Chapter 1
Introduction

Bilal Jan, Haleem Farman and Murad Khan

Abstract Recently, deep learning techniques are widely adopted for big data analytics. The concept of deep learning is favorable in the big data analytics due to its efficient use for processing huge and enormous data in real time. This chapter gives a brief introduction of machine learning concepts and its use in the big data. Similarly, various subsections of machine learning are also discussed to support a coherent study of the big data analytics. A thorough study of the big data analytics and the tools required to process the big data is also presented with reference to some existing and well-known work. Further, the chapter is concluded by connecting the deep learning with big data analytics for filling the gap of using machine learning for huge datasets.

List of Acronyms

AI	Artificial intelligence
ANN	Artificial neural networks
HDFS	Hadoop Distributed File System
M2M	Machine to machine
IoT	Internet of things
CPS	Cyber physical systems
ICN	Information-centric networking
WSN	Wireless sensor network

1.1 Machine Learning

The need of machine learning was felt when artificial intelligence (AI)-based systems were facing difficulties with hard-coded programs, and it was suggested that machines should be able to extract patterns from the data by itself without the involvement of human or programs for specific tasks. The idea of machine learning was

M. Khan et al., *Deep Learning: Convergence to Big Data Analytics*, SpringerBriefs in Computer Science, https://doi.org/10.1007/978-981-13-3459-7_1

introduced in 1959 by Arthur Samuel (field expert) that instead of programming machines for specific tasks, computers should be able to learn themselves (BMC blog 2018). Machine learning is the subset of artificial intelligence, in which system can adjust its activities and react to specific situation when provided with large amount of data (Ahmad et al. 2018a, b). In machine learning, systems are trained to act accordingly. These systems are provided with many examples specifically related to task, and statistical structures are identified that leads system to define rules for that particular task. Machine learning deals with large amount of datasets (Khumoyun et al. 2016), for instance, medical dataset containing millions of patient images for different diseases. There are many applications such as recommendation and navigation systems using machine learning giving more accurate and efficient result as compared to hard-coded programs. Machine learning is classified into supervised learning and unsupervised learning (BMC blog 2018).

1.1.1 Supervised Learning

In supervised learning, the machine is trained according to the given input data and output is drawn from it accordingly. The relation between input data and expected output is approximated through machine learning algorithms. In simple words, supervised learning is based on the prior knowledge that what our output will be. Normally supervised learning is done in two ways, either as classification or regression. In classification, the input is mapped to output tags, while in regression the input is mapped to a continuous output. The goal of both ways is to find structures in input data that can be transformed into effective, accurate, and correct output. In real world, it is not always the case that the data labels will always be correct. Incorrect data labels can lead to ineffective output and will clearly affect the effectiveness of the model. Most common algorithms are naive bayes, artificial neural networks, support vector machines, and logistic regression (Towards Data Science 2018).

1.1.2 Unsupervised Learning

Unsupervised learning works on input data only by not considering the labeled output. It finds structures in the data without having expected output. Unsupervised learning algorithms are complex and time-consuming because of generating output without prior knowledge. Clustering, dimensionality reduction, and representation learning are the most commonly used tasks in unsupervised learning. The most common algorithms used in unsupervised learning are K-means clustering, autoencoder, and principal component analysis. Dimensionality reduction and exploratory analysis are the two most common use cases for unsupervised learning. In dimensionality analysis, feature or columns are reduced to represent data, while in exploratory analysis, structure in data can be automatically identified using unsupervised learning.

The trend toward machine learning is getting bigger and bigger as it is assisting business a great deal such as machine learning in healthcare, financial services, vehicle automation, and retails. Data in digital form is generated in huge volume and variety. It is not important to gather more data, but how effectively and efficiently can we use this data is the problem to be solved. Machine learning systems are really helping companies to manage, analyze, and produce output from such huge and diverse data. Machine learning systems can identify the hidden patterns in data, and customer preferences can be identified to boost business, market trends, and many more ways in which business can be boost. As the data is growing data by data, companies such as Facebook and Google are more interested in customer behavior to improve their service. To handle big data, a subset of machine learning known as deep learning is used, in order to use the available data in more effective and efficient manner.

