Iraqi Journal for Computer Science and Mathematics

Journal Homepage: http://journal.esj.edu.iq/index.php/IJCM
ISSN: 2788-7421

A Brief Review of Big Data Analytics Based on Machine Learning

Ahmed Hussein Ali[1], Mahmood Zaki Abdullah[2], Shams N. Abdul-wahab[3,*], Mohammad Alsajri[4]

[1]Department of computer, College of Education, AlIraqia University, Baghdad, Iraq

[2]Department of computer engineering, Baghdad, Iraq

[3]Department of Computer Technical Engineering, Alsalam University College, Baghdad, Iraq

[4]Faculty of Computing, College of Computing and Applied Sciences, Universiti Malaysia Pahang, Malaysia

*Corresponding Author: Shams N. Abd-Alwahab

DOI: https://doi.org/10.52866/ijcsm.2020.01.02.002
Received Aprile 2020; Accepted June 2020; Available online July 2020

ABSTRACT: Owing to the exponential expansion in the data size, fast and efficient systems of analysis are extremely needed. The traditional algorithms of machine learning face the challenge of learning bottlenecks such as; human participation, time, and the accuracy of prediction. But, the efficient and fast methods of dynamic learning offer considerable advantages like lower human participation, rapid algorithms of learning, and easiness implementation. This review paper presents the researches with a brief display for recently existing works in big data analytics and the effective algorithms of machine learning, furthermore, the issues of resources allocation in big data.

Keywords: Machine Learning; Bigdata; Classification.

1. INTRODUCTION

The algorithms of machine learning represent one of the significant components of artificial intelligence (AI). Within industrial applications, the algorithms of machine learning can collect, analyze, and predict data. These algorithms are auto-learned from the data and hence behaved with no clearly programmed. The data should be analyzed for evaluating and estimating the advantages of organizational aims. The analytics of data needed data mining, artificial intelligence, machine learning, mechanism of prediction, etc. for evaluating the data evaluation inside the organization. The algorithms of machine learning can provide joint tools for data reusability evaluation, hence, it is preferable to predict and estimate the developed analytics [1]. This review paper presents a brief display for recently existing works in big data analytics based on effective algorithms of machine learning, besides the issues of resources allocation.

2. THE ANALYTICS OF BIG DATA BASED ON MACHINE LEARNING

Alvaro Brandon Hernandez et al. [2], proposed a machine learning that selects the best parameters in task parallelization for the workloads of big data. The method is addressed the problem of the basis of big data technologies and infrastructure that causes under-use and failure of accessible resources and applications. Lixia Yang et al. [3], proposed new learning methods called extreme learning machines of incremental Laplacian regularization. These methods are developed for online learning (semi-supervised) that accept unlabeled and labeled examples. Compared with the other related methods,

