International Journal of Advanced and Innovative Research (2278-7844)/
Volume 11 Issue 1

Classification of Big Data in Cloud Environment: A Survey

Mukesh Kumar
Phd Schalor
Department of Computer Science & Engineering
Rabindranath Tagore University, Raisen, Madhya Pradesh, India
goutam.mukesh@gmail.com

Dr Sitesh Kumar Sinha
Professor
Department of Computer Science & Engineering
Rabindranath Tagore University, Raisen, Madhya Pradesh, India
siteshkumarsinha@gmail.com

Abstract— ***Big Data concern large-volume, large-variety, large-veracity growing at very high speed which very complex and generated from different sources such as IoT devices, social media, videos, financial transactions, and customer logs. Earlier technologies were not able to handle storage and processing of huge data thus Big Data concept comes into existence. This is a tedious job for users to identify accurate data from huge unstructured data. So, there should be some mechanism which classifies unstructured data into organized form which helps user to easily access required data. Classification techniques over big transactional database provide required data to the users from large datasets more simple way. There are two main classification techniques, supervised and unsupervised. In this paper we focused on to study of different type of classification algorithm for unstructured data. Further this paper shows application of each technique and their advantages and limitations.***

Keywords—big Data, classification technique

INTRODUCTION

Today‘s scenario, due to advancement in digitalization and computerization, very huge amount of data is available through different sources across in various field‘ which is increasing at an exponential rate contributing to the pool of data. These huge and bulky data sets proved as an effective tool for knowledge discovery and decision support system. There lies no benefits in having such a huge amount of data until it is not analyzed and utilized properly. For analyzing and effectively using this huge amount of data, earlier ―hands-on approaches used for data analysis, results in inefficient outcomes, those techniques required support from automated data processing tools. Infeasibility of hands-on approaches for large datasets provides a road map for an approach which bridges the gap between statistics and database management system by providing a new way of storing, processing and executing different methods for large datasets[1]. This approach fetches the valuable insights, uncovers hidden patterns and identifies the relationships from the huge datasets, for the effective decision making, improved planning, cost reduction, competitive advantage, improved customer relationship and customer satisfaction. Finally, a name is assigned to this introduced approach as data mining[2], which is understood in simple terms as digging deep into the elements and understanding their different aspects. Different terms are used in the literature to define data mining including data analysis, knowledge extraction, and knowledge mining from datasets.

LITERATURE REVIEW

A. Big Data

Big data computers include request applications and large amounts of data production. Big Data shows architectures of data and provides the necessary tools for large data management. Traditional data is a less structured database volume operating on a single server. Big Data is, on the other hand, a massive collection of data, including audio and video, structured and unstructured data including log files, data sensors and social media. As an integral and non-volatile data collection, Inmon [3] has described traditional data as helping analysts in decision making. Traditional databases lack the solutions to large volumes of rapidly changing data that are unstructured.

Big Data Mining was focused by Wei Fan and Albert Bifet, [4] who could obtain useful information. In the past, large amounts of data could not be mined. However, it is now possible with the assistance of software such as Apache Hadoop. The authors concluded that, besides Apache Hadoop, there is also a large data tool such as R, MOA, Strom and Vow pal Wabbit. Pegasus and Graph Lab are a large-scale open source tool. In order to develop intelligent cities for better services and customer experience, big data mining applies for business, technology and healthcare.

B. *Big Data Analytics*

Sachidanand and Nirmala [5] discussed the concept of big data and several market tools for the exploration of unstructured data. They also provided detailed information on Big Data analyses and their importance in various fields, such as medical, public, retail, manufacturing, and others. Story, management and processing issues were discussed by Stephen Kaisler et al. .[6], Sadhana and Savitha Shetty [7] have tried and proposed a prediction model for an analysis of facts concerning Diabetic datasets. In the past, Relational Data Base Management Systems were used as small data, but at present RDBMS is not possible because the data is large. Vikram Phaneendra & Madhusudhan Reddy [8] suggested HDFS as a suitable tool for treating the current situation, because of its enormous size and complexity.

