International Journal of Engineering & Technology, 7 (2.33) (2018) 666-669

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper



An understanding of machine learning techniques in big data analytics: a survey

S. Josephine Isabella [1] *, Sujatha Srinivasan [2]

[1] Research Scholar, School of Computing Sciences, VISTAS Chennai, India
[2] HOD, Dept. of IT, School of Computing Sciences, VISTAS Chennai, India
**Corresponding author E-mail: josephineisabella@yahoo.co.in*

Abstract

Big data is a Firing Term in the recent era of the modern world, due to the information exploita-tion; there is an enormous amount of data produced. Big data is a powerful momentum of infor-mation and communication technology field due to the effect of growing data in healthcare, IOT, cloud computing, online education, online businesses, and public management. The produced data is not only large but also complex. Big data has a large amount of unstructured data so that there is a need to develop advanced tools and techniques for han-dling big data. Machine Learning is a prominent area of Artificial Intelligence. It makes the system to make intelligent resolutions by giving the knowledge to achieve the goals. This study reviews the various challenges and innovative ideas for big data analytics with machine learning in different fields over the past ten years. This paper mainly organized to identify the research projects based on the discussions over machine learning techniques for big data analytics and provide suggestions to develop the new projects.

Keywords: *Big Data; Big Data Analytics; Machine Learning; Classification; Clustering; SVM.*

1. Introduction

Big data is a greater volume of unstructured and structured data that is too large that it is hard to handle with the traditional tech-niques and tools. Big data analytics means that to examine the data produced by big data sources to upgrade their businesses or stud-ies. The primary goal of big data analytics manages the too large volume, velocity, and variety using various techniques based on intelligence (called Machine Learning). Big data analytics applied to analyze the data presented by various big data platforms.
Machine Learning is a method of giving knowledge to the system with the relevant information and being processed to get the re-sults of some studies. The main theme of machine learning is to combine the concepts of statistical methods, computing knowledge and cross-disciplinary areas to find out the solutions and is applicable to solve the big data problems. Machine learning techniques like classification and clustering play a major role in big data areas.
This paper presents a clear understanding of big data and its tools and the application of various machine learning techniques for big data analytics. Section 1 shows the introduction part of this study. Review about the understanding of big data, big data analytics and machine learning and its techniques are discussed in Section 2. In Section 3 we discuss and summarize the research works of various originators in the field of big data analytics using machine learn-ing. Section 4 has the conclusion part gives the end of the above discussions.

2. Big data - an overview

2.1. Meaning of big data

Big data is a term that sounds everywhere in the world. Due to the growth of internet technology large or enormous amount of infor-mation produced by various devices called big data. A report pro-duced by IDC in 2011 gives that the information explosion gave 1.8 Zeta Bytes and will be increased to nine times during the next five years [1]. Cisco approximated as 6.6 zeta bytes of data pro-duced by sharing of information at the end of this year [2]. From these two reports, we identified that the growth of data to be too large.

2.2. Sources of big data

In the recent era, big data plays an important role in social media such as Google, Face book, and Twitter. Google manages almost 100 PB (PetaByte) and Face Book yields more than ten PetaBytes of log data per month. Some of the big data key elements are Text/Graph/Images produced by Social networking, health care records/images produced by electronic equipment like MRI scan, Genetic information in the field of healthcare, Videos generated by Surveillance of CCTV Cameras, Anomaly detection of Sensor data, Log Files produced by Machine etc.,

2.3. Big data characteristics- v's of big data

The fundamental characteristics of big data have been discussed below:
Volume—This means that the quantity of a large amount of data produced by the digital world and is very critical to find how to store, how it has been processed, analyzed and how it will be pre-sented [2].
Velocity— It relates to incremental of information i.e., speed or rate of production at which data flows within an organization (e.g., organizations dealing with financial information have the ability to deal with this).

Copyright © 2018 S. Josephine Isabella, Sujatha Srinivasan. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Veracity –It means that the available information is reliable and accountable in nature.
Variety—It refers to the types of data in regard to structured and unstructured.
Value—It is very important because how the data being processed or it is peculiar to show how it is going to utilize in each and every sector of production[3].

2.4. Big data tools

During this information sharing world, there is a need for respective tools to handle the big data. Hadoop system having more tools and has to reflect the big data challenges with corresponding solutions as the mirror. They are Apache Hadoop and MapReduce, Storm, Dryad, Splunk, Apache Mahout, Apache Spark [4].

