International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 5 Issue: 6 1294 – 1299

An Analysis of Clustering Algorithms for Big Data

B. Bikku
Assistant Professor in Computer science,
SDLCE, Kakatiya University,
Warangal, Telangana, India.
bhukyabikku9@gmail.com

P. Praveen
Research Scholar & Assistant professor,
Department of Computer Science,
SR Engineering College,
Warangal, Telangana, India
prawin1731@gmail.com

Abstract: Clustering is an important data mining and tool for reading big records. There are difficulties for making use of clustering strategies to huge data duo to new challenges which might be raised with massive records. As large information is relating to terabytes and peta bytes of information and clustering algorithms are come with excessive computational costs, the question is the way to take care of with this hassle and how to install clustering techniques to big information and get the outcomes in a reasonable time. This study is aimed to review the style and progress of agglomeration algorithms to cope with massive knowledge challenges from first projected algorithms until modern novel solutions. The algorithms and the centered demanding situations for generating stepped forward clustering algorithms are introduced and analyzed, and later on the viable future path for extra superior algorithms are based on computational complexity. In this paper we discuss clustering algorithms and big data applications for real world things.

Index terms*: Big Data, Clustering Algorithms, Computational complexity, Partition based Algorithms, Hierarchical Algorithms*

*****

I. INTRODUCTION

In the current digital era, in line with (as far) huge progress and development of the net and on-line world technologies like massive and powerful knowledge servers, we tend to face a large volume {of information of knowledge|of knowledge} and data day by day from many various resources and services that weren't accessible to group simply some decades past. huge quantities of information square measure created by and concerning individuals, things, and their interactions. various teams argue concerning the potential edges and prices of analyzing data from Twitter, Google, Verizon, 23andMe, Facebook, Wikipedia, and each area wherever massive teams of individuals leave digital traces and deposit knowledge[2]. This knowledge comes from accessible totally different on-line resources and services that are established to serve their customers. Services and resources like sensing element Networks, Cloud Storages, Social Networks and etc., turn out massive volume {of knowledge|of knowledge|of information} and additionally have to be compelled to manage and utilize that data or some analytical aspects of the information. though this huge volume of information will be very helpful for individuals and firms, it will be problematic similarly. Therefore, {a massive|an enormous|a giant} volume {of knowledge|of knowledge|of information} or big data has its own deficiencies similarly. they have massive storages and this volume makes operations like analytical operations, method operations, retrieval operations, terribly tough and massively time overwhelming. a way to beat these tough issues is to possess massive knowledge clustered during a exceedingly|in a very} compact format that's still an informative version of the whole knowledge. Such cluster techniques aim to supply an honest quality of clusters/summaries. Therefore, they would hugely benefit everyone from ordinary users to researchers and people in the corporate world, as they could provide an efficient tool to deal with large data such as critical systems (to detect cyber attacks)[6].

The main goal of this paper is to provide readers with a proper analysis of the different classes of available clustering techniques for big data by experimentally comparing them on real big data. The paper does not refer to simulation tools. However, it specifically looks at the use and implementation of an efficient algorithm from each class. It also provides experimental results from a variety of big datasets. Some aspects need careful attention when dealing with big data, and this work will therefore help researchers as well as practitioners in selecting techniques and algorithms that are suitable for big data[8]. [Math Processing Error] olume of data is the first and obvious important characteristic to deal with when clustering big data compared to conventional data clustering, as this requires substantial changes in the architecture of storage systems. The other important characteristic of big data is [Math Processing Error] elocity. This requirement leads to a high demand for online processing of data, where processing speed is required to deal with the data flows. [Math Processing Error] ariety is the third characteristic, where different data types, such as text, image, and video, are produced from various sources, such as sensors, mobile phones, etc. These three V (Volume, Velocity, and Variety) are the core characteristics of big data which must be

taken into account when selecting appropriate clustering techniques[7].

Despite a vast number of surveys for clustering algorithms available in the literature [1] and [8] for various domains (such as machine learning, data mining, information retrieval, pattern recognition, bio-informatics and semantic ontology), it is difficult for users to decide a priori which algorithm would be the most appropriate for a given big dataset. This is because of some of the limitations in existing surveys: (i) the characteristics of the algorithms are not well studied; (ii) the field has produced many new algorithms, which were not considered in these surveys; and (iii) no rigorous empirical analysis has been carried out to ascertain the benefit of one algorithm over another. Motivated by these reasons, this paper attempts to review the field of clustering algorithms and achieves the following objectives:

- To propose a categorizing framework that systematically groups a collection of existing clustering algorithms into categories and compares their advantages and drawbacks from a theoretical point of view.