1.2 Deep Learning

Deep learning uses artificial neural networks (ANN) that got inspiration from the neuron present in human brain. It consists of layers, and the word deep refers to the depth of layers (Ahmad et al. 2018a, b). Initially, the word deep referred to very few layers, but due to the use deep learning in complex problems, the number of layers is in hundreds and even thousands. Deep learning is very successful in many domains such as image processing, healthcare, transportation, and agriculture; with the help of deep learning, more and more data can be utilized in best possible and it is getting popularity due to the availability of trained dataset, for instance, ImageNet (Deng et al. 2009) that contains thousands of images. Secondly, the low-cost GPUs are in more use to train data and can avail the services of cloud as well. Giant companies such as Facebook, Amazon, Google, and Microsoft are using deep learning techniques to analyze the huge amount of data on daily basis.

Deep learning got high attention not only from researcher but from techno-companies as well. Social media companies such as Facebook, Twitter, and YouTube generate large volume of data on daily basis due to the number of users they have (Jan et al. 2017). It is very important for them to handle this huge data often termed as "Big Data." By using traditional data analysis tools, it is very difficult and almost impossible to have a good insight of the data and extract the meaningful data from it. Even machine learning technique will not work that much efficient as required; therefore, deep learning is used in order to analyze deep in the network to extract meaningful, accurate, and precise information. In the coming chapter, we will be studying deep learning in detail.

1.3 Conventional Data Processing Techniques

Data processing is a sequence of steps for transforming raw data into meaningful information or insights as an output in the form of plaintext file, table/spreadsheet, charts and graphs, maps/vector, or image file leading to the solution of a problem or improving an existing situation. It leads to solving a problem or improving an existing situation. Data collection, preparation as suitable input, processing, output and interpretation and storage are some core steps of data processing. It is important to process data for businesses and scientific operations where large volumes of output is required out of business data after repeated processing while scientific data requires fast-generated outputs governing numerous computations. There are many traditional data processing techniques briefly discussed as follows (BMC blog 2018; Ahmad et al. 2018a, b; Khumoyun et al. 2016; Towards Data Science 2018):

Manual Data Processing: In this method of data processing, all logical operations, calculations, data transfer from one place to another, and required results are obtained without intervention of any tool or machine. Manual data processing is performed in small firms and government institutes for their tasks manually but is avoided to be used due to error-pruned, time-consuming, and labor-intensive nature of this processing. The absence of technology or its high cost favored the use of manual data processing, but advancement of technology has drastically decreased the dependency on manual techniques.

Mechanical Data Processing: This method is more accurate and faster than manual data processing by using devices such as mechanical printers, typewriters, or other mechanical devices to work for printing press or examination boards. More sophisticated and better computing powered machines surpassed them.

Electronic Data Processing: It is a modern data processing technique, very fast and accurate, with the involvement of a computer that automatically processes the data according to the provided set of instructions along with it. Computer is another name of electronic data processing machine. Batch processing, online processing, real-time processing, multiprocessing, and time sharing and distributed processing are some methods of data processing by electronic means. Some methods are briefly discussed as follows:

- Batch Processing: It is the simplest form of data processing which works for large volume of data if data could be clumped into one or two categories; e.g., daily or weekly transactions of a store can batch-processed, and results are sent to the head office.
- Online Processing: In this method, directly attached equipment to a computer and Internet connections are utilized to allow data to be stored at one place and to get it used for processing at different places. This type of processing is used in cloud computing.
- Real-Time Processing: For an instant turnaround, this method is faster than batch processing; e.g., instant update of record is required in case of canceling a reservation or buying an airline ticket.

- Distributed Processing: It is widely used in a scenario where one server is connected with remotely located workstations, e.g., ATMs where fixed software is located at a particular place while all the end machines make use of that same set of instructions and information.
- Data Mining: Data mining collects data from different sources and combines it to analyze to find patterns, to group similar observations, to predict class label of previously unknown instances based on historical data, and to identify deviated values from the normal ones.
- Neural networks, nonlinear data modeling tools, are automated to store, identify, and find patterns in database or to decipher complex relationships between inputs and outputs.

