International Journal of Computer Science and Software Engineering (IJCSSE), Volume 9, Issue 7, July 2020
ISSN (Online): 2409-4285 www.IJCSSE.org Page: 39-42

A Big Data Analytics Literature Survey Using Machine Learning Algorithms

Senthil Raman[1], Dr. Narayanan B[2] and Dr. Velmurugan K[3]

[1] Dept. of Computer Science & Engineering, Annamalai University, Chidambaram, Tamilnadu, India

[2] Assistant Professor, Dept. of Computer Science & Engineering Wing, D.D.E, Annamalai University, Chidambaram, Tamilnadu, India

[3] Professor, Anjalai Ammal-Mahalingam Engineering College, Thiruvarur, Tamilnadu, India

ABSTRACT

Big data analytics is a methodology to analyze large volumes of data sets involved in various sectors and extract information that helps to predict proper conclusions. These concluded results are then used to predict the future of business. The data analysed comes in various volumes, velocities, and varieties. The Volume is the amount of data generated every second. Many organizations like healthcare, airlines, banking, and geospatial applications, etc. collect large volumes of data. The rate or speed at which the data is generated is referred to as Velocity. Social media plays a vital role as enormous amount of data is collected when anyone uploads or comments on any posts. Variety is defined as data coming in different forms like numerical data, texts, images, pdfs, videos, JSON, XML, etc. Many frameworks are analyzed in this survey paper to process these enormous data. Various complex algorithms are explained which helps developers and analysts to analyze and predict the future results.

Keywords: *Big Data Analytics, Dimensionality Reduction technique, Machine Learning algorithms and Data Mining.*

1. INTRODUCTION

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 [1] 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.

2. BIG DATA ANALYTICS AND MACHINE LEARNING

Nowadays, Big Data Analytics is widely found in various domains, namely IT Sectors, banking, hospitals, healthcare, insurance sector, universities, social media, agriculture, bioinformatics, industrial sectors, entertainment industry, geo spatial related applications, etc. Big data, along with machine learning helps in analysis of humungous data to trace out unfamiliar patterns and other important information. Machine learning [2] is considered to be an approach used to create models to be used across various sectors for prediction, for storing humongous data volumes, analyzing the predicted information. Machine learning can be categorized into three types of learning, such as supervised machine learning, unsupervised machine learning, and reinforcement machine learning. Supervised machine learning works with well-labelled training data. Each of the labelled training data consists of input and output value. With that data, the supervised learning algorithm completes the analysis and creates a function that is used to map new values. Wherein, unsupervised machine learning technique deals with unlabeled datasets, from where hidden insights are derived. Reinforcement machine learning technique works with trial and error method and with the results obtained the learner knows which method to utilize.

Jundong Li et al. have illustrated the Feature Selection algorithm [4], which is one of the dimension reduction techniques for Big Data Analytics using high-dimensional data. The author also explored the advantage of using specific core algorithm from the background of error data, which is proved to be more efficient and effective in handling high dimensional data. Many real-world applications have started using Feature selection algorithms to remove unwanted and irrelevant data to feature only relevant information. The author distinguishes the challenges on feature selection across various features selection algorithms like Linked Data, Multisource, Multiview data, and streaming data. Post learning tasks like classification and clustering could be improved when there is a basic knowledge of feature selection algorithms. This is helpful when feature selection algorithms are applied to datasets having structured features. MapReduce and MPI are the distributed programming frameworks implemented by the authors to evaluate very large-scale datasets using parallel feature selection. Data velocity and data variety are the biggest challenges of Big Data, which the author concluded to face by developing sophisticated feature selection algorithms.

Mingda Li et al. have demonstrated research on Conditional Functional Dependency [5] algorithm rules about Big Data. One of the powerful Data cleaning tools used to find the hidden relationship between different kinds of items is Conditional Functional Dependency (CFD) discovery algorithms. Data cleaning systems with CFDs are considered to be effective based on the defined High-quality rules. Several algorithms are used for performing modeling and data cleaning tasks. CFD algorithms like the one-pass sampling algorithm, Multiple-pass scan algorithm, DFCFD algorithm, and the Integration of three algorithms are presented in their research. The tools and libraries offered in the study are namely Kibana, Weka, R-cran, MOA, Elastic Search, Python and Sci-Kit. Java is the technology used to implement the algorithms, and once it is developed, it is validated on a machine having Intel Core i7 with Ubuntu configuration.