Current large data are a group of structured, semi-Structured, homogenous, unstructured and hetrogeneous data described in Kiran Kumara Reddi & Indira [9]. They proposed the transfer of a large number of data over the network and recommended the transfer of Big Data with new algorithms. Albert Bifet [10] recently stressed the importance of efficient and rapid tools for the analysis of the data set in real time. At the moment, huge amount of information is created via many sources using web servers into an enormous data set, a challenge in extracting useful information (Mrigank Mridul, et.al) [11]. In their Big Data review, Sagiroglu and Sinanc [12] described Big Data content, security and safety. Data are becoming more and more complex, every day, which poses many challenges. Garlasu et.al [13] discussed the importance of grid computing which provides storage capabilities with advantages. Raghupathi and Raghupathi [14] discussed the importance of platforms and tools to speed up its processing time in the health care industry.

Big data comprises such a large volume of data, and traditional databases and software techniques are difficult to handle. A technical barrier is encountered in the use of large data applications when data are moved across different locations, which is very expensive and requires large main store for holding computer data. Big data includes data transactions and interactions based on the size and complexity of data sets that exceed regular technical ability in cloud data collection, organization and processing It is processed in high-performance, data-intensive real-time. Big data applications are handled for structured and unstructured data sharing by effectively collecting the data to achieve a faster response and less time for classification. Existing research concentrates on Big Data mining using a heterogeneous and independent evolving theorem that improves cloud security and privacy [15]. Flex Analytics is also a prototype method designed to increase data transmission bandwidth.

In data applications, a centralized control unit is used to identify large amounts of data for attacks and malfunctions. Cloud computing is a distributed parallel computing system that has become an often-used big data analytics computing application. Both methods do, however, not address space and time-related problems.

The classification performance is handled by a distributed framework named MapReduce for prototype reduction. MapReduce divides data based on the applications of large data. The prototype avoids the data set, which reduces processing time. However, because the imbalance in data, which leads to a class imbalance issue, the learning process gets complicated. The prototype cut allows for the classification of big data and analyses the accuracy of classification and processing for information sharing with big data applications. However, for the prototype reduction technique to efficiently manage big data, the nearest neighbor rule is necessary. Unbalanced large data are used to achieve scalable and parallel large data deployment with random forest technology. MapReduce framework parallels the processing of Big Data information to develop cloud-based scalable and error acceptance applications. Although a small sample of data classification is not possible with large numbers of mapper in Big Data applications, random forest with big data algorithm is clustered to two different phases, namely a map and a reducing phase. For the purpose of big data computing and cloud-based sharing information, the Parallel Symmetric Matrix-based Predictive Bayes Classifier (PSM-PBC) model [5] is proposed. There are three processes for sharing information in the proposed PSM-PBC model. At first, Tridiagonal symmetric matrix is built in parallel with distributed Big Data applications to allow faster calculation of data removal and data sharing through cloud paradigms through transformation of the householder. Next, the Bayes Classification Model is developed in order to evaluate actual diagonal search data for their results. This increases the prediction rate by the results of every user request. Finally MapReduce is enhanced by Bay classes that provide forecasting big data analytics for improved computing and sharing of information.

Similarly, it was proposed to offer an efficient Big Data Calculation and Information Sharing in the Cloud Computing Environment to discretized support vehicle classification and prediction (DSV-CP) [16] model. It Pre-processing is initially carried out in the DSV-CP model, based on discretizing equivalence interval that helps to remove noise from various sources and incompatible data. The computation time and space complexity are reduced while removing the noise and inconsistency present in the data. In addition, the DSV-CP [16] model uses a support vector forecast classifier to classify data using parallel hyperplanes based on the user query request, with the purpose of improving user request information classification precision on big data. The proposed DSV-CP model predicts the information on the user request with the classified data accurately. The framework called Canopy MapReduce Linguistic Fuzzy Rules (LFR CM) [5] is designed to classify and share information in the cloud. The parallel mechanism based on the MapReduce parallel programming model is used with linguistic fluid rules. The parallel mechanism guarantees minimum runtime. The resulting language fogging rules are also used to form another sample set to accelerate classification precision and guarantee a minimum runtime for big data classification. The algorithm of canopy shuffle is also applied. The canopy-fuzzy MapReduce algorithm's convergence rate will also be accelerated [5]. In order to enhance classification time and precision in parallel in fugitive knowledge base and canopy fugitive MapReduce algorithm, hybrid classification models are also developed. With the Amazon EC2 Stanford Large Network cloud dataset collection,

.COMPARISION & ANALYSIS

PSM-PBC model initially performs mapping to provide efficient Big Data Calculation and reduces computational time by 34.11 percent. Next, the DSV-CP[38] model removes the noise and non-consistent data with the data pre-treatment task. In addition, the canopy shuffle of the MapReduce algorithm in the LFR-CM[28] frame enhances the accuracy of the mapping factor by 20.85%.

