International Journal of Electrical and Computer Engineering (IJECE)
Vol. 10, No. 1, February 2020, pp. 575~580
ISSN: 2088-8708, DOI: 10.11591/ijece.v10i1.pp575-580

A survey of big data and machine learning

Surender Reddy Salkuti
Department of Railroad and Electrical Engineering, Woosong University, Republic of Korea

Article Info

Article history:

Received Apr 3, 2019
Revised Sep 12, 2019
Accepted Sep 27, 2019

Keywords:

Big data
Distribution systems
Machine learning
Microgrid
Power and energy
Smart grid

ABSTRACT

This paper presents a detailed analysis of big data and machine learning (ML) in the electrical power and energy sector. Big data analytics for smart energy operations, applications, impact, measurement and control, and challenges are presented in this paper. Big data and machine learning approaches need to be applied after analyzing the power system problem carefully. Determining the match between the strengths of big data and machine learning for solving the power system problem is of utmost important. They can be of great help to plan and operate the traditional grid/smart grid (SG). The basics of big data and machine learning are described in detailed manner along with their applications in various fields such as electrical power and energy, health care and life sciences, government, telecommunications, web and digital media, retailers, finance, e-commerce and customer service, etc. Finally, the challenges and opportunities of big data and machine learning are presented in this paper.

Corresponding Author:

Surender Reddy Salkuti,
Department of Railroad and Electrical Engineering,
Woosong University,
17-2, Jayang-Dong, Dong-Gu, Daejeon 34606, Republic of Korea.
Email: surender@wsu.ac.kr

1. INTRODUCTION

Artificial intelligence (AI) technologies can improve the conversation and cooperation of human and machine. These technologies are used for better interactions between human and machines. Now, we are living in a time of huge information, i.e., an age described by quick gathering of pervasive data. In numerous enterprises, it is developing and giving a way to enhance and streamline business. Numerous fields and segments, going from financial and business exercises to open organization, from national security to logical research in numerous territories, are associated with huge information issues [1]. Huge information has changed the world as far as anticipating client conduct. The introduction of huge information cannot abstain from specifying another current prominent term, interpersonal organizations and the connection between the two is self-evident, yet convoluted. During the last few years, wind, solar, hydro and nuclear power companies have greatly benefited from the power of AI, big data, machine learning and predictive models. They used these technologies to make better predictions, to increase their portfolio's rate of return and to lower their costs [2].

AI came into picture in early fifties and sixties. It was mostly about enabling machines to do things on their own in programming machines which later increased into something like robotics and then in early 90s till 2010 we had this machine learning coming into picture where so many different kind of algorithms and approaches and different kind of theories were discovered and rented in order to begin machine to start learning on their own and then from 2010 onwards a new field which is a subset of artificial intelligence (AI), and machine learning (ML) is the subset of AI, and deep learning is a subset of ML which started in early 2010. ML is an interdisciplinary field which allows us to achieve some sort of AI by using statistical techniques [3].

Huge information and interpersonal organizations are reliant, on the grounds that the majority of present information is produced from person to person communication destinations, yet enormous information is not generally helpful. The real test of huge information is not in gathering it, however in overseeing it and also comprehending it. A few instruments are being intended to better comprehend the part of gigantic measure of information in enhancing business. Analysts and specialists are endeavouring to investigate the eventual fate of huge information to extricate more advantages. Numerical conventional approaches are computationally expensive, and hence it is difficult to use for the on-line security assessment. ML approaches with their learning capabilities, high speed of identifying the potential security boundaries and pattern recognition can offer an alternative method [4].

Reference [5] presents the definition of big data application scenarios through examples in different segments of transportation and energy sectors. Reference [6] proposes a platform which can provide a technical solution to multidisciplinary cooperation of smart grid (SG) monitoring and big data technology. An assessment of distinguished aspects in big data analytics developments in the domain of power systems has been presented in [7]. A new core-broker-client system architecture for big data analytics is proposed in [8]. An overview and potential of big data for smart energy management system is presented in [9].