1.4 Data Mining Techniques: A Big Data Analysis Approach

Availability of gigantic amount of data due to enormous expansion of information technology arises the strong demand for data mining techniques which play a vital role in the analysis of data using techniques of statistics, artificial intelligence, database system, and machine learning. Data mining is a computing process to explore large datasets to discover interesting patterns. It has mainly three steps (Towards Data Science 2018; Jan et al. 2017; Five Data 2018):

Walking Through: Nature of the data is determined, and it is converted into a suitable form after going through all the collected data. Educated guesses and different types of imputation method such as regression substitution, average imputation, and multiple imputation are used to substitute values to replace missing data.

Identifying Patterns: By understanding what information to extract, patterns are checked that could lead to make predictions.

Outcome Planning: Here, desired outcome is planning having patterns on hand.

On completion of data mining process, one can utilize it in a number of applications such as to recommend new products based on the patterns discovered within a data and helps to identify clusters showing the similarities and dissimilarities within the data and classifying the data to draw useful insights from it.

There are a number of data mining techniques for achieving business goals (Ahmad et al. 2018a, b).

Association Rule Learning: Here, the frequently occurring items or variables and association or relationship of two or more items are analyzed in order to uncover the hidden patterns in the large datasets. Various applications can be drawn utilizing this technique; e.g., the trend of buying bananas along with cereal in customers' purchases record can help to make better decision of placing bananas closer to cereals in the store shelf to boost selling these items together. Also, effect of coupons or sales offer on selling can be tracked.

Clustering Analysis: Similar objects exhibiting same behaviors are grouped and analyzed with this old yet popular unsupervised technique of data mining. K-means clustering, agglomerative clustering, and hierarchical clustering are some well-known clustering approaches.

Classification Analysis: It is supervised way of classifying large sets of data by predicting a class label to unknown instances based on the values in the historical database. In other words, it helps to find the categories the data belongs to. The classification methods make use of decision tree and neural network techniques. Models are built by dividing the data into two datasets—one for training or building the model and another to test the built models which are then compared with predicted values in order to be evaluated. For example, classification analysis is performed by our email provider to classify our incoming email as legitimate or spam based on some specific words of the email or its attachment.

Regression Analysis: It explores the dependency between variables b assuming a one-way causal effect of a variable in response of another. In contrast to correlation, here dependency is not necessarily from both sides; one variable can be dependent on other but not vice versa.

Anomaly or Outlier Detection: Various techniques are governed to detect, analyze, and understand anomaly, outlier, or an exception in the data that deviates from a particular dataset.

Data mining evolving nature of data has prompted the need to develop new analytic techniques by expanding already existing data mining techniques (Khumoyun et al. 2016). Social media data in the form of comments, likes, browsing behaviors, tweets, opinions, e-commerce, and medical data is some digital sources of the big data. Significant contribution of numerous sensors also played a vital role in the growth of big data (Jan et al. 2017).

Life cycle of big data analytics consists of various steps such as problem definition, problem tackling based using existing techniques, Human Resource requirement estimation, data acquisition, munging and storage, then performing exploratory analysis of data in order to select decision model and finally converting the designed and evaluated model into implementation (Deng et al. 2009; Data Processing and Data Processing Method 2018).

Traditional database management tools are not capable to store and process tremendous volume of big data (Jan et al. 2017). Hadoop and Mahout are popular big data analytics tools.

Apache Hadoop is a framework based on MapReduce programming model that provides the solution by supporting distributed data processing. Large datasets are processed in parallel across clusters of computers. Hadoop Distributed File System (HDFS) and MapReduce are two main parts of Hadoop where former is a distributed file system consisting thousands of nodes used for big data storage primarily, while latter is software that processes these nodes in parallel or simultaneously and retrieves data at a faster rate. Classification algorithms are implemented in Apache Mahout, which are applied against the MapReduce paradigm to estimate the values of the subjects under consideration.

MapReduce is a programming model encompassing map and reduce as two operations defined by the users to generate and process massive datasets stored in HDFS by allowing parallel large computations boosting the speed and reliability. Divide and conquer is the basic technique behind it where smaller chunks of the given big data problem are made, shuffled, and reduced to acquire the desired result as output. More specifically, all the complex business rules, logic, or costly code are specified first in the Map phase; then, lightweight processing, e.g., summation or aggregation, is specified in the second Reduce phase. There are multiple phases of data processing with different components by Hadoop MapReduce discussed as follows (Jan et al. 2017; Explain three methods 2018).