REFERENCES

[1] Varatharajan, R., Manogaran, G., and Priyan, M.K., 2018. A big data classification approach using LDA with an enhanced SVM method for ECG signals in cloud computing. *Multimedia Tools and Applications*, *77*(8), pp.10195-10215.

[2] Qi, G., Zhu, Z., Erqinhu, K., Chen, Y., Chai, Y. and Sun, J., 2018. Fault-diagnosis for reciprocating compressors using big data and machine learning. *Simulation Modelling Practice and Theory*, *80*, pp.104-127.

[3] Sadhana and Savitha Shetty, "Analysis of Diabetic Data Set Using Hive and R", International Journal of Emerging Technology and Advanced Engineering, 4(7), 2014.

[4] Vikram Phaneendra,S.& E.Madhusudhan Reddy, "Big Data- solutions for RDBMS problems-A survey". In12th IEEE/IFIP Network Operations & Management Symposium. 2013

[5] V. Vennila, A. and Rajiv Kannan "Symmetric Matrix-based Predictive Classifier for Big Data computation and information sharing in Cloud"

[6] Kiran kumara Reddi & Dnvsl Indira "Different Technique to Transfer Big Data :survey" IEEE Transactions. 52(8):2013.

[7] Albert Bifet, "Mining Big Data In Real Time" Informatica 37.2013.]

[8] Ahmed, ESA & Saeed, RA 2014, 'A survey of big data cloud computing security', International Journal of Computer Science and Software Engineering (IJCSSE), vol. 3, no. 1, pp. 78-85.

[9] Assunção, MD, Calheiros, RN, Bianchi, S, Netto, MAS & Buyya, R 2015, 'Big Data computing and clouds: Trends and future directions', Journal of Parallel and Distributed Computing, vol. 79- 80, pp. 3-15.

[10] Aydin, G, Hallac, IR & Karakus, B 2015, 'Architecture and implementation of a scalable sensor data storage and analysis system using cloud computing and big data technologies', Journal of Sensors, vol. 2015.

[11] Ayma, VA, Ferreira, RS, Happ, P, Oliveira, D, Feitosa, R, Costa, G, Plaza, A & Gamba, P 2015, 'Classification algorithms for big data analysis, a map reduce approach', International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives, vol. 40, no. 3W2, pp. 17-21.

[12] Barkhordari, M & Niamanesh, M 2015, 'ScaDiPaSi: An Effective Scalable and Distributable MapReduce-Based Method to Find Patient Similarity on Huge Healthcare Networks', Big Data Research, vol. 2, no. 1, pp. 19-27.

[13] Baro, E, Degoul, S, Beuscart, Rg & Chazard, E 2015, 'Toward a literature-driven definition of big data in healthcare', BioMed Research International, vol. 2015.

[14] Bautista Villalpando, L, April, A & Abran, A 2014, 'Performance analysis model for big data applications in cloud computing', Journal of Cloud Computing: Advances, Systems and Applications, vol. 3, no. 1, pp. 19-38.

[15] Sheikh Md. Zubair, Md. Zahoor, and Dr. Rajiv Yadav, "A Study of Prediction Classifier of Big Data Techniques in Cloud Setting", JARIIE-ISSN(O)-2395-4396 Vol-4 Issue-2 2018

[16] G. Kalyani, Dr M.V.P.Chandra Sekhara Rao. "Privacy Preserving Classification Rule Mining for Balancing Data Utility and Knowledge Privacy using Adapted Binary Firefly Algorithm", Arabian Journal of Science and Engineering, Springer, ISSN: 2193-567X, 2017. [Free Journal, indexed by: SCI, SCOPUS, Google Scholar, etc., Impact Factor: 1.32].