The ongoing work of application of ML on dynamic security assessment of power systems is addressed in [10]. A comprehensive review of applications of deep learning approaches on machine health monitoring tasks is presented in [11]. Machine learning for hourly solar forecasting application is proposed in [12]. Reference [13] reviews the ML models that are used for condition monitoring in wind turbines. Reference [14] presents an attack detection model for power systems based on ML that can be trained by using information and logs collected by phasor measurement units. A method to address the problem of suggesting the most suitable components for each user by creating a recommender system using intelligent data analysis is proposed in [15]. Reference [16] trains a ML model to predict the duration of big data workloads.

First, this paper presents the concepts of big data analytics and ML techniques. One need to determine the power system problem where we need to apply the ML or big data. If the problem can be solved with classical methods with desired accuracy within the time frame, then there is no requirement of these AI techniques. Generally, the ML is used for forecasting, as there is an availability of lot of data. If we have a rich data set, the ML can narrow down the data based on model based approach or observed data. ML will not give satisfactory results for highly complex and dynamic problems of power systems, such as load flow, contingency analysis, transient stability analysis, etc.

2. CONCEPTS OF BIG DATA AND MACHINE LEARNING

2.1. Big data

Big data is the large sets of data that are computationally analyzed to reveal the trends and patterns related to a certain aspect of data. It is the process to deliver decision-making insights. It uses technology and people to analyze large amounts of data of various types (structured, unstructured and semi-structured data) quickly from a variety of sources to produce a stream of actionable knowledge. Generally, the big data is found in 3 forms, i.e., structured, unstructured and semi-structured [17]. The data which can be stored, accessed and processed in the form of fixed format is defined as structured data. Whereas, the unstructured data is in unknown form/structure, and as the size of this data is huge, it poses several challenges in terms of its processing for extracting value out of it. Semi-structured data can contain both structured and unstructured forms of data [18].

Big data analytics examines the large data sets that contain various data types to reveal unseen patterns, market trends, hidden correlations, customer preferences, and other useful information. The access and utilization of this huge data can be split into six parts, and they are data extraction, storage, cleaning, mining, analysis, and finally data visualization. The characteristics of big data include volume, variety, velocity, variability and value [19]. Volume is the size of data content generated that needs to be analyzed. Velocity is the speed at which new data is generated, and the speed at which data moves. Variety is the types of data that can be analyzed. Veracity is the trustworthiness of the data. Value is the ability to turn big data into clear business value, which requires access and analysis to produce meaningful output.

2.2. Machine learning (ML)

ML is the science which give the computers the ability to learn and predict from the experience without explicitly programmed. If a computer program can improve its performance by learning from previous experience then one can say that it has learned. Machine learning is more closed to data analysis rather than AI. Machine learning uses algorithms that allow computers to iteratively learn from data. In past decades, ML has reached to a new level. ML has given us self-driving car, effective web search, human voice

recognition, image recognition and many more [20]. Every day we use it several times without knowing it. The process of ML is depicted in Figure 1.

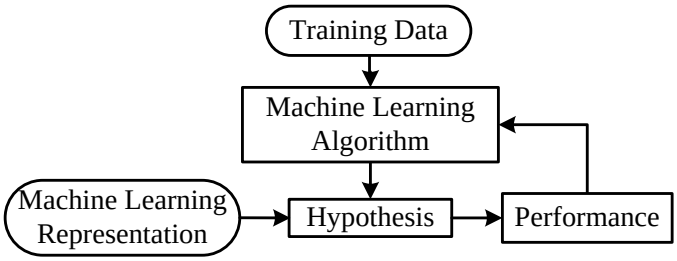

Figure 1. Machine learning (ML) process

It is like predicting the future. The companies would already know what kind of decision you are going to take any given certain situations. If certain parameters, they are provided to you what would be your reaction because they know in past you have done something like this and you have a set of data your data is with them that every time when you have into such kind of environment, you do certain kind of actions companies can predict it and this is what is done as a part of machine learning, where data is provided feature extraction takes place. Then a predictive model is calculated and this predictive model is then rolled out for the users for which data was taken [21].

