ROOTINDEXING
JOURNAL ABSTRACTING AND INDEXING SERVICE
INTERNATIONAL ADVANCE JOURNAL OF ENGINEERING, SCIENCE AND MANAGEMENT (IAJESM)
January-June 2023, Submitted in February 2023, iajesm2014@gmail.com, ISSN -2393-8048
Multidisciplinary Indexed/Peer Reviewed Journal. SJIF Impact Factor 2023 =6.753

Study On Review of Literature Hybrid Data Clustering Technique In Big Data Using Machine Learning

Anju,Research Scholar, Department Of Computer Science, Monad University, Hapur, Uttar Pradesh (India)
anjupanwar2793@Gmail.Com

Abstract

Big data actually implies an assortment of very large datasets which cannot be processed easily by implementing traditional computing methods. Big data is not merely a data, it has rather transformed into a comprehensive topic, which encompasses number of tools, procedures and frameworks. In general, big data is a datasets that could not be observed, attained, managed, and administered with hold-style IT and software/hardware components within a bearable period. Big Data technologies pronounce a novel origination of equipment's and constructions, designed to assist numerous organizations to cautiously abstract value from very huge bulks of a widespread diversity of the data by facilitating high-velocity acquisition, innovation, and/or exploration. This realm of Big Data have need of a modification in the computing manner, so that the clients can control both the data saving necessities and the hefty server processing needed to economically evaluate massive volumes of data. Much of this data explosion is the consequence of an intense rise in the equipment's that are sited at the border of the network comprising implanted sensors, smart phones, and tablet computers. This data produces novel chances to "abstract more value" in healthcare, human genomics, funding, oil and gas, exploration, investigation, and several other regions.

Keywords: Review of Literature, Hybrid Data Clustering Technique, Big Data using Machine Learning

Introduction

With the emergence of 5G technologies, a tremendous amount of data is being generated very quickly, which turns into a massive amount that is termed as Big Data. The attributes of Big Data such as huge volume, a diverse variety of data, high velocity and multivalued data make data analytics difficult. Moreover, extracting meaningful information from such volumes of data is not an easy task (Bhadani & Jothimani, 2016). As an indispensable tool of data mining, clustering algorithms play an essential role in big data analysis. Clustering methods are mainly divided into density-based, partition-based, hierarchical, and model-based clustering.

All these clustering methods are developed to tackle the same problems of grouping single and distinct points in a group in such a way that they are either similar to each other or dissimilar to points of other clusters. They work as follows: (1) randomly select initial clusters and (2) iteratively optimize the clusters until an optimal solution is reached (Dave & Gianey, 2016). Clustering has an enormous application. For instance, clustering is used in intrusion detection system for the detection of anomaly behaviours (Othman et al., 2018; Hu et al., 2018). Clustering is also used extensively in text analysis to classify documents into different categories (Fasheng & Xiong, 2011; Baltas, Kanavos & Tsakalidis, 0000). However, as the scale of the data generated by modern technologies is rising exponentially, these methods become computationally expensive and do not scale up to very large datasets. Thus, they are unable to meet the current demand of contemporary data-intensive applications (Ajin & Kumar, 2016). To handle big data, clustering algorithms must be able to extract patterns from data that are unstructured, massive and heterogeneous.

Big Data Analytics helps in examining large volumes of data to discover unknown patterns, hidden associations, meaningful developments, and other perceptions for making data-driven decisions in the process of tracking down better results through various tools and techniques. Big Data Analytics involves applying an algorithm or mechanical process to derive intuitions running through several data sets to look for meaningful associations between each other. As of today,

most of the data are generated from social networking sites. Hence Social networks and Big Data are mutually dependent on each other. So, Big Data industry look for the solutions of best machine learning techniques to predict data as well as purify the data based on different source systems of resources. Several industries have started using Big Data analytics, which helps allow organizations and companies to make better decisions to verify or challenge existing theories or models.