it can obtain good performance when the labeled examples are restricted. The prediction accuracy of these methods was raised by 9% for the classification problem. Mohammad Habib Ur Rehman et al. [4], proposed a novel method called (CCM) concentric computing model for deploying of big data applications in industrial IoT. The industrial IoT combines parallel distributed systems with machine learning like grids, clouds, clusters for the storage of big data, processing, and analytics. The proposed method (CCM) aid in processing, integration, and analysis of industrial data. Xiaoming He et al. [5], presented a novel big data framework for enhancing the performance of QoE to the smart city. The new architecture includes several main plans: data storage, processing, and application. Simulation results by applying three machine learning algorithms (SVM, DL, and KNN) and measuring output in term of recall, precision, and accuracy indicate that the proposed architecture achieves high performance of QoE. Shanjiang Tang et al. [6], presented an equitable resource allocation method for analyzing big data over the environment of cloud computing. This method called (LTRF) Long-Term Resource Fairness. Extended to (LTRF) they proposed (H-LTRF) Hieratical Long-Term Resource Fairness for hieratical resource allocation in the cloud. Finally, they develop open source (LTYARN) by combine LTRF and H-LTRF in YARN. The experimental results show superior resource fairness comparing with current schedulers of YARN. Cen Chen et al. [7], proposed a method of parallel hierarchal extreme learning machine depend on in-memory cluster computing Flink. This method exploits the GPUs to accelerate Flink and extend ELM from a single hidden layer feed-forward network to Multi-Layer Perceptron (MLP). This proposed method is capable of accelerating considerable algorithms of machine learning. The obtained outcomes illustrated that this framework provides high performance, speedup for large-scale big data. Zhun-Ga Liu et al. [8], proposed a new hybrid classification method for uncertain data. This method exploits several kinds of classification; Fuzzy, Hard, and Credal classifications. Among these classifications, the appropriate one will be chosen regarding the context. Hard classification is chosen when the object is classified clearly. If the object is difficult to classify or the object near to the border of much diverse class, the credal classification will be adopted. S. Kamaruddin et al. [9], proposed a fast method of parallel evolving clustering. This method uses the framework of apache-spark processing with the storage of the distributed data. The experimental results are illustrated on the datasets of credit card fraud. The method performance was about 2.6x faster than other related clustering methods. L. Wilkinson [10], proposed a new algorithm to detect unusual features in big data called hdoutlier. Hdoutlier address many problems of outlier detection such as non-normal distribution, scalability, and high dimensionality. The experimental results demonstrate that hdoutliers algorithm reduces the risk of making a false outlier detection for a board class. Diego Marron et al. [11], presented a combination method for big data steam classification. The method of combining three machine-learning algorithms (KNN, Hoeffding-Tree, and Gradient Descent method) and utilizing GPUs for data stream learning due to their high ability. The experimental results show that the new method of obtaining powerful predictive performance. Xiaoyong Li et al. [12], proposed a parallel and fast scheme of trust-computing. This scheme depends on analyzing big data for trustworthy cloud services with high speed and low overhead. This method overcomes the problem of user fear in interactive and uploads most sensitive data to cloud data centers. The experimental results verified the effectiveness and feasibility of the proposed scheme. Mehdi Mohammadi et al. [13], proposed a framework of three-level learning to big data created via smart cities. The framework utilizes a combination of labeled and unlabeled data by semi-supervised deep reinforcement learning style. The results demonstrate that the new framework provides better data recycling, efficient sampling, and fast prediction. Deepak Puthal [14], presented a model of static lattice to control information through stream of big sensing data. The proposed model addresses the problem of information flow control. The model evaluation demonstrates low latency of incoming big data. Mingxing Duan et al. [15], proposed a new parallel multi-classification algorithm based on spark framework called spark extreme learning machine (SELMs). SELMs include three parallel sub-algorithm. Extensive experiments have been conducted to the proposed algorithm showing that SELMs achieves 33.81x speedup with 35 nodes. Yuantao Chen et al. [16], presented a novel method for online learning depend on the support vector machine. The new algorithm addresses the low execution effectiveness problem for SVM within large scale training data. The experimental results demonstrate that the online learning algorithm based on SVM supplies effective speed of training and accurate rates of classification. Georgios Ghatzigeorgakidis et al. [17], proposed a novel distributed processing framework(FML-KNN). The method performs regression and classification applied in Apache Flink. Comparing to other methods that similarly utilize Apache Hadoop and Apache Spark, this method was executed considerably faster with similar workloads. B. Yadranjiaghdam et al. [18], worked on developing a framework for processing big data in the online traffic data stream. The Apache Kafka was utilized in this framework as a component for data ingestion, additionally, the spark streaming was utilized as a component for processing data to supply additional high-level and developed analytics of data stream. G. Liu et al. [19], presented a method of partition to process live data stream. This method solves the unbalanced intermediate data stream problem within spark streaming. C. Misale et al. [20], presented a method to processing real-time data called PiCo. PiCo (pipeline composition) was proposed for enhancing big data stream analysis process comparing with spark and Kafka streaming.

3. CONCLUSIONS

This review paper presented an expansion display for the significant methods of Big data streaming that were utilized in the literature and the main issues concerning these methods, and the appropriate solutions. Furthermore, several recently existence methods in Big data streaming based on the algorithms of machine learning are also reviewed.

4. ACKNOWLEDGEMENT

Many thanks are going for the ICCI, Informatics Institute for Postgraduate Studies (IIPS_IRAQ) to their ethical assist. Also, Special thanks for the university of Al-Mustansiriyah to its encourage for finishing this review paper.