ML consists of two phases, i.e., learning and prediction phases. Figure 2 depicts learning phase of ML process. Supervised, unsupervised, semi-supervised and reinforcement learnings are the some popular machine learning types. Supervised learning requires humans to train by providing inputs and desired output [22]. Unsupervised learning is opposite to the supervised learning, which learns by its own without any labeled response. In the supervised and unsupervised leanings there is either labeled data or unlabeled data, whereas semi-supervised learning uses both labeled and unlabeled data for training. Reinforcement learning algorithms learns by trial and error method in which actions yield greatest reward. This algorithm is a ML as well as branch of AI. The prediction phase of ML uses the developed model and the new data is fed to this model. The predicted data is going to be available by using the ML algorithm [23].

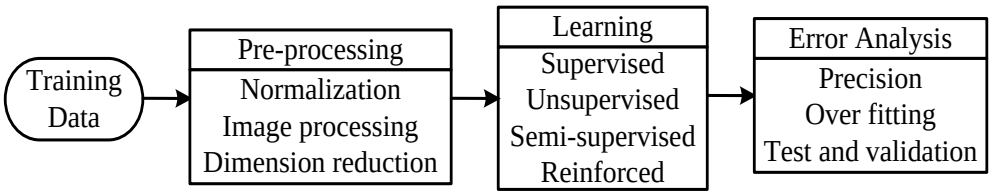

Figure 2. Learning phase of machine learning (ML) process

Some of the common ML algorithms used include regression analysis, clustering, association rule mining and collaborative filtering. Regression analysis helps to predict the relation between two variables (i.e., dependent and independent variables). In the clustering algorithm, the classifier tries to find some structure from given data set without known classification. Association rule mining tries to find the association between objects, also known as affinity analysis. It is used to find what goes with what [24-25]. It is basically a set of analytical techniques that is used to uncover connections and associations between objects. Collaborative filtering is a technique of predicting user preferences based on the preference of group of users. This technique is used when we need more complex information than keyword.

3. BIG DATA AND MACHINE LEARNING (ML) APPLICATIONS

Industries which uses large amounts of data have recognized the value of big data and ML. By working with big data and ML organizations can work efficiently and can gain advantage over competitors [18]. The applications of big data and ML are presented in the following subsections.

3.1. Big data applications

The major applications of big data are presented in Table 1 [26].

Table 1. Applications of big data in various industries

Industries	Applications of Big data
Electrical power and energy	Generation systems, distribution and utilities systems: electricity theft detection, detection of electric vehicles, phase connectivity identification, transformer to customer association; smart grid applications such as wide area situational awareness, event classification and detection, transient power prediction, fault detection/ prevention, forecasting weather, load demand, wind speed and solar irradiation data, control problems, state estimation and to support the participation of market agents in electricity markets [27].
Health care and life sciences	Disease pattern analysis, chain management, clinical trials data analysis, drug discovery and development analysis, patient care quality and program analysis. Big data uses patient's data that healthcare organizations have to improve pharmaceutical sales, patient analysis, and allow for better payer solutions.
Government	Government uses big data to handle the security threats, population dynamics, budgeting and finance operations, etc.
Telecommunications	Big data is required for network capacity planning and optimization by correlating network usage, subscriber density, along with traffic and location data. Call detail record analysis, mobile user location analysis, customer churn prevention, revenue assurance and price optimization and network performance and optimization.
Web and digital media	Abuse and click-fraud prevention, large scale click stream analytics, campaign management and loyalty programs, social graph analysis and profile segmentation.
Retail/consumers	Big data is required for analyzing customer buying behaviour, current products, pricing, and promotions. Market and consumer segmentations, event and behaviour-based targeting, supply chain management and analytics, merchandizing and market based analysis.
Finance and fraud services	Fraud detection and security analytics, risk analysis and management, credit risk, scoring and analysis, compliance and regulatory reporting.
E-commerce and customer service	Event analytics, right offer at right time, next best option or next best action, cross-channel analytics.