2.5. Big data analytics

The technique used to analyze the too large unstructured or structured, batch or stream data and various sizes of data sets ranging from terabytes to zeta bytes is called Big Data Analytics. Big Data is the data sets that are peculiar in type or size (large size) when compared to traditional data. Analytics refers to extract needful business paradigms or decision designs from a well-defined information/statistics that is to be preprocessed [5].

2.5. Machine learning

Machine learning is a science of understanding the information, analyze the given data set and make the system as intelligent based on the algorithms without human intervention. It is an interdisciplinary area having the key ideas from various different kinds of areas like Artificial Intelligence, knowledge representation, statistical problems. It can also provides solution to mathematics, engineering and scientific problems [6]. Machine learning can be successfully applied to find relationships between multiple features and to find and correct the mistakes occurred during analyses [7]. In recent years, the concept of Machine learning explored to a greater extent due to the growing demand for automating the model of enormous amounts of data. Nowadays people can interact with numerous utilities based on machine learning techniques such as web search engines, speech/face recognition, handwriting recognition, GPS-Navigation tools/sensors and knowledgeable Robots [8].

2.6. Machine learning techniques

Machine learning techniques are the methodologies or algorithms in data mining. Classification and clustering are the two main divisions in the machine learning techniques and are discussed below.
Classification: Classification is a Machine Learning technique and it is used to predict and classify the data for the particular class based on the group of predetermined data[9] [7]. For example, the credit risk for loan applicants is categorized by using the classification model in order to reduce low, medium and high credit values [9].
Major classification algorithms are given by,

1) Support Vector Machine (SVM): SVM used in classification and regression of data sets and used in pattern recognition, data mining and text mining and suitable for optimal dataset problems [9].
2) C4.5 and C5.0 Algorithms: C4.5 is an extension of ID3 classification and produces a decision tree for the given data set. It uses Depth-First growing strategy and suitable for analyzing numeric values, missing values, and noisy data. Also used to prune the growing decision trees. C5.0 algorithm is an extension of both C4.5 algorithm ID3 and is more applicable for big data set and multivalue attribute [10] .
3) Naive Bayes algorithm: It is based on Bayes theorem and is having high scalability and used to build very large data sets in an easy way [9]. The researchers developed algorithms by combining the naive Bayes algorithm along with Map Reduce. Some of the examples are an algorithm developed for sentiment classification with scalability and another one developed for sentimental data of Twitter [1].
4) K-nearest neighbor classifier (KNN) Algorithm: K-NN is a simple algorithm like classification and Regression algorithms and peculiar to instance-based learning and Machine Learning [9].

Clustering: In general clustering means grouping. In the machine-learning trend also have reflected the same meaning as cluster the items based on similar properties or objects. The clustering algorithms categorized into the following manner.

1) Clustering algorithms for partitioning- K-means, K-medoids, K-modes etc., [11].
2) Clustering algorithm for Hierarchical structure- BIRCH, CURE, ROCK, chameleon etc., [11]..
3) Clustering algorithm for Density-DBSACN,DBCLASD, GDBSCAN etc., [11].

Clustering algorithms for big data are used to find the clusters based on similar properties and it is used to split the data set into a number of clusters based on their resemblance. More familiar clustering algorithms are k-means, k-modes, and k-medoids [1].

3. A review of ML techniques for big data analytics

[12] Observed that the introduction of Deep Mining data in big data assist the people to establish decisions in betterment, e.g., In Nov. 2012, the campaign team of Barack Obama scanned the big data(real-time facts) and made the survey about the voters and their tendencies and suggested Obama to be concentrated to arrange the resources to receive voter's attention against to beat Romney in various regions in the U.S. presidential election and also discussed various challenges such as complexity of data, system, and computations leads to find an integrated solution with the iterative algorithms and top-down approach model should be developed with Machine Learning algorithms.
[4] Suggested to follow a workflow diagram of big data projects and gave a sketch of IOT research area. The author briefly discussed various challenges like the storage facility, computational complications, cloud and distributed computing, data visualization, privacy and security for the multilevel data model to be developed in big data analytics and also insisted that to develop tools for extracting data from IoT devices using machine learning. Big data researchers extended to the areas of cloud computing, the computational study of Biological applications and quantum computing.
[1] Found a pathway to the researchers by introducing various big data architectures and big data machine algorithms – clustering and classification algorithms to overlook the big data challenges in healthcare in an efficient manner. He referenced that IKM (Incremental k-means algorithm) with MapReduce produce a solution for multiple scans big data set with one scan data set is a new concept for the research field.
[13] Introduced a research question and conducted a study of finding various error values, the accuracy of classification and the Kappa values by applying various machine learning algorithms and found that PART algorithm gives the better performance than other algorithms. And he suggested that find the issues related to sensor data by applying fuzzy rule based induction algorithms and also handles the online streaming data.
[14] Experimented with an untruthful review spam with an example by introducing the Feature Engineering along with bags of words approach and also introduce Review centric spam detection with machine learning techniques.
[15] Designed a framework for a real-time network traffic anomaly detection system using Machine Learning algorithms like Naive Bayesian, SVM, and Decision Tree. This system is designed with 4 switches along connection window features(A, B,

