SN Computer Science (2021) 2:160
https://doi.org/10.1007/s42979-021-00592-x

REVIEW ARTICLE

Machine Learning: Algorithms, Real-World Applications and Research Directions

Iqbal H. Sarker[1,2]

Received: 27 January 2021 / Accepted: 12 March 2021 / Published online: 22 March 2021

Abstract

In the current age of the Fourth Industrial Revolution (4*IR* or Industry 4.0), the digital world has a wealth of data, such as Internet of Things (IoT) data, cybersecurity data, mobile data, business data, social media data, health data, etc. To intelligently analyze these data and develop the corresponding *smart and automated* applications, the knowledge of artificial intelligence (AI), particularly, *machine learning (ML)* is the key. Various types of machine learning algorithms such as supervised, unsupervised, semi-supervised, and reinforcement learning exist in the area. Besides, the *deep learning*, which is part of a broader family of machine learning methods, can intelligently analyze the data on a large scale. In this paper, we present a comprehensive view on these *machine learning algorithms* that can be applied to enhance the intelligence and the capabilities of an application. Thus, this study's key contribution is explaining the principles of different machine learning techniques and their applicability in various real-world *application* domains, such as cybersecurity systems, smart cities, healthcare, e-commerce, agriculture, and many more. We also highlight the challenges and potential *research directions* based on our study. Overall, this paper aims to serve as a reference point for both academia and industry professionals as well as for decision-makers in various real-world situations and application areas, particularly from the technical point of view.

Keywords Machine learning · Deep learning · Artificial intelligence · Data science · Data-driven decision-making · Predictive analytics · Intelligent applications

Introduction

We live in the age of data, where everything around us is connected to a data source, and everything in our lives is digitally recorded [21, 103]. For instance, the current electronic world has a wealth of various kinds of data, such as the Internet of Things (IoT) data, cybersecurity data, smart city data, business data, smartphone data, social media data, health data, COVID-19 data, and many more. The data can be structured, semi-structured, or unstructured, discussed briefly in Sect. "Types of Real-World Data and Machine Learning Techniques", which is increasing day-by-day. Extracting insights from these data can be used to build various intelligent applications in the relevant domains. For instance, to build a data-driven automated and intelligent cybersecurity system, the relevant cybersecurity data can be used [105]; to build personalized context-aware smart mobile applications, the relevant mobile data can be used [103], and so on. Thus, the data management tools and techniques having the capability of *extracting insights or useful knowledge* from the data in a timely and intelligent way is urgently needed, on which the real-world applications are based.

Artificial intelligence (AI), particularly, *machine learning (ML)* have grown rapidly in recent years in the context of data analysis and computing that typically allows the applications to function in an intelligent manner [95]. ML usually provides systems with the ability to learn and enhance from experience automatically without being specifically programmed and is generally referred to as the most popular

This article is part of the topical collection "Advances in Computational Approaches for Artificial Intelligence, Image Processing, IoT and Cloud Applications" guest edited by Bhanu Prakash K N and M. Shivakumar.

✉ Iqbal H. Sarker
msarker@swin.edu.au

[1] Swinburne University of Technology, Melbourne, VIC 3122, Australia

[2] Department of Computer Science and Engineering, Chittagong University of Engineering & Technology, 4349 Chattogram, Bangladesh

latest technologies in the fourth industrial revolution (4*IR* or Industry 4.0) [103, 105]. "Industry 4.0" [114] is typically the ongoing automation of conventional manufacturing and industrial practices, including exploratory data processing, using new smart technologies such as machine learning automation. Thus, to intelligently analyze these data and to develop the corresponding real-world applications, *machine learning algorithms* is the key. The learning algorithms can be categorized into four major types, such as supervised, unsupervised, semi-supervised, and reinforcement learning in the area [75], discussed briefly in Sect. "Types of Real-World Data and MachineLearning Techniques". The popularity of these approaches to learning is increasing day-by-day, which is shown in Fig. 1, based on data collected from Google Trends [4] over the last five years. The *x-axis* of the figure indicates the specific dates and the corresponding popularity score within the range of 0 (*minimum*) to 100 (*maximum*) has been shown in *y-axis*. According to Fig. 1, the popularity indication values for these learning types are low in 2015 and are increasing day by day. These statistics motivate us to study on *machine learning* in this paper, which can play an important role in the real-world through Industry 4.0 automation.