REFERENCES

[1] J. Qiu, Q. Wu, G. Ding, Y. Xu, and S. Feng, "A survey of machine learning for big data processing," *EURASIP Journal on Advances in Signal Processing*, pp. 1–16, 2016.
[2] Álvaro Brandón Hernández, M. S. Perez, S. Gupta, and V. Muntés-Mulero, "Using machine learning to optimize parallelism in big data applications," *Future Generation Computer Systems*, vol. 86, pp. 1076–1092, 2018.
[3] L. Yang, S. Yang, S. Li, Z. Liu, and L. Jiao, "Incremental laplacian regularization extreme learning machine for online learning," *Applied Soft Computing*, vol. 59, pp. 546–555, 2017.
[4] M. H. ur Rehman, E. Ahmed, I. Yaqoob, I. A. T. Hashem, M. Imran, and S. Ahmad, "Big Data Analytics in Industrial IoT Using a Concentric Computing Model," *IEEE Communications Magazine*, vol. 56, no. 2, pp. 37–43, 2018.
[5] X. He, K. Wang, H. Huang, and B. Liu, "Qoe-driven big data architecture for smart city," *IEEE Communications Magazine*, vol. 56, pp. 88–93, 2018.
[6] S. Tang, B. S. Lee, and B. He, "Fair resource allocation for data-intensive computing in the cloud," *IEEE Transactions on Services Computing*, 2016.
[7] C. Chen, K. Li, A. Ouyang, Z. Tang, and K. Li, "GPU-Accelerated Parallel Hierarchical Extreme Learning Machine on Flink for Big Data," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 47, no. 10, pp. 2740–2753, 2017.
[8] Z.-G. Liu, Q. Pan, J. Dezert, and G. Mercier, "Hybrid Classification System for Uncertain Data," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 47, no. 10, pp. 2783–2790, 2017.
[9] S. Kamaruddin, V. Ravi, and P. Mayank, "Parallel Evolving Clustering Method for Big Data Analytics Using Apache Spark: Applications to Banking and Physics," *in International Conference on Big Data Analytics*, pp. 278–292, 2017.
[10] L. Wilkinson, "Visualizing Big Data Outliers through Distributed Aggregation," *IEEE transactions on visualization and computer graphics*, vol. 24, pp. 256–266, 2018.
[11] D. Marrón, J. Read, A. Bifet, and N. Navarro, "Data stream classification using random feature functions and novel method combinations," *Journal of Systems and Software*, vol. 127, pp. 195–204, 2017.
[12] X. Li, J. Yuan, H. Ma, and W. Yao, "Fast and Parallel Trust Computing Scheme Based on Big Data Analysis For Collaboration Cloud Service," *IEEE Transactions on Information Forensics and Security*, vol. 13, pp. 1917–1931, 2018.
[13] M. Mohammadi and A. Al-Fuqaha, "Enabling Cognitive Smart Cities Using Big Data and Machine Learning: Approaches and Challenges," *IEEE Communications Magazine*, vol. 56, no. 2, pp. 94–101, 2018.
[14] D. Puthal, "Lattice-Modeled Information Flow Control of Big Sensing Data Streams for Smart Health Application," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 1312–1320, 2019.
[15] M. Duan, K. Li, X. Liao, and K. Li, "A parallel multiclassification algorithm for big data using an extreme learning machine," *IEEE transactions on neural networks and learning systems*, 2017.
[16] Y. Chen, J. Xiong, W. Xu, and J. Zuo, "A novel online incremental and decremental learning algorithm based on variable support vector machine," *Cluster Computing*, vol. 22, no. S3, pp. 7435–7445, 2019.
[17] G. Chatzigeorgakidis, S. Karagiorgou, S. Athanasiou, and S. Skiadopoulos, "FML-kNN: scalable machine learning on Big Data using k-nearest neighbor joins," *Journal of Big Data*, vol. 5, no. 1, pp. 4–4, 2018.
[18] B. Yadranjiaghdam, S. Yasrobi, and N. Tabrizi, "Developing a Real-time Data Analytics Framework For Twitter Streaming Data," *2017 IEEE International Congress on*, pp. 329–336, 2017.
[19] G. Liu, X. Zhu, J. Wang, D. Guo, W. Bao, and H. Guo, "SP-Partitioner: A novel partition method to handle intermediate data skew in spark streaming," *Future Generation Computer Systems*, 2017.
[20] C. Misale, M. Drocco, G. Tremblay, and M. Aldinucci, "Pico: a novel approach to stream data analytics," *European Conference on Parallel Processing*, pp. 118–128, 2017.