- To present a complete taxonomy of the clustering evaluation measurements to be used for empirical study.

- To make an empirical study analyzing the most representative algorithm of each category with respect to both theoretical and empirical perspectives.

Therefore, the survey presents taxonomy of clustering algorithms and Big data applications framework that covers major factors in the selection of a suitable algorithm for big data.. The rest of this paper is organized as follows. Section II provides a review of clustering algorithms categories. We group and compare different clustering algorithms based on computational Section II introduces the taxonomy of clustering evaluation measurements[3].

As there are so many clustering algorithms, this section introduces a categorizing framework that groups the various clustering algorithms found in the literature into distinct categories. The proposed categorization framework is developed from an algorithm designer's perspective that focuses on the technical details of the general procedures of the clustering process. Accordingly, the processes of different clustering algorithms can be broadly classified follows[4]

A. Partitioning-based: In such algorithms, all clusters ar determined promptly. Initial teams are given and reallocated towards a union. In different words, the partitioning algorithms divide knowledge objects into variety of partitions, wherever every partition represents a cluster. These clusters ought to fulfill the subsequent requirements: (1) every cluster should contain a minimum of one object, and (2) every object should belong to precisely one cluster. within the K-means formula, as an example, a middle is that the average of all points and coordinates representing the expectation. Within the K-medoids formula, objects that ar close to the middle represent the clusters. There ar several different partitioning algorithms like K-modes, PAM, CLARA, CLARANS and FCM [7].

B. Hierarchical-based: Data area unit organized during a stratified manner betting on the medium of proximity. Proximities area unit obtained by the intermediate nodes. A dendrogram represents the datasets, wherever individual knowledge is conferred by leaf nodes. The initial cluster step by step divides into many clusters because the hierarchy continues. stratified bunch ways is clustered (bottom-up) or discordant (top-down). AN clustered bunch starts with one object for every cluster and recursively merges 2 or additional of the foremost acceptable clusters. A discordant bunch starts with the dataset mutually cluster and recursively splits the foremost acceptable cluster. the method continues till a stopping criterion is reached (frequently, the requested variety [Math process Error] of clusters). The stratified methodology includes a major disadvantage although, that relates to the very fact that after a step (merge or split) is performed, this can't be undone. BIRCH, CURE, ROCK and Chameleon area unit a number of the well-known algorithms of this class[11].

C. Density-based: Here, information objects area unit separated supported their regions of density, property and boundary. they're closely associated with point-nearest neighbours. A cluster, outlined as a connected dense element, grows in any direction that density results in. Therefore, density-based algorithms area unit capable of discovering clusters of whimsical shapes. Also, this provides a natural protection against outliers. so the general density of a degree is analyzed to work out the functions of datasets that influence a selected datum. DBSCAN, OPTICS, DBCLASD and DENCLUE area unit algorithms that use such a way to filtrate noise (ouliers) and see clusters of whimsical form[10].

D. Grid-based: The house of the information objects is split into grids. the most advantage of this approach is its quick time interval, as a result of it goes through the dataset once to cypher the applied mathematics values for the grids. The accumulated grid-data create grid-based clump techniques freelance of the amount {of information|of knowledge|of information} objects that use a homogenous grid to gather regional applied mathematics data, then perform the clump on the grid, rather than the information directly[5]. The performance of a grid-based methodology depends on the scale of the grid, that is typically abundant but the scale of the information. However, for extremely irregular information distributions, employing a single uniform grid might not be

comfortable to get the desired clump quality or fulfill the time demand. Wave-Cluster and STING area unit typical samples of this class[9][14].

E. Model-based: Such a way optimizes the work between the given information and a few (predefined) mathematical model. it's supported the belief that the information is generated by a mix of underlying likelihood distributions. Also, it results in how of mechanically decisive the amount of clusters supported commonplace statistics, taking noise (outliers) under consideration and so yielding a sturdy clump methodology. the model-based method: applied mathematics and neural network approaches[10]. MCLUST is perhaps the known model-based rule, however there area unit different smart algorithms, like EM (which uses a mix density model), abstract clump (such as COBWEB), and neural network approaches (such as self-organizing feature maps). The applied mathematics approach uses likelihood measures in decisive the ideas or clusters. Probabilistic descriptions area unit usually wont to represent every derived idea. The neural network approach uses a group of connected input/output units, wherever every affiliation features a weight related to it. Neural networks have many properties that create them common for clump. First, neural networks area unit inherently parallel and distributed process architectures. Second, neural networks learn by adjusting their interconnection weights therefore on best work the information. this permits them to normalize or epitome. Patterns act as options (or attributes) extractors for the assorted clusters. Third, neural networks method numerical vectors and need object patterns to be portrayed by quantitative options solely. several clump tasks handle solely numerical information or will remodel their information into quantitative options if required. [11][12].