Review of Literature:

Fairness is a fundamental concept of education, whereby all students must have an equal opportunity in study or be treated fairly regardless of, e.g., their household income, assets, gender, race, or knowledge and domain-specific abilities. Fairness in the education system is reflected in a wide range of education-related activities, such as assessment and measurement (Dorans & Cook, 2016; ZlatkinTroitschanskaia et al., 2019), students' group work and group assignment (Rezaeinia, Góez, & Guajardo, 2021; Miles & Klein, 1998; Ford & Morice, 2003), graduate school admission (Song, 2018), or predicting student performance (Xiao, Ji, & Hu, 2021). One of the kernel demands for justice is education; therefore, having a fair education system is crucial to achieving justice in society (Meyer, 2014). In EDS, machine learning (ML) has been used in a wide variety of decisionmaking and learning analytics (LA) and educational data mining tasks (PeñaAyala, 2014; Dutt, Ismail, & Herawan, 2017; Romero & Ventura, 2020; McFarland, Khanna, Domingue, & Pardos, 2021); for example, student dropout prediction (Del Bonifro, Gabbrielli, Lisanti, & Zingaro, 2020; Kemper, Vorhoff, & Wigger, 2020), education admission decisions (Song, 2018) or forecasting on-time graduation of students (Livieris, Tampakas, Karacapilidis, & Pintelas, 2019; Hutt, Gardner, Duckworth, & D'Mello, 2019). The results of these ML models are the basis for building applications in EDS, such as student data analysis, learning support, and decision support systems. In fact, different ML-based decisions can be made based on protected attributes (i.e., the attributes for which the model is likely to exhibit bias), such as gender or race, leading to discrimination (Ntoutsi et al., 2020). Hence, improving fairness w.r.t. protected attributes in the results of ML models is imperative while maintaining the performance of the models. In other words, ensuring fairness in ML models also contributes to equity in educational systems. Research on fairness-aware ML has been carried out in various domains such as finance, education, healthcare, criminology and social issues (Mehrabi, Morstatter, Saxena, Lerman, & Galstyan, 2021; Le Quy et al., 2022). However, along with the development of the EDS research community, there are more and more studies on ensuring the fairness of ML models applied in education. These studies mainly focus on supervised learning models on the students' data (Gardner, Brooks, & Baker, 2019; Riazy, Simbeck, & Schreck, 2020; Bayer, Hlosta, & Fernandez, 2021). Recently, there have been several surveys on algorithmic bias and fairness in education, in which the authors presented the theory of fairness and summarized the classification and predictive methods (Baker & Hawn, 2021; Kizilcec & Lee, 2022).

R. Tamilselvi et al [2013], provided an overview of the data mining, types of data mining and different application involved with the data mining. Data mining is also called as knowledge discovery which discovers the patterns and delicate relationships in data and infers rule that allows future predictions. Afterwards, several tasks such as Association rules, Decision tree, clustering, and certain related algorithms are included along with the merits and demerits of those related algorithms. Finally, the different application domains of the decision trees and clustering algorithms are presented. This survey has not covered all classification techniques for understanding the estimate of the accuracy of the classification rules.

Adil Fahad et al [2014], presented the notions and algorithms associated with the clustering, altogether with a brief survey on the up-to-date algorithms of clustering as well as their

comparative analysis based on both theoretical and empirical perception. A categorizing framework is developed on the basis of the properties that are highlighted in the previous researches. Afterwards, experiments are conducted comparing the most representative algorithm from all categories utilizing large number of real data sets. The effectiveness of the algorithm is measured through several performance metrics. In conclusion the best performing clustering algorithms for the big data are highlighted in this work. The investigations of ensembles for single and individual clustering by means for the accuracy and stability has not been covered. The author has not defined the setting of parameters for every clustering algorithm.

Lisbeth Rodríguez-Mazahua et al [2016], offered a comprehensive analysis of Big Data mechanisms for identification of the primary complications, tools, presentation area and evolving classes of Big Data. The study within the field of big data is emerging immensely these days. In order to meet the objective of this study, the authors have studied 457 papers to investigate and categorize the concepts that already exist in the field of big data. This evaluated research work recommends associated material to researchers concerning significant functioning in study and Big Data application in miscellaneous practical regions. The research has not covered the aspects like provision of more specialised framework review and PLs (Programming language) for the analysis of big data for extending the information for the limitations, advantages for every framework and the language for giving the guidelines with the concept proof for exposing the functionality of every PL or some framework.