Step 1. A Master process is forked by the user and processed by a number of workers.

Step 2. The Map and Reduce tasks are assigned to the worker processes by the Master process.

Step 3. The input file is split into chunks of 16–64 MB by the user program of the MapReduce library. One or more input file(s) chunks are assigned to each Map task.

Step 4. Then, the data chunks are turned into a sequence of key–value pairs by these Map tasks for reading by the mapper.

Step 5. Mapper produces intermediate key–value pairs from the input record produced in the previous phase. The intermediate output is totally different from the input pair and is forwarded to the combiner for next processing.

Step 6. Combiner acts as mini-reducer to aggregate mapper's output locally by minimizing the data transfer between reducer, and mapper then forwards the output to the next practitioner step.

Step 7. Here, partitioning is performed on the output of the combiner on the basis of the key in MapReduce to evenly distribute the map output over the reducer.

Step 8. Physical movement of the data, also known as shuffling, over the network to the reducer nodes is done at this stage, and intermediate output is merged and sorted and then is shipped as input to the reduce stage.

Step 9. In this phase, intermediate key–value pairs generated by the mappers are considered as the input by the reducer and reducer function is applied on each of them to produce the final output stored on HDFS.

Step 10. In the last stages, RecordWriter writes the produced output key–value pairs from the reducer phase to the output files. Output format specified the method of writing these output key–value pairs on HDFS.

Virtual machine is a big data processor platform for analysis and processing the data (Data Processing 2018). It is a platform for machine learning and analytics optimized for the cloud.

1.5 Big Data Analytics

During the past 20 years, data has been enormously generated due to the interconnecting devices in various forms such as machine to machine (M2M), Internet of things (IoT), cyber physical systems (CPS), and information-centric networking (ICN).

However, the very formal definition was first introduced by D. Laney, META Group in 2001, by defining the three Vs architecture of big data as volume, velocity, and variety (Laney 2001). For a decade, the researchers follow the three Vs architecture; however, it was changed after 2010 when Dave Beulke and Associates introduces the five Vs architecture such as volume, velocity, value, veracity, and variety in 2011 (Associates 2011). The definition of big data consists of 6 Vs architecture that is put forward by Enterprise Strategy Group in 2012 (Group 2012). Similarly, the concept of 8 Vs, i.e., volume, velocity, value, variability, veracity, viscosity, virality, and variety, is introduced by Vorhies (2014). However, all types of data do not follow any of the aforementioned architecture as standard. It entirely depends on the researchers, scientists, and academia, who are extracting data for various purposes. Similarly, big data is not only used to process data and come up with necessary results. It is also used to find the correct tools and identify the applications for big data in smart environments such as smart homes, smart cities, smart e-health systems. Big data is defined for huge amount of data, and thus, conventional data tools and processing applications are not sufficient to analyze the big data. The market of big data involves in three layers of processing, i.e., 1) the infrastructure layer (particularly related to hardware and physical equipment), 2) the data management, organization, and analytics layer, and 3) the services layer. The infrastructure layer consists of hardware components such as cloud facilities, networking services, communication technologies. The gathering, collection, and acquisition of data normally happen at the infrastructure layer. It is sometime done with the help of sensors which further needs an optimized and efficient wireless sensor network (WSN). After the data is collected at the infrastructure layer, the data is communicated with the management layer with the help of underlying technologies. The management and analytics layer processes the data for various activities such as storing, decision-making, pattern recognition. Finally, the data is disseminated among the users with the help of services layer. The services layer also offers other functionalities such as defining the applications and interfaces where user can connect to the system to retrieve the required information. Figure 1.1 shows the functionalities and concepts provided by all the three layers.