Sunil Kumar et al. have presented a research work based out of machine learning techniques using the hybrid clustering [6] concept to extract hidden information from complex data. This hybrid clustering algorithm overcomes all the disadvantages of k-means and Hierarchical algorithms. The weather report is preferred for the analysis, which is defined in standard ASCII format. The datasets are defined by 3 sections, namely additional data section, mandatory section, and control section. The author recommended a highly scalable clustering algorithm comparing a new hybrid algorithm with the present algorithm based on the recall, precision, execution time, F-measure, and accuracy of results. Out of this observation, it is derived that the accuracy of the implemented hybrid clustering algorithm which produces the best parameter values such as recall, precision, and F-measure values. The author implemented the hybrid clustering approach using Hadoop MapReduce with an observation on NCDC weather data files and the k-means algorithm, which explored more on the structure of a data set.

Lingyun Xiang et al. have presented their work on advanced unsupervised learning techniques on predicting required information from a huge volume of unlabeled heterogeneous data. It provides solutions for the huge complicated data modelling samples in Big Data considering various approaches by using the method of MKL, multiple kernel learning algorithms, which gives way for retrieving cohesive information from diverse data. The process implemented by the author is a two-stage unsupervised multiple kernel extreme data learning algorithm (TUMK-ELM [7], for short) is implemented for getting accurate results. This research paper evaluates efficiency, convergence speed, clustering loss, the same clustering performance level with various existing clustering techniques. The experimental results depict the learning speed achieved, which is a 1,40, 000 times faster than LMKKM, 1000 times faster than RMKKM, 8,500 times faster than MKC-LKAM overall.

Hakeem A. Owolabi et al. published a paper describing a few other learning techniques on Big Data analytics. Some of the examples of such learning techniques are Regression [8] as Random Forrest, learning Problem, kernel regression, Regression Trees, and Support Vector Machines (SVMs), logistic regression, etc. Big Data Analytics is also a predictive modeling technique to predict the risk in advance. Several predictive models have been for mutated and assessed using the above learning techniques for completion risk prediction. The study also recognizes the misappropriate classification of the wastes as mixed wastes under the current surviving waste management approaches. The comparison of the experimental results shows that the Random Forrest is best at the least RSS(residual error) value as 1.23, RSS value is considered to be high at 469.56 for DNN, and medium-range RSS value is at 23.20 for Linear Regression. The RF comparatively gives high predictive performance accuracy than the above-discussed analytics techniques.

Mohammadhossein Ghahramani[9] et al. have discussed the research on the massive amounts of various mobile data such as structured data, unstructured data, and semi-structured data generated by mobile data utilities. The results are depicted using the exploratory data analysis algorithm. The behavior of the systems which are interacted overwhelmed by the exploratory spatial data analysis algorithm. The permutation-based analysis technique is implemented to test large volumes of data.

In recent days, large volumes of data are analysed using new cloud computing technologies. In this paper, Novel spatial analysis is explained to help in detecting the spatial distribution of mobile phones. Various spatial and non- spatial objects are analysed to determine the correlation between them. Kernel density analysis is the method used to monitor and evaluate plans at any level and makes the necessary improvement satisfying the requirements of the user. This approach also proved to be more effective and long-lasting. The research output of this author benefits in planning and development and also to upgrade infrastructure in the telecommunication industry.

Sunil Kumar et al.[10] have illustrated another paper on the impacts of Big Data Analytics in the health care domain using Machine Learning Algorithms. Four primary attributes associated with Big Data are namely: velocity, volume, veracity, and various parameters are discussed in this research in a detailed manner along with healthcare departments. The author discussed the useful platforms for Big Data analytics such as Apache Hadoop, HDFPS, Apache HBase, Apache Oozie, Apache Pig, MapReduce, Apache Hive, Apache Yarn, Apache Zookeeper and Apache Sqoop. With the results of the analysis of different available amounts of Medical data using Hadoop based terminologies, the authors have finally proposed an architecture that mainly focusses on solving health care problems. This architecture also explains the development of methods used to analyze medical data.

Table 1 summarizes the various Big Data machine learning algorithm approaches already used by different researchers that are analysed for deriving the best results based on other criteria like big high dimensional data, architecture model, and accuracy. In this survey paper, the main objects discussed comprises of the best methods derived by various authors from their perspective. Every author's best-derived approach [3] is described separately in every paragraph, which clears the understanding in determining the suitable approach to handle Big Data's machine learning techniques, which can be used for future research. The following paragraphs explain the few important literature reviews pertaining to this research area.