In general, the effectiveness and the efficiency of a *machine learning* solution depend on the nature and characteristics of *data* and the performance of the *learning algorithms*. In the area of machine learning algorithms, classification analysis, regression, data clustering, feature engineering and dimensionality reduction, association rule learning, or reinforcement learning techniques exist to effectively build data-driven systems [41, 125]. Besides, *deep learning* originated from the artificial neural network that can be used to intelligently analyze data, which is known as part of a wider family of machine learning approaches [96]. Thus, selecting a proper learning algorithm that is suitable for the target application in a particular domain is challenging. The reason is that the purpose of different learning algorithms is different, even the outcome of different learning algorithms in a similar category may vary depending on the data characteristics [106]. Thus, it is important to understand the principles of various machine learning algorithms and their applicability to apply in various real-world application areas, such as IoT systems, cybersecurity services, business and recommendation systems, smart cities, healthcare and COVID-19, context-aware systems, sustainable agriculture, and many more that are explained briefly in Sect. "Applications of Machine Learning".

Based on the importance and potentiality of "Machine Learning" to analyze the data mentioned above, in this paper, we provide a comprehensive view on various types of *machine learning algorithms* that can be applied to enhance the intelligence and the capabilities of an application. Thus, the key contribution of this study is explaining the principles and potentiality of different machine learning techniques, and their applicability in various real-world application areas mentioned earlier. The purpose of this paper is, therefore, to provide a basic guide for those *academia and industry* people who want to study, research, and develop data-driven automated and intelligent systems in the relevant areas based on machine learning techniques.

The key contributions of this paper are listed as follows:

- To define the scope of our study by taking into account the nature and characteristics of various types of real-world data and the capabilities of various learning techniques.
- To provide a comprehensive view on machine learning algorithms that can be applied to enhance the intelligence and capabilities of a data-driven application.

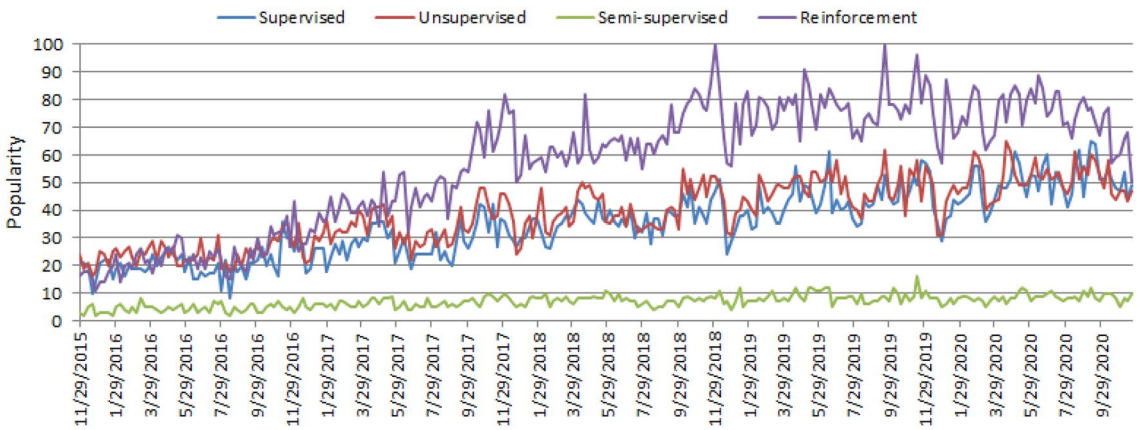

Fig. 1 The worldwide popularity score of various types of ML algorithms (supervised, unsupervised, semi-supervised, and reinforcement) in a range of 0 (min) to 100 (max) over time where x-axis represents the timestamp information and y-axis represents the corresponding score

– To discuss the applicability of machine learning-based solutions in various real-world application domains.
– To highlight and summarize the potential research directions within the scope of our study for intelligent data analysis and services.