There are various challenges involved in big data analytics such as processing data in real time, disseminating data with users, and providing them interfaces. Similarly, most of the recent literature used big data for different applications such as energy management in smart homes, cities, and buildings and urban planning. However, that is no generic standard available for analyzing big data apart from above three layers. There is need of standard solutions and regulation for analyzing and processing big data. The world famous firms such as ACM and IEEE are working hard to come up with common and generic standard for big data. However, at this stage, we will appreciate the efforts of researchers, scientist, and academia in this regard. Further, the big data is a part of data science and we cannot say that one particular algorithm can be used to process big data. Big data is a linear process, and at each step, one can use a particular type of algorithm to solve a problem. With the passage of time, various algorithms and methods have been employed by the researchers to make the process of analyzing big data more efficient and convenient for the users and

Fig. 1.1 Three layers of big data

practitioners (Khan et al. 2018a, b). The processing of data in real time required high computing resources and efficient algorithms such as deep learning, machine learning, and transfer learning. However, implementing these techniques in real time is a challenging job and the researchers are looking for various computing mechanisms such as artificial intelligence and quantum computing. As the computing system is advanced every day, there is a possibility in future to reach to certain level of processing data in real time. For example, if a space shuttle or space vehicles sent to Mars for collecting data and processing it in real time would require high-end programming and pattern matching techniques. On top of the computation capabilities, the data analytics are quite important which is however impossible with conventional data analytics. Similarly, the conventional MapReduce architecture used in the Hadoop ecosystem is mainly used to process off-line data, which is therefore nearly impossible to be used for processing real-time data. The Hadoop ecosystem alongside Spark somehow addresses the processing of data in real time. However, when it comes to huge amount of data specifically in petabyte, the performance of the Spark also degraded exponentially. Therefore, a cluster system combined of Spark can be used to process the data in petabytes (Silva et al. 2017). Similarly, integrating data sources for solving a problem or even answering a question requires big data analytics. For example, understanding the pattern of why the rainfall rate is dropping lower than the average rainfall every year may require obtaining the data from satellites, sensors, and airborne and fusing them together for a better approach. Similarly, obtaining the

data of vehicles moving on some particular roads and combining them for analytics might resulting various patterns such as why the number of cars are high on a particular time of the day and so on. All these challenges discussed above can be answered with proper data analytic techniques, tools, and methods.

1.6 Deep Learning in Big Data Analytics

Machine learning is widely studied recently for processing big data in real time. Similarly, the machine learning techniques can be considered an essential part of the overall process of big data analytics. In such process, a computer system is trained with a pattern and then the computer system finds the same pattern or similar pattern in the entire data. This particular type of learning where the computer system is taught at the beginning is called supervised learning. However, training a computer system in real time is challenging task to accomplish with even hundreds and thousands of examples. On the hand in an unsupervised learning methodology, a process is applied to classify and manage the information in unstructured documents. Similarly, an analysis can be carried out to find natural divisions and clustering in the data. Further, the unsupervised learning looks for inherent sequence and patterns in the data and groups them together. Furthermore, in such processes, the outliers can be identified for grouping the data falls outside of a particular sequence or pattern. Similarly, the unsupervised learning can be applied to the existing data to find real-time solutions to disasters, heavy rainfall, etc. A number of approaches have been proposed to use machine learning for big datasets alongside Hadoop and MapReduce architectures (Dean and Ghemawat 2004; Shvachko et al. 2010). Similarly, the online and deep learning is also adopted alongside machine learning to overcome the challenges of big data. Summarizing the challenges present with big data using machine and deep learning, we come across unstructured data obtained from heterogeneous sources, high and fast streaming data, noisy and poor-quality data, high-dimensional data, and data with limited labels (Najafabadi et al. 2015; Sukumar 2014). Before applying machine learning techniques to big data analytics, one can have enough knowledge of statistical methods and signal processing techniques. The signal processing techniques play an important role in classifying data obtained from heterogeneous sources and inputs. Qiu et al. presented a work with identifying signal processing techniques to address various challenges such as large-scale, different data types, high-speed data, incomplete and uncertain data, and density of data (Qiu et al. 2016). However, dealing with large-scale data using machine learning requires high memory. Therefore, efficient systems are needed to overcome the issues of high memory requirements. A review of such systems is presented by Al-Jarrah et al. for processing large-scale data with less memory requirements (Al-Jarrah et al. 2015). The authors were mainly interested in the analytical aspects of processing big data. However, they did not address the computation complexity which is one of the main

components of high memory requirements. Therefore, a more comprehensive study is still needed to properly address the challenges present in the area of using machine learning for big data analytics.