3. TABLES, FIGURES AND EQUATIONS

3.1 Tables and Figures

Table 1 summarizes the various Big Data machine learning algorithm approaches already used by different researchers that are analysed for deriving the best results based on other criteria like big high dimensional data, architecture model, and accuracy. In this survey paper, the main objects discussed comprises of the best methods derived by various authors from their perspective. Every author's best-derived approach [3] is described separately in every paragraph, which clears the understanding in determining the suitable approach to handle Big Data machine learning techniques, which can be used for future research. The following paragraphs explain the few important literature reviews pertaining to this research area.

Table 1: Comparison of different algorithms/techniques

S. No	Author (s) Name	Algorithm/Techniques	Implementation approach
1	Jundong Li and Huan Liu"	The Supervised Feature Selection algorithm	They discussed various supervised feature selection algorithms using MapReduce techniques.
2	Mingda Li, Hongzhi Wang, and Jianzhong Li	CFD discovery algorithm	The tools and libraries presented in the research are namely Kibana, Weka, R-cran, MOA, Elastic Search, and Python Sci-Kit.
3	Sunil Kumar and Maninder Singh	Hybrid algorithm	The author discussed finding a solution to extract fruitful information from unformatted data.
4	Rua-Huan Tsaih and, Biing-Shen Kuo	machine-learning modelling-RLEM	Discussed a model to the machine-learning model is implemented using TensorFlow and GPU.
5	Sunil Kumar	Conceptual Architecture using Hadoop	The author discussed the conceptual architecture to predict healthcare remedies using Hadoop-based technologies.
6	Lingyun Xiang, Unhang Zhao, Qian Li, Wei Hao and Feng Li	TUMK-ELM	This research paper evaluates efficiency, convergence speed, clustering loss, a same clustering performance level with various existing clustering techniques
7	Mohammad Hossein Ghahramani, MengChu Zhou and Chi Tin Hon	K-means, Hierarchical, Agglomerative clustering and devise clustering	The authors suggest a Kernel density method for more effective and long-lasting results.
8	Hakeem A. and Muhammad Bilal et.al	The Learning techniques discussed as Random Forrest, SVM, and Regression trees.	The author identified that the random forest method is the best method concerning RSS value among discussed all four techniques.

4. CONCLUSIONS

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

[1] Mohammad Sultan Mahmud, Joshua Zhexue Huang, Salman Salloum, Tamer Z. Emara and Kuanishbay Sadatdi "A survey of data partitioning and sampling methods to support big data analysis", Big Data Mining and Analytics, Vol.3, Number 2, 2020, pp. 85-101.

[2] [2] Zhiqiang Ge, Zhihuan Song, Steven X. Ding and Biao Huang "Data Mining and Analytics in the Process Industry: The Role of Machine Learning", IEEE Access, 2017, Vol.5, pp. 20590-20616

[3] [3] Alexandra L'Heureu, KatarinaGrolinger, Hany F. Elyamany, Miriam A.M. Capretz "Machine Learning with Big Data: Challenges and Approaches", IEEE Access, 2017, pp 7776-7797

[4] [4] Jundong Li and Huan Liu "Challenges of Feature Selection for Big Data Analytics", IEEE Intelligent Systems, Vol.32, Issue.2, 2017, pp. 9-15

[5] [5] Mingda Li, Hongzhi Wang and Jianzhong Li, "Mining conditional functional dependency rules on big data", IEEE Big Data Mining and Analytics, Vol.3, 2020, Issue.1, pp. 68-84

[6] [6] Sunil Kumar and Maninder Singh, "A Novel Clustering Technique for Efficient Clustering of Big Data in Hadoop Ecosystem", IEEE Big Data Mining and Analytics, Vol.2, Issue.4, 2019, pp. 240-247

[7] [7] Lingyun Xiang, Unhang Zhao, Qian Li, Wei Hao, and Feng Li "TUMK-ELM: A Fast Unsupervised Heterogeneous Data Learning Approach" IEEE Access, Vol.6, 2018, pp. 35305-35315.

[8] [8] Hakeem A. Owolabi and Muhammad Bilal, "Predicting Completion Risk in PPP Projects Using Big Data Analytics", IEEE Transactions on Engineering Management, vol.67, Issue.2, 2020, pp. 430-453.

[9] [9] Mohammadhossein Ghahramani, MengChu Zhou, and Chi Tin Hon, "Mobile Phone Data Analysis: A Spatial Exploration towards Hotspot Detection", IEEE Transactions on Automation Science and Engineering, Vol.16, Issue.1, 2019, pp. 351-362.

[10] [10] Sunil Kumar and Maninder Singh, "Big data analytics for the health care industry: impact, applications, and tools", Big Data Mining and Analytics, Vol.2, Issue.1, 2019, pp. 48-57.