide quality of clustering process
3	NSO	This method will perform better and will solve the issues related to search process in DM mining	Generic	It uses real time datasets	Good	Good	It is a promising method based on real time problem phenomena and is very easy to implement nt	Result are in the form of generic, but improvement has to be for better performance and assistance
4	Ranking based algorithm	addresses the problem of anomalies. It operates using two techniques: intrinsic clustering and value approaches	Generic	Public do-main categorical dataset	Good	Very good	Effective to find out different number of outliers and computational complexity are not affected by outliers	It can be applied to categorical datasets only
5	Clustering approaches to constrained Binary Matrix Factorization (cBMF)	Deals with reduction of dimensionality in high dimensional data	Generic	Large high dimensional datasets	Very good	Very good	Results are accurate and provided in better time	NA

ISSN PRINT 2319 1775 Online 2320 7876
Research paper **Volume 11, Iss 4 , 2022**

6	DMM-Stream	addresses the problem of streaming data in real time. It makes use of the notion of mini-micro clusters	Density based	Real and Synthetic Datasets	Very good	Very good	Ascertain the appropriate number of higher-quality clusters while preserving temporal complexity. reduces noise in the data	Implementing it on all real-world datasets is a little complicated.
7	Clustering based on Cuckoo search optimization (CSO)	It is used to find the optimal or near-optimal solutions to optimization problems	CSO can help optimize the partitioning of data points into clusters, seeking to find the best clustering solution based on certain criteria	From UCI machine learning repositories datasets	Very good	good	Its computing performance is good and its implementation is simple	NA
8	DBCUREMR	It deals with problem of big data clustering with varying densities and parallelized with Map Reduce	Density based	Synthetic data and real time data	Very good	Very good	It is easy to parallelize	Much computational complexity
9	ACA-DTRS and FACA-DTRS	DTRS extension that automatically determines the number of clusters	Hierarchical	Synthetic data and real time data	Very good	Very good	Without human intervention, it accurately counts the number of clusters without sacrificing function quality. expedite execution time as well	It cannot work for bound ary region
10	SOHAC		Hierarchical	Three real	Very	Good	Queries can	Limited to tick

ISSN PRINT 2319 1775 Online 2320 7876
Research paper **Volume 11, Iss 4 , 2022**

		It deals with size of tick data which is growing rapidly		world da-tasets by investment bank	good		efficiently run. Clusters can be found in significant time	data only
11	HGCUDF	minimizes data space and narrows the search scope for hierarchical grids by using a divide and conquer strategy	Hierarchical	Vast com puterised datasets	Very good	Good	It can be applying on parallel plat-forms and speedup spatial data mining	NA
12	SWIFT	This clustering method is used in dealing with huge dimensional datasets	Hierarchical	It uses syntheticand FC large datasets	Good	Very good	It can detect the datasets of population category	It is based on task limited

Table 1. Analysis of comparison of various clustering method applied on Bigdata

3. Conclusions

Our paper presents a comparative analysis based on literature survey of the available techniques of clustering done on heterogeneous high dimensional data. Clustering method used for BDA are compared based on demerits and merits which are presented with results in a tabular form. At present big data has posted a buzz in the social, economic and other related application. Among them has created a challenging aspect in BDA in the aspect of designing and developing new methods for clustering the big data sets. Clustering methods are used in analysing large dataset from which similar object clusters are gathers which will help in solution the problems of weather, business and forecasting etc….

ISSN PRINT 2319 1775 Online 2320 7876
Research paper Volume 11, Iss 4 , 2022

References

1. Chun-Wei Tsa, Bo-Chi Huang,andMing-Chao Chiang "A Novel Spiral Optimization for Clustering" Mobile, Ubiquitous, and Intelligent Computing pp 621-628 DOI: 10.1007/978-3-642-40675-1_92, © Springer-Verlag Berlin Heidelberg 2014
2. Peng Jiang, Jiming Peng, Michael Heath, and Rui Yang "A Clustering Approach to Constrained Binary Matrix Factorization"Data Mining and Knowledge Discovery for Big Da- ta pp 281-303 ISBN: 978-3-642-40836-6, Springer, 2014
3. N.N.R.Ranga Suri, M.Narasimha Murthy and G.Athithan "A ranking-based algorithm for detection of outliers in categorical data" International Journal of Hybrid Intelligent Systems, Volume 11 Issue 1,January2014 Pages 1-11
4. Uranus Kazemi "Clustering methods in Big data" Journal of Embedded Systems and Processing Volume 3 Issue 2 December 2017.
5. AvitaKatal, Mohammad Wazid, and RH Goudar. (2013). Big data: Issues, challenges, tools and good practices, In Contemporary Computing(IC3), Sixth International Conference on IEEE , pp. 404-409. 2
6. Tingting Hu, Haishan Chen, Lu Huang, and Xiaodan Zhu. (2012). A survey of mass data mining based on cloud-computing, In AntiCounterfeiting, Security and Identi_cation (ASID) International Conference, pp. 1-4.
7. Jun, S., Park, S.-S., Jang, D.-S.: Document clustering method using dimension reduction and support vector clustering to overcome sparseness. Expert Syst. Appl. 41, 3204–3212 (2014)
8. Naim I, Datta S, Rebhahn J, Cavenaugh JS, Mosmann TR, Sharma G. "SWIFT-scalable clustering for automated identification of rare cell populations in large, high-dimensional flow cytometry datasets" Conference Paper (PDF Available) Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on · March 2010.
9. AminehAmini, HadiSaboohi, Teh Ying Wah, and TututHerawan.(2014). Dmm-stream: A density mini-micro clustering algorithm for evolving datastreams. In Proceedings of the First International Conference on AdvancedData and Information Engineering (DaEng2013), pp. 675682.
10. Seema Maitreya,C.K.Jhab "MapReduce: Simplified Data Analysis of Big Data" Procedia Computer Science 57 (2015) 563 – 571 Elsevier
11. Younghoon Kim, Kyuseok Shim,Min-SoengKim, June Sup Lee "DBCURE-MR: An efficient density-based clustering algorithm for large data using MapReduce" Journal Information Systems Volume 42, June, 2014 Pages 15-35