II. BIG DATA

Laney [5], first proposed three dimensions Volume, Velocity and Variety (3Vs) that distinguishes the opportunities and challenges of increasing huge data volumes. These 3Vs have been generally used to depict big data. The another new dimension called veracity is added along with 3Vs to depict data excellence and integrity. Further Vs are also been proposed like validity, volatility, variability, value, visibility and visualization. But, the quality of the data can be determined without the necessity of these Vs and these further dimensions of Vs are not useful to understand the ‘‘big” of big data directly, but these Vs clarify concepts of data collection, processing and presentation as a operational sequence of big data[6].

The big data environment works based on cloud computing technique which provides the shared pool of services by distributed computing resources which is convenient for different applications with simple management effort [15]. Bayer et al[1] explained about the importance of Big data with characteristics and way of processing for process optimization enhanced decision making and insight discovery. Hadoop is designed to provide a reliable, distributed storage and analysis environment for the user community. Dittrich et al [3] explained about layouts and indexes of several data management techniques, starting from job optimization to physical data management for efficient data processing in Hadoop Mapreduce[12].

When evaluating clustering methods for big data, specific criteria need to be used to evaluate the relative strengths and weaknesses of every algorithm with respect to the three-dimensional properties of big data, including *Volume*, *Velocity*, and *Variety*. In this section, we define such properties and compiled the key criterion of each property[16].

- **Volume** refers to the ability of a clustering algorithm to deal with a large amount of data. To guide the selection of a suitable clustering algorithm with respect to the *Volume* property, the following criteria are considered:

  (i) size of the dataset,
  (ii) handling high dimensionality and
  (iii) handling outliers/ noisy data.

- **Variety** refers to the ability of a clustering algorithm to handle different types of data (numerical, categorical and hierarchical). To guide the selection of a suitable clustering algorithm with respect to the *Variety* property, the following criteria are considered: (i) type of dataset and (ii) clusters shape[12].

- **Velocity** refers to the speed of a clustering algorithm on big data. To guide the selection of a suitable clustering algorithm with respect to the *Velocity* property, the following criteria are considered: (i) complexity of algorithm and (ii) the run time performance[18].

 ISSN: 2321-8169
1294 – 1299

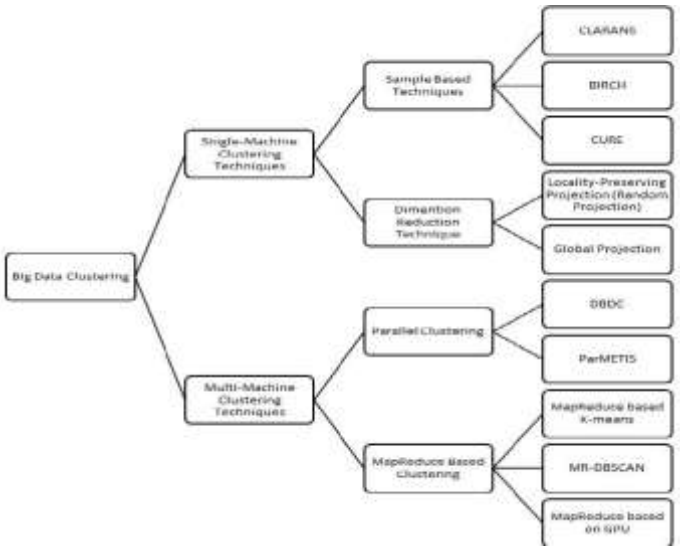

Fig: Taxonomy of Clustering Algorithms for Big Data

In what follows, we explain in detail the corresponding criterion of each property of big data:

1. ***Type Of Dataset:*** Most of the traditional clustering algorithms are designed to focus either on numeric data or on categorical data. The collected data in the real world often contain both numeric and categorical attributes. It is difficult for applying traditional clustering algorithm directly into these kinds of data. Clustering algorithms work effectively either on purely numeric data or on purely categorical data; most of them perform poorly on mixed categorical and numerical data types.

2. ***Size Of Dataset:*** The size of the dataset has a major effect on the clustering quality. Some clustering methods are more efficient clustering methods than others when the data size is small, and vice versa.

3. ***Input Parameter****:* A desirable feature for "practical" clustering is the one that has fewer parameters, since a large number of parameters may affect cluster quality because they will depend on the values of the parameters.