C,&D) and experimented with UMKC network data produces the feature group C, D of switch 3 having the high accuracy performance.

[16] Discussed about application ML in different categories in health care like Free-text Notes of doctors, stroke prediction diagnoses, CT scan diagnosis and teaches how to apply ML in healthcare application in a stepwise refinement and impressed to develop the health care system based on machine learning techniques in order to fulfill the challenges of biomedical research.

[17] Explored that Magnetic Flux Leakage(MFL) sensors to examine defect nature of the oil and gas pipelines. The MFL sensors spread in every 3 millimeters over the pipelines produces more MFL signals growing as big data and a system developed with ANN algorithm to find different defect types by using Magnetic Flux Leakage(MFL) sensors to examine the oil and gas pipelines. The results obtained that best defect depth estimation accuracy and with better error percentage found in sample data. They motivated to develop the same system with very large data sets(big data).

4. Discussion

The above literature study mainly focuses on the discussions of literature and experimental views of various authors in different fields like health care, education, agriculture, spam detection, education, network anomaly detection using machine learning techniques in big data analytics and also finds some challenges in big data analytics and may be answered by the applying the machine learning techniques to them. The observation shows many of them talk about Hadoop and MapReduce. Both Hadoop and MapReduce play a significant role in big data analytics and machine learning.

Table1 illustrates that some of the studies concentrates to develop a sophisticated system by giving the intelligence Information using ML techniques to the extension of big data.

Table 1: Summary of ML Techniques and Observations

Sl.no	Reference Paper	Algorithm/ Methodology	Dataset	Observation obtained	Future work
1	(Acharjya 2016)	Hadoop, Spark, Storm etc.,	Unstructured data	Improved solutions to big data challenges identified	Techniques to be developed to improve the efficiency and scalability
2	(Daniel 2015)	Learning Analytical framework	Online educational resources	Improvements achieved in educational field	System to be developed to find the consistent policies
3	(Revathy and Lawrance 2017)	C4.5 and C5.0 algorithms	crop dataset	c5.0 produced better accuracy and less memory utilization	System should be developed using c5.0 and MapReduce for extensive data
4	(Zhao et al. 2015)	Naive Bayesian, SVM and Decision Tree	network data from center of UMKC	An anomaly detection system found better accuracy in 4 switches	Anomaly detection system for network management system and complex networks
5	(Crawford et al. 2015)	SLM, SVM, PU-Learning, KDE	Online reviews from Amazon	The online fake reviews leads to imbalanced datasets was found	Feature selection in the big data domain should be developed
6	(Manogaran and Lopez 2017)	Decision tree, SVM, Naïve Bayes etc	Unstructred data,clinical data	Key areas that produces bigdata and algorithms for BDA were found	ML techniques would be applied to health care and medical issues
7	(Mehdiyev et al. 2015)	One R,RIPPER, PART, Ridor,DTNB	Sensor data from accelerometers (phone)	A machine learning model replaces rule patterns identified by manual system	Found the issues related to sensor data by applying fuzzy rule based induction algorithms
8	(Wiebe et al. 2015)	NN, Nearest Centroid	ML benchmark dataset(half-moon)	Computed distance metrics (in quantum computers)	Extended to find quantum K-NN or other learning algorithms
9	(Jin et al. 2015)	Deep Mining, predictive analysis	real-time data, Google search queries	Predictive analysis produced	System to be developed with distributed computations in streaming data
10	(Mohamed et al.2015)	ANN algorithm	MFL signals	Best defect depth estimation accuracy found	System to be developed to increase the defect depth accuracy estimation rate
11	(Haupt and Kosovic 2015)	Sun Cast Solar Power Forecasting System of NCAR	Meteorological data and time duration	NCAR Forecasting system produced the successful forecast results	A forecasting system with renewable energy should be developed