T. Sanjana et al [2016], analysed a number of algorithms that are used for the purpose of clustering in the big data processing. As per the research work, it has been discovered that ORCLUS, BIRCH and CLIQUE are the algorithms of clustering that can be utilized to detect outliers in big data. It has also been suggested that the algorithms like CURE and ROCK can be applied on the categorical data to generate clusters having arbitrary shape. Also, the non-convex shaped clusters can be obtained by utilizing the algorithms like COBWEB and CLASSIT, on the model based statistical data. At last, this work has also discovered that the algorithms like 1STING, OPTIGRID, PROCLUS and ORCLUS can be applied on the spatial data to acquire arbitrary shaped cluster. This survey has analyzed different algorithms for the data clustering for the processing in Big Data. The survey has focused on the identification of the outliers in large data sets by using the different algorithms but some parameters like variety, velocity and value are not according to the requirement and there are a lot of possibilities to improve that factor by using a classifier along with the clustering algorithm.

V. W. Ajin et al [2016], provided a complete study about the various clustering algorithms bearing the big data principles. The foremost objective of the clustering methods is to sort the data into different groups so that the related data entities can be assembled inside the similar group based on similarity, potential and activities. Furthermore, several varieties of the clustering methods are deliberated within this study, along with reasonable investigation of the utmost frequently employed and efficient algorithms i.e. FCM, BIRCH, K-Means, CLIQUE, etc. are disclosed based on the Big Data views. The study has not covered the single clustering algorithms being accurate and stable as compared to the individual and single clustering for the ensembles. The incorporation of distribution system for performance improvement and existing algorithms efficiency for big data has not been taken place. For every clustering algorithm, the appropriate parameter setting has not been discussed.

Ahmed Oussous et al [2017], deliberated a survey which is entirely based on the recent technologies that are emerged for big data. The big data characteristics have been thoroughly studied together with the trials that are raised up in the big data computing systems. The constituents and tools that are utilized in every layer of big data platform have been 8accentuated and furthermore equated the tools and allocations which is primarily based on their proficiencies,

profits and confines. Further big data a system are categorized on the basis of their service area and features that are delivered to consumers. In conclusion, this survey offers a wide-ranging knowledge about the structural design, practices and methodologies that are presently tailed in the computing system of big data. The author has lacked in covering the creation of next generation infrastructure, different areas like platform tools, domain specific tools and data organization.

Conclusion

In this survey paper, several research papers are studied to collect more precise information about Big Data Analytics using machine learning algorithms. This paper has a detailed analysis of machine learning algorithms and its best usage across various requirements based on criteria such as accuracy, architecture model, and storage data. This research paper will be more beneficial to anyone who is currently working on machine learning practices using Big Data Analytics. The difficulty of handling big complex data in big data technology using traditional learning algorithms has been explored. Various efficient, intellectual, and advanced learning algorithms are necessary to address the humongous volumes of diverse datasets. As a result of surveying these research papers, the resulted information gathered from these analytics provides more reliable and effective solutions. Many day-to-day real-world problems

References

Raj Kumar, and Rajesh Verma, "Classification algorithms for data mining - A survey", In the International Journal of the Innovations in Engineering and Technology (IJIET), vol. 1, no. 2, pp: 7-14. 2012.

Ahmed Oussous, Fatima-Zahra Benjelloun, Ayoub Ait Lahcen, and Samir Belfkih, “Big Data Technologies: A Survey”, Journal of King Saud University-Computer and Information Sciences, 2017.

Adil Fahad, Najlaa Alshatri, Zahir Tari, Abdullah Alamri, Ibrahim Khalil, Albert Y. Zomaya, Sebti Foufou, and Abdelaziz Bouras, “A survey of clustering algorithms for big data: Taxonomy and empirical analysis”, IEEE transactions on emerging topics in computing, Vol. 2, no. 3, pp: 267-279, 2014

G. Kesavaraj, and S. Sukumaran. "A study on classification techniques in data mining." In IEEE Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), pp. 1-7. 2013.

R. Tamilselvi and S. Kalaiselvi, "An Overview of Data Mining Techniques and Applications", International Journal of Science and Research (IJSR), Vol. 2, No. 2, pp. 506-509, 2013.

Praful Koturwar, Sheetal Girase, and Debajyoti Mukhopadhyay, "A survey of classification techniques in the area of big data", arXiv preprint arXiv: 1503.07477, 2015.

V. W. Ajin, and Lekshmy D. Kumar, "Big data and clustering algorithms", In IEEE International Conference on Research Advances in Integrated Navigation Systems (RAINS), pp. 1-5. 2016.

Ahmed Oussous, Fatima-Zahra Benjelloun, Ayoub Ait Lahcen, and Samir Belfkih, “Big Data Technologies: A Survey”, Journal of King Saud University-Computer and Information Sciences, 2017.