4. ***Handling Outliers/Noisy Data:*** A successful algorithm will often be able to handle outlier/noisy data because of the fact that the data in most of the real applications are not pure. Also, noise makes it difficult for an algorithm to cluster an object into a suitable cluster. This therefore affects the results provided by the algorithm.

5. ***Time Complexity:*** Most of the clustering methods must be used several times to improve the clustering quality. Therefore if the process takes too long, then it can become impractical for applications that handle big data.

6. ***Stability****:* One of the important features for any clustering algorithm is the ability to generate the same partition of the data irrespective of the order in which the patterns are presented to the algorithm.

7. ***Handling High Dimensionality****:* This is particularly important feature in cluster analysis because many applications require the analysis of objects containing a large number of features (dimensions). For example, text documents may contain thousands of terms or keywords as features. It is challenging due to the curse of dimensionality. Many dimensions may not be relevant. As the number of dimensions increases, the data become increasingly sparse, so that the distance measurement between pairs of points becomes meaningless and the average density of points anywhere in the data is likely to be low.

8. ***Cluster Shape:*** A good clustering algorithm should be able to handle real data and their wide variety of data types, which will produce clusters of arbitrary shape[11].

III. BIG DATA APPLICATIONS AND COMPUTATIONAL COMPLEXITY

1) Banking

With large amounts of information streaming in from countless sources, banks are faced with finding new and innovative ways to manage big data. While it's important to understand customers and boost their satisfaction, it's equally important to minimize risk and fraud while maintaining regulatory compliance. Big data brings big insights, but it also requires financial institutions to stay one step ahead of the game with advanced analytics[17].

2) Education

Educators armed with data-driven insight can make a significant impact on school systems, students and curriculums. By analyzing big data, they can identify at-risk students, make sure students are making adequate progress, and can implement a better system for evaluation and support of teachers and principals[12].

3) Government

When government agencies are able to harness and apply analytics to their big data, they gain significant ground when it comes to managing utilities, running agencies, dealing with traffic congestion or preventing crime. But while there are

many advantages to big data, governments must also address issues of transparency and privacy.

4) ***Health Care***

Patient records. Treatment plans. Prescription information. When it comes to health care, everything needs to be done quickly, accurately – and, in some cases, with enough transparency to satisfy stringent industry regulations. When big data is managed effectively, health care providers can uncover hidden insights that improve patient care.

5) ***Manufacturing***

Armed with insight that big data can provide, manufacturers can boost quality and output while minimizing waste – processes that are key in today's highly competitive market. More and more manufacturers are working in an analytics-based culture, which means they can solve problems faster and make more agile business decisions.

6) ***Retail***

Customer relationship building is critical to the retail industry and the best way to manage that is to manage big data. Retailers need to know the best way to market to customers, the most effective way to handle transactions, and the most strategic way to bring back lapsed business. Big data remains at the heart of all those thing.

(I=no. of iterations, k=no. of clusters, n=no. of objects)

Method name	Algorithm	Time Complexity	Dataset size	Dataset type	Cluster shape	Handl outlie
Partitioning	K-means	O(Ikn)	Huge	Numeric	Spherical	No
	k-medoids	$O(n^2I)$	Small	Categorical	Spherical	Yes
	k-prototype	O(n)	Small	Numeric & categorical	Spherical	No
Hierarchical	BIRCH	O(n)	Huge	Numeric	Spherical	Yes
	CURE	$O(n^2 \log n)$	Huge	Numeric	Arbitrary	Yes
	CHAMELEON	$O(n^2)$	Huge	All type of data	Arbitrary	Yes
Density based	DBSCAN	O(n log n)	Huge	Numeric	Arbitrary	Yes
	OPTICS	O(n log n)	Huge	Numeric	Arbitrary	Yes
	DENCLUE	O(n log n)	Huge	Numeric	Arbitrary	Yes
Grid based	STING	O(n)	Huge	Special	Arbitrary	Yes
	Wave Cluster	O(n)	Huge	Special	Arbitrary	Yes
Model based	EM	O(n)	Huge	Special	Spherical	No
	SOM	$O(n^2m)$	Small	Multivariate	Spherical	No

Table: Clustering Analysis of Time complexities

In above Table is represents the clustering algorithms to find computational complexity and check for whether it is suitable for small or large data set at the same time it will check handling outliers.