3.2. Machine learning (ML) applications

As explained earlier, ML is a type of AI that provides computers the ability to learn without being explicitly programmed. It is a subset of AI, which has behavioural rules by examining and comparing large data sets to determine the common patterns [28]. This approach is very effective for solving classification problems. Nowadays, sensors generate a lot of data, and it is very difficult to analyze this data for practical applications using classical approaches. In this case, ML is the best option. It is used for developing stochastic models. ML has many practical applications that results in time and money saving. With ML things can be done more quickly and efficiently. It automates the task such as changing password, checking balance which would otherwise need a live agent, which as a result save the valuable time of the agent that can be used to focus on other important tasks that humans perform best [29].

In most classical models, there are underlying assumptions that they are made to lend the problem to classical analysis. ML makes the performance better by using heuristics to replace these assumptions. ML technique is applied to a data driven estimated model. Therefore, accuracy/error limit is an important factor [30]. The key factors which increased the importance of machine learning are the data availability and computation power. These two factors have increased the excessive applications of ML across various fields. The major applications of machine learning for various fields are presented in Table 2 [31].

Table 2. Applications of machine learning (ML) in various industries

Industries	Applications of Machine Learning
Electrical Power and Energy	Load modelling, anomaly detection and control, forecasting (load, wind and solar PV), image recognition, nonlinear programming, recommendation, fault location estimation, load and generator circuit model, dynamic model parameter estimation, exciter and governor models, etc.
Transportation	Transportation is the main industry which is very much affected by ML. Many big companies like Google, Tesla and others are investing too much in self-driving cars. Tesla cars are already equipped with autopilot feature which uses a system of 8 cameras and 12 sensors which provides 360 degree view to a range of 250 meters [32].
Education	ML helps to organize and optimize the content module and track student knowledge and recommend next step. ML and AI can easily identify the learning disability. It can collect student's work and analyze the difficulties faced by the students. It can be used to generate tests and assignments.
Financial Services	Financial services like banks and other business used ML to increase profit, investment and to prevent fraud. ML algorithms can easily detect frauds and can flag them to the security team.
Marketing and Sales	ML is useful for fraud detection in transactions between buyers and sellers. E-commerce websites uses ML to recommend items based on previous items purchased or searched.
Healthcare	ML is useful for early detection of disease and to understand risk factors for a disease. These algorithms can easily identify diseases within a fraction of time taken by doctors. It takes years of training for doctors to identify any disease a machine can learn to identify it in hours. AI also works as virtual doctor. It listens to the patient's problem and suggests the treatment [33].

4. CHALLENGES AND OPPORTUNITIES OF BIG DATA AND MACHINE LEARNING (ML)

Big data brings new challenges, and requires new approaches to deal with the challenges. The challenges to big data include performance, data federation, data cleansing, security and time to value. Most common actions performed on data sets using big data analysis includes capturing data, storing data, data analysis, updating, querying and visualizing the data. Challenges of big data for electrical power and energy industry include privacy, data mining, integration of data, cyber security, and demand prediction through analytics processing in smart grid (SG) applications, data quality and cost balance, industrial fault diagnosis using big data and quantum cryptography for data security in smart grids. Some challenges lying ahead in the terms of SG big data technology includes multisource data integration and storage, real-time data processing technology, data compression, big data visualization technology, and data privacy and security [34]. State of art, current status and recent developments of big data are:

a. Big data technique to handle a large amount of information in a short time using meter data management.
b. Big data requirements and enhancements throughout the power network dispatching and planning.
c. Schedulable capacity forecasting technique for thermostatically controlled load by big data analysis.
d. Concept of device electrocardiogram in fault diagnosis using big data.
e. Tensor based big data management scheme in smart grid (SG) systems.
f. Artificial neural network (ANN) approach in efficient electricity generation forecasting.