The rest of the paper is organized as follows. The next section presents the types of data and machine learning algorithms in a broader sense and defines the scope of our study. We briefly discuss and explain different machine learning algorithms in the subsequent section followed by which various real-world application areas based on machine learning algorithms are discussed and summarized. In the penultimate section, we highlight several research issues and potential future directions, and the final section concludes this paper.

Types of Real-World Data and Machine Learning Techniques

Machine learning algorithms typically consume and process data to learn the related patterns about individuals, business processes, transactions, events, and so on. In the following, we discuss various types of real-world data as well as categories of machine learning algorithms.

Types of Real-World Data

Usually, the availability of data is considered as the key to construct a machine learning model or data-driven real-world systems [103, 105]. Data can be of various forms, such as structured, semi-structured, or unstructured [41, 72]. Besides, the "metadata" is another type that typically represents data about the data. In the following, we briefly discuss these types of data.

– *Structured:* It has a well-defined structure, conforms to a data model following a standard order, which is highly organized and easily accessed, and used by an entity or a computer program. In well-defined schemes, such as relational databases, structured data are typically stored, i.e., in a tabular format. For instance, names, dates, addresses, credit card numbers, stock information, geolocation, etc. are examples of structured data.
– *Unstructured:* On the other hand, there is no pre-defined format or organization for unstructured data, making it much more difficult to capture, process, and analyze, mostly containing text and multimedia material. For example, sensor data, emails, blog entries, wikis, and word processing documents, PDF files, audio files, videos, images, presentations, web pages, and many other types of business documents can be considered as unstructured data.
– *Semi-structured:* Semi-structured data are not stored in a relational database like the structured data mentioned above, but it does have certain organizational properties that make it easier to analyze. HTML, XML, JSON documents, NoSQL databases, etc., are some examples of semi-structured data.
– *Metadata:* It is not the normal form of data, but "data about data". The primary difference between "data" and "metadata" is that data are simply the material that can classify, measure, or even document something relative to an organization's data properties. On the other hand, metadata describes the relevant data information, giving it more significance for data users. A basic example of a document's metadata might be the author, file size, date generated by the document, keywords to define the document, etc.

In the area of machine learning and data science, researchers use various widely used datasets for different purposes. These are, for example, cybersecurity datasets such as NSL-KDD [119], UNSW-NB15 [76], ISCX'12 [1], CIC-DDoS2019 [2], Bot-IoT [59], etc., smartphone datasets such as phone call logs [84, 101], SMS Log [29], mobile application usages logs [137] [117], mobile phone notification logs [73] etc., IoT data [16, 57, 62], agriculture and e-commerce data [120, 138], health data such as heart disease [92], diabetes mellitus [83, 134], COVID-19 [43, 74], etc., and many more in various application domains. The data can be in different types discussed above, which may vary from application to application in the real world. To analyze such data in a particular problem domain, and to extract the insights or useful knowledge from the data for building the real-world intelligent applications, different types of machine learning techniques can be used according to their learning capabilities, which is discussed in the following.

Types of Machine Learning Techniques

Machine Learning algorithms are mainly divided into four categories: Supervised learning, Unsupervised learning, Semi-supervised learning, and Reinforcement learning [75], as shown in Fig. 2. In the following, we briefly discuss each type of learning technique with the scope of their applicability to solve real-world problems.

– *Supervised:* Supervised learning is typically the task of machine learning to learn a function that maps an input to an output based on sample input-output pairs [41]. It uses labeled training data and a collection of training examples to infer a function. Supervised learning is carried out when certain goals are identified to be accom-