IV. CONCLUSION

This survey provided a comprehensive study of the clustering algorithms proposed in the literature. In order to reveal future directions for developing new algorithms and to guide the selection of algorithms for big data, This paper analyzed different clustering algorithms required for processing Big Data. The study revealed that to identify the outliers in large data sets and to analyze big data even future clustering algorithms could be incorporated into the framework according to the computational complexities . Furthermore, the most representative clustering algorithms of each category have been empirically analyzed over a vast number of evaluation metrics and traffic datasets.

III. REFERENCES

[1] A. Abbasi, M. Younis, "A survey on clustering algorithms for wireless sensor networks", Comput. Commun., vol. 30, no. 14, pp. 2826-2841, Oct. 2007.

[2] C. C. Aggarwal, C. Zhai, "A survey of text clustering algorithms", Mining Text Data., pp. 77-128, 2012.IK. Elissa, "Title of paper if known," unpublished.

[3] A. Almalawi, Z. Tari, A. Fahad, I. Khalil, "A framework for improving the accuracy of unsupervised intrusion detection for SCADA systems", Proc. 12th IEEE Int. Conf. Trust Security Privacy Comput. Commun. (TrustCom), pp. 292-301, Jul. 2013.

[4] . Almalawi, Z. Tari, I. Khalil, A. Fahad, "SCADAVT-A framework for SCADA security testbed based on virtualization technology", Proc. IEEE 38th Conf. Local Comput. Netw. (LCN), pp. 639-646, Oct. 2013.

[5] M. Ankerst, M. M. Breunig, H.-P. Kriegel, J. Sander, "Optics: Ordering points to identify the clustering structure", Proc. ACM SIGMOD Rec., vol. 28, no. 2, pp. 49-60, 1999.

[6] J. Brank, M. Grobelnik, D. Mladenić, "A survey of ontology evaluation techniques", Proc. Conf. Data Mining Data Warehouses (SiKDD), 2005.

[7] P.Praveen,B.Rama," A Novel Approach to Improve the Performance of Divisive Clustering-BST" Third Springer International Conference on Computer & Communication Technologies (IC3T 2016), DOI,10.1007/978-981-10-3223-3_53.

[8] P.Praveen, B. Rama, ,Uma Dulhare," A study on monothetic Divisive Hierarchical Clustering Method" International Journal of Advanced Scientific Technologies ,Engineering and Management Sciences (IJASTEMS-ISSN: 2454-356X) Volume.3,Special Issue.1,March.2017.

[9] A. Fahad, Z. Tari, A. Almalawi, A. Goscinski, I. Khalil, A. Mahmood, "PPFSCADA: Privacy preserving framework for SCADA data publishing", Future Generat. Comput. Syst., vol. 37, pp. 496-511, Jul. 2014.

[10] A. Fahad, Z. Tari, I. Khalil, I. Habib, H. Alnuweiri, "Toward an efficient and scalable feature selection approach for internet traffic classification", Comput. Netw., vol. 57, no. 9, pp. 2040-2057, Jun. 2013.

[11] P. Praveen B. Rama 2016," An Empirical Comparison of Clustering using Hierarchical methods and K-means" "International Conference on Advances in Electrical, Electronics ,Information, Information, Communications and Bio-Informatics (AEEICB2016), 978-1-4673- 9745-2 ©2016 IEEE.

[12] P. Praveen , B. Rama ,Ch. Jayanth Babu 2016," Big data environment for geospatial data analysis" International Conference on Communication and Electronics Systems (ICCES2016),DOI: 10.1109/CESYS.2016.7889816.

[13] S. Guha, R. Rastogi, K. Shim, "Cure: An efficient clustering algorithm for large databases", Proc. ACM SIGMOD Rec., vol. 27, no. 2, pp. 73-84, Jun. 1998.

[14] S. Guha, R. Rastogi, K. Shim, "Rock: A robust clustering algorithm for categorical attributes", Inform. Syst., vol. 25, no. 5, pp. 345-366, 2000.

[15] Han, M. Kamber, Data Mining: Concepts and Techniques, San Mateo, CA, USA:Morgan Kaufmann, 2006.

[16] 16. A. Hinneburg, D. A. Keim, "An efficient approach to clustering in large multimedia databases with noise", Proc. ACM SIGKDD Conf. Knowl. Discovery Ad Data Mining (KDD), pp. 58-65, 1998.

[17] A. Hinneburg, D. A. Keim, "Optimal grid-clustering: Towards breaking the curse of dimensionality in high-dimensional clustering", Proc. 25th Int. Conf. Very Large Data Bases (VLDB), pp. 506-517, 1999.

[18] Z. Huang, "A fast clustering algorithm to cluster very large categorical data sets in data mining", Proc. SIGMOD Workshop Res. Issues Data Mining Knowl. Discovery, pp. 1-8, 1997.