Some of the benefits from big data analytics in SG includes increased system stability and reliability, increased asset utilization and efficiency, and better customer experience and satisfaction [35]. Machine learning is a subset of AI which includes abstruse statically approaches that enable machines to improve at tasks with experience. The category includes deep learning. Hence, it is a subset of AI in which machine they get improved kind of decision making experience depending upon the training or the data they have and it is based upon deep learning [36]. The main challenges of machine learning are: generative vs discriminative learning, beyond classification and regression, learning from non-vectorial data, machine learning bottlenecks, intelligible models, combining learning methods, distributed data mining, unsupervised learning comes of age, and more informed information access.

5. CONCLUSIONS

A detailed analysis of big data and machine learning (ML) in electrical power and energy sector including the smart grid has been presented in this paper. Big data analytics involves the processes of searching a database, mining, and analysing data dedicated to improve the performance of the company. ML focuses on the development of computer programs that can teach themselves to grow and change when exposed to the new data. Applications of big data and ML in various industries such as electrical power and energy including smart grid, transportation, health care, education, e-commerce, financial services, marketing and sales, etc. Various challenges and opportunities related to big data and machine learning are also reviewed in this paper.

ACKNOWLEDGEMENTS

This research work has been carried out based on the support of “Woosong University's Academic Research Funding - 2019”.

REFERENCES

[1] R. J. Bessa, “Chapter 10 - Future Trends for Big Data Application in Power Systems,” *Big Data Application in Power Systems*, pp. 223-242, 2018.
[2] Y. Zhang, *et al*., “A big data driven analytical framework for energy-intensive manufacturing industries,” *Journal of Cleaner Production*, vol. 197, pp. 57-72, 2018.
[3] “How Machine Learning, Big Data, & AI Are Changing Energy,” [Online], Available: https://rapidminer.com/blog/machine-learning-big-data-ai-energy/
[4] N. V. Tomin, *et al.*, “Machine Learning Techniques for Power System Security Assessment,” *IFAC-PapersOnLine*, vol. 49, pp. 445-450, 2016.
[5] S. Rusitschka and E. Curry, *Big Data in the Energy and Transport Sectors*, in Cavanillas J., Curry E., Wahlster W. (eds), “New Horizons for a Data-Driven Economy,” Springer, Cham, 2016.
[6] Y. Guo, *et al*., “Complex Power System Status Monitoring and Evaluation Using Big Data Platform and Machine Learning Algorithms: A Review and a Case Study,” *Complexity*, vol. 2018, pp. 1-21, 2018.
[7] H. A. Hejazi and H. M. Rad, “Power systems big data analytics: An assessment of paradigm shift barriers and prospects,” *Energy Reports*, vol. 4, pp. 91-100, 2018.
[8] T. Wilcox, *et al.*, “A Big Data platform for smart meter data analytics,” *Computers in Industry*, vol. 105, pp. 250-259, 2019.