References

Ahmad J, Muhammad K et al (2018a) Visual features based boosted classification of weeds for real-time selective herbicide sprayer systems. Comput Ind 98:23–33

Ahmad J, Muhammad K et al (2018b) Efficient conversion of deep features to compact binary codes using fourier decomposition for multimedia big data. IEEE Trans Industr Inf 14(7):3205–3215

Al-Jarrah OY, Yoo PD, Muhaidat S, Karagiannidis GK, Taha K (2015) Efficient machine learning for big data: a review. Big Data Res 87–93

Associates DB (2011, Nov) Big data impacts data management: agreement: the 5 Vs of big data. Retrieved from http://davebeulke.com/big-data-impacts-data-management-the-five-vs-of-big-data/

BMC blog (2018) https://www.bmc.com/blogs/machine-learning-data-science-artificial-intelligence-deep-learning-and-statistics/

Data Processing & Data Processing Methods (2018) https://planningtank.com/computer-applications/data-processing-data-processing-methods

Data processing: meaning, definition, steps, types and methods (2018) https://planningtank.com/computer-applications/data-processing

Dean J, Ghemawat S (2004) MapReduce: simplified data processing on large clusters. In: Proceedings of 6th symposium on operating systems design and implementation, pp 137–149

Deng J, Dong W et al (2009) Imagenet: a large-scale hierarchical image database. In: Paper presented at IEEE conference on computer vision and pattern recognition, Miami, FL, USA, pp 20–25

Explain three methods of data processing (2018) https://www.kenyaplex.com/questions/16977-explain-three-methods-of-data-processing.aspx

Five data mining techniques that help create business value (2018) https://datafloq.com/read/data-mining-techniques-create-business-value/121

Group ES (2012, August) The 6 vs: the BI/analytics game changes so microsoft changes excel. Retrieved from http://www.esg-global.com/blogs/the-6-vs-the-bianalytics-game-changes-so-microsoft-changes-excel/

Jan B, Farman H et al (2017) Deep learning in big data analytics: a comparative study. Comput Electr Eng. https://doi.org/10.1016/j.compeleceng.2017.12.009

Khan M, Han K, Karthik S (2018a) Designing smart control systems based on internet of things and big data analytics. Wirel Pers Commun 1683–1697

Khan M, Iqbal J, Talha M, Arshad M, Diyan M, Han K (2018b) Big data processing using internet of software defined things in smart cities. Int J Parall Program 1–14

Khumoyun A, Cui Y et al (2016) Spark based distributed deep learning framework for big data applications. Paper presented at international conference on information science and communications Technologies, Tashkent, Uzbekistan, pp 2–4

Laney D (2001) 3D data management: controlling data volume, velocity, and variety. Application Delivery Strategies, META Group

Najafabadi MM, Villanustre F, Khoshgoftaar TM, Seliya N, Wald R, Muharemagic M (2015) Deep learning applications and challenges in big data analytics. J Big Data 1–21

Qiu JW, Ding G, Xu Y, Feng S (2016) A survey of machine learning for big data processing. EURASIP J Adv Signal Process 1–16

Shvachko K, Kuang H, Radia S, Chansler R (2010) The Hadoop distributed file system. In: Proceedings of the 2010 IEEE 26th symposium on mass storage systems and technologies 1–10

Silva BN, Khan M, Han K (2017) Big data analytics embedded smart city architecture for performance enhancement through real-time data processing and decision-making. Wirel Commun Mob Comput 1–12

Sukumar SR (2014) Machine learning in the big data era: are we there yet? In: Proceedings of the 20th ACM SIGKDD conference on knowledge discovery and data mining: workshop on data science for social good

Towards Data Science (2018) https://towardsdatascience.com/supervised-vs-unsupervised-learning-14f68e32ea8d

Vorhies W (2014, October) How many V's in big data? The characteristics that define big data. Data Science Central. Retrieved from http://www.datasciencecentral.com/profiles/blogs/how-many-v-s-in-big-data-the-characteristics-that-define-big-data