The above discussion gives the following innovative ideas,

a) ML techniques should be developed to analyze the unstructured data and to increase its efficiency and scalability.
b) There is a need to find anomaly detection system for network management system and complex networks in the future.
c) Find the issues related to sensor data by applying fuzzy rule-based induction algorithms and also handles the online streaming data should be developed.
d) Using machine learning techniques identify the mislabeled data, missing values, noisy data and feature selection method used in highly dimensioned data and also find imbalanced data will be a new development for the information growing environment.
e) Big data algorithms (ML techniques) will be applied to healthcare and medical field.
f) Feature selection in the big data domain should be developed using Machine Learning Techniques.
g) A system should be developed by combining C5.0 with Map Reduce to improve the results in analyzing of crop pest data in the agriculture field.
h) A system developed with Machine Learning algorithms to increase the defect depth accuracy estimation rate with minimum error-prediction in big data
i) Future system has to be developed with distributed computations in streaming data for big data.
j) A system with renewable energy should be developed for forecasting.

5. Conclusion

In recent years big data gradually gathered in several fields like web crawling on health care, retail industries, businesses, education, and interdisciplinary research areas like biochemistry, microbiology. This gathering gives the new chances to develop perceptional based innovations using machine learning techniques.

Finally, big data has made a greater impact in every industry and growing into a massive level. From the above understanding, there is a need for research in Big Data Analytics. Based on the above discussions it is essential to develop systems that manage big data in various perspectives such as searching on the internet, business, scientific calculation, and the flow of network data, machine

learning, health care and many more. The traditional machine learning algorithms have applied to handle big data and are used for big data analytics introduces key areas for the new scholarly researchers. The research scholars have an open way to machine learning techniques for Big Data Analytics as a new emerging trend and will make it success.

References

[1] G. Manogaran and D. Lopez, "A survey of big data architectures and machine learning algorithms in healthcare," Int. J. Biomed. Eng. Technol., vol. 25, no. 2/3/4, p. 182, 2017.

[2] A. A. Tole, "Big Data Challenges," Database Syst. J., vol. IV, no. 3, 2013.

[3] B. Daniel, "Big Data and analytics in higher education: Opportunities and challenges," Br. J. Educ. Technol., vol. 46, no. 5, pp. 904–920, Sep. 2015.

[4] D. P. Acharjya, "A Survey on Big Data Analytics : Challenges , Open Research Issues and Tools," vol. 7, no. 2, 2016.

[5] B. Baesens, "Analytics in a Big Data World." p. 232, 2014.

[6] J. Qiu, Q. Wu, G. Ding, Y. Xu, and S. Feng, "A survey of machine learning for big data processing," EURASIP J. Adv. Signal Process., vol. 2016, no. 1, p. 67, Dec. 2016.

[7] T. N. Phyu, "Survey of Classification Techniques in Data Mining," Int. MultiConference Eng. Comput. Sci., vol. I, pp. 18–20, 2009.

[8] N. Wiebe, A. Kapoor, K. S.-Q. I. and, and undefined 2015, "Quantum nearest-neighbor algorithms for machine learning," microsoft.com.

[9] R. Christy Pushpaleela, "Performance Comparison of SVM and C4.5 Algorithms for Heart Disease in Diabetics," Int. J. Control Theory Appl.

[10] R. Revathy and R. Lawrance, "Comparative Analysis of C4.5 and C5.0 Algorithms on Crop Pest Data," Int. J. Innov. Res. Comput. Commun. Eng., vol. 5, no. 1, 2017.

[11] T. Sajana, C. M. Sheela Rani, and K. V. Narayana, "A survey on clustering techniques for big data mining," Indian J. Sci. zTechnol., vol. 9, no. 3, pp. 1–12, 2016.

[12] X. Jin, B. W. Wah, X. Cheng, and Y. Wang, "Significance and Challenges of Big Data Research," Big Data Res., vol. 2, no. 2, pp. 59–64, 2015.

[13] N. Mehdiyev, J. Krumeich, D. Enke, D. Werth, and P. Loos, "Determination of Rule Patterns in Complex Event Processing Using Machine Learning Techniques," Procedia Comput. Sci., vol. 61, pp. 395–401, 2015.

[14] M. Crawford, T. M. Khoshgoftaar, J. D. Prusa, A. N. Richter, and H. Al Najada, "Survey of review spam detection using machine learning techniques," J. Big Data, vol. 2, no. 1, p. 23, Dec. 2015.