[9] K. Zhou, *et al*., "Big data driven smart energy management: From big data to big insights," *Renewable and Sustainable Energy Reviews*, vol. 56, pp. 215-225, 2016.
[10] E. M. Voumvoulakis, *et al*., "Application of Machine Learning on Power System Dynamic Security Assessment," *International Conference on Intelligent Systems Applications to Power Systems*, Toki Messe, Niigata, pp. 1-6, 2007.
[11] R. Zhao, *et al*., "Deep learning and its applications to machine health monitoring," *Mechanical Systems and Signal Processing*, vol. 115, pp. 213-237, 2019.
[12] G. M. Yagli, *et al*., "Automatic hourly solar forecasting using machine learning models," *Renewable and Sustainable Energy Reviews*, vol. 105, pp. 487-498, 2019.
[13] Stetco, *et al*., "Machine learning methods for wind turbine condition monitoring: A review," *Renewable Energy*, vol. 133, pp. 620-635, 2019.
[14] D. Wang, *et al*., "Detection of power grid disturbances and cyber-attacks based on machine learning," *Journal of Information Security and Applications*, vol. 46, pp. 42-52, 2019.
[15] A. J. F. García, *et al*., "A recommender system for component-based applications using machine learning techniques," *Knowledge-Based Systems*, vol. 164, pp. 68-84, 2019.
[16] Á. B. Hernández, *et al*., "Using machine learning to optimize parallelism in big data applications," *Future Generation Computer Systems*, vol. 86, pp. 1076-1092, 2018.
[17] R. Arghandeh and Y. Zhou, *Big Data Application in Power Systems*, Elsevier Science, 2018.
[18] J. Lee, *et al*., "Data Analysis for Solar Energy Generation in a University Microgrid," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 8, pp. 1324-1330, 2018.
[19] https://www2.microstrategy.com/producthelp/10.10/WebUser/WebHelp/Lang_1033/Content/mstr_big_data.htm
[20] M. Farhadi and N. Mollayi, "Application of the least square support vector machine for point-to-point forecasting of the PV power," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 9, pp. 2205-2211, 2019.
[21] V. Malbasa, *et al*., "Voltage Stability Prediction Using Active Machine Learning," *IEEE Transactions on Smart Grid*, vol. 8, pp. 3117-3124, 2017.
[22] Rahman, *et al*., "Power disaggregation of combined HVAC loads using supervised machine learning algorithms," *Energy and Buildings*, vol. 172, pp. 57-66, 2018.
[23] Md. A. Rahman, *et al*., "A Survey of Machine Learning Techniques for Self - tuning Hadoop Performance," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 8, pp. 1854-1862, 2018.
[24] S. Preda, *et al*., "PV Forecasting Using Support Vector Machine Learning in a Big Data Analytics Context," *Symmetry*, vol. 10, 2018.
[25] C. S. Sindhu and N. P. Hegde, "A Novel Integrated Framework to Ensure Better Data Quality in Big Data Analytics over Cloud Environment," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 7, pp. 2798-2805, 2017.
[26] R. Eskandarpour and A. Khodaei, "Machine Learning Based Power Grid Outage Prediction in Response to Extreme Events," *IEEE Transactions on Power Systems*, vol. 32, pp. 3315-3316, 2017.
[27] G. Bathla, *et al*., "A Novel Approach for Clustering Big Data based on MapReduce," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 8, pp. 1711-1719, 2018.
[28] W. Xiang, *et al*., "Machine learning based optimization for vehicle-to-infrastructure communications," *Future Generation Computer Systems*, vol. 94, pp. 488-495, 2019.
[29] C. Tu, *et al*., "Big data issues in smart grid – A review," *Renewable and Sustainable Energy Reviews*, vol. 79, pp. 1099-1107, 2017.
[30] B. A. S. Leech, *et al*., "Big Data issues and opportunities for electric utilities," *Renewable and Sustainable Energy Reviews*, vol. 52, pp. 937-947, 2015.
[31] R. Y. Zhong, *et al*., "Big Data for supply chain management in the service and manufacturing sectors: Challenges, opportunities, and future perspectives," *Computers & Industrial Engineering*, vol. 101, pp. 572-591, 2016.
[32] D. Radhika and D. A. Kumari, "Misusability Measure Based Sanitization of Big Data for Privacy Preserving MapReduce Programming," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 8, pp. 4524-4532, 2018.
[33] R. A. Archana, *et al*., "A Study on Big Data Privacy Protection Models using Data Masking Methods," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 8, pp. 3976-3983, 2018.
[34] E. Hossain, *et al*., "Application of Big Data and Machine Learning in Smart Grid, and Associated Security Concerns: A Review," *IEEE Access*, vol. 7, pp. 13960-13988, 2019.
[35] T. Yuan, *et al*., "HyperOXN: A Novel Data Center Topology Driven by Machine Learning," *13th APCA International Conference on Automatic Control and Soft Computing*, pp. 573-578, 2018.
[36] M. K. Saggi and S. Jain, "A survey towards an integration of big data analytics to big insights for value-creation," *Information Processing & Management*, vol. 54, pp. 758-790, 2018.