The analytics and applications of big data and machine learning

Mingze Tang

Department of statistics, the Ohio State University, Columbus, USA

*Corresponding author: tang.1272@buckeyemail.osu.edu

Abstract. Economic and social activities in today's world are becoming increasingly digital. The Internet and the assistance of artificial intelligence technologies have led to massive amounts of data from these activities. These data come from different sources and in various forms, both structured and unstructured. Some data have large sample sizes, while there are also high-dimensional, enormous data in which the dimensionality of the explanatory variables surpasses the sample size. These enormous data sets are valuable, and they could drive a variety of economic activities. Big data production, machine learning, and statistics are deeply interrelated. This article discusses the concepts and methods of big data and machine learning from the characteristics of big data and the nature of machine learning. In particular, the four characteristics of big data are deeply analyzed and discussed. In addition, the analysis discusses the relationship between machine learning and big data. The article concludes with a summary and outlook of the whole article.

Keywords: Artificial neural networks; machine learning; big data; mathematical statistics; data analytics.

1. Introduction

Nowadays, statistics and Big Data analysis are very closely connected. As a methodological science, statistical data analysis provides reliable methods and tools for research in social and natural sciences. But because of the continuous development of the Internet and smart devices, Big Data greatly impacts statistical methods [1-4]. Compared with general data, Big Data is huge in volume, wide in source and variety; moreover, much of Big Data is generated in real-time. And the uninterrupted data generation rush forces the birth of new statistical methods, the Machine learning is one of them. So, what are big data and machine learning? What is the connection between them? What is the main difference between machine learning and statistical modeling as an essential tool for big data analysis? What are the methods of machine learning?

This paper attempts to answer these questions. In Section 2, this paper will discuss the primary sources and main characteristics of Big Data, especially economic Big Data. In Section 3, the nature of machine learning and several important machine learning methods will be discussed. Section 4 is the conclusion.

2. The characteristics of big data

The exponential growth of Internet devices and sensors is the main reason for generating and collecting massive amounts of Big Data. These growths have led to many big data sources, including smartphones, social media, search data, sensors and images, and traffic data. In financial markets, various offline online commodity trading platforms, and electronic payment systems. Helpful information is also collected by various corporate and government websites, especially by large technology companies such as Google, Amazon, and Apple. These companies collect user experience data to optimize their products. So big data as a new factor of production can, in turn, further boost the economy. Based on the existence and forms of activities of Big Data mentioned, it can conclude that Big Data has the following four significant characteristics.

1) Volume: The volume of Big Data is so large that the storage security of information collected from various sources is critical. Such a large volume of data has two implications: on the one hand,

the sample size of many big data can be in the millions. Suppose the sample size of big data is larger than the dimensionality of the explanatory or predictive variables. In that case, such big data is called "Tall Big data." The large sample size means that much new information can be obtained from big data, especially unstructured data, which can improve the statistical inference of DGP (data generating process). Generally speaking, only a tiny fraction of large data is used for feasibility statistical analysis due to the limitation of computer capacity and computational speed [2]. On the other hand, the enormous nature of big data does not inevitably imply a huge sample size. It may also refer to many descriptions of DGPs from different dimensions in a given time. Big data possesses a high-dimensional set of potential explanatory or predictor variables. This provides excellent possibilities and flexibility for exploring critical explanatory variables. When the dimensionality of the potential explanatory or predictive variables exceeds the sample size, it poses a considerable challenge for statistical modeling and statistical inference, which is called "dimensional disaster" in statistics, and big data with this characteristic is called "Fat Big data." Many explanatory variables may not affect the dependent variable for a high-dimensional set of explanatory variables. There may be multicollinearity among many explanatory variables. Therefore, developing various feasible methods for variable selection, essentially a dimension reduction or data reduction, is necessary.

2) Variety: Big data comes in various forms, and structured digital data is shared. Text documents, emails, images, videos, and audio are unstructured data. Unstructured data provides wealthy new information that is not available in traditional data and will dominate. The diversity of Big Data leads to the diversity of its sources [5]. It may have different storage addresses and methods, making integrating and cleaning various data between different systems difficult. However, because Big Data provides more valuable information than traditional data, it can be used to develop more efficient statistical inference methods and tools. For example, there is an increasing interest in social media data. This information is usually unstructured or semi-structured data, which are generated daily and are good training data. On the other hand, combining unstructured data with traditional structured data can better infer the essential features of DGP.

3) Velocity. Big data is generated and spread quickly, so it is essential to have a matching collection and analysis speed. Many sensors and intelligent charging systems enable the need to process large amounts of data in near real-time. However, more realistically, in many cases, big data is not generated at a uniform rate but rather at a cyclical rate over time. For instance, there is a distinct cyclical pattern in stock market trading, with larger volumes during the market's opening and closing and lower volumes around lunchtime [6]. Alternatively, perhaps the comment data on a website forum also has a cyclical pattern. Daily cyclical peak data based on event triggers are complicated to manage. In order to make timely data analysis and prediction possible, it is necessary to build high-frequency, real-time data collection speed. For example, in economic statistics, high-frequency macroeconomic variables can be constructed to allow a timely understanding of macroeconomic trends and improve economic policy interventions' timeliness. The current practice of economic statistics can only obtain monthly time series data such as consumer price index and producer price index. However, based on Internet information and extensive data analysis, the government is fully capable of constructing daily data of CPI and PPI, and even the sampling frequency can be higher. In addition, the availability of high-frequency data in time series analysis can avoid the missing information caused by temporal aggregation. If the model parameters change slowly over time, it may need higher-frequency observations to infer the parameter values at arbitrary time points.

4) Veracity: The authenticity of big data means much noise in big data, including false information and distorted data. Therefore, it is necessary to perform data cleaning and processing. Effectively extracting meaningful and accurate information from the data is the essence of statistical analysis. So classical statistical methods help process big data, in any case. However, due to the large volume, high dimensionality, and low information density of big data, it is also necessary to develop new tools. The computer algorithm-based adequacy principle is helpful in data imputation.

3. The essence of Machine learning

Machine learning has emerged as the essential data science analytical technique in the age of big data. Machine learning has proliferated due to big data and cloud computing, but machine learning cannot replace statistical analysis [7]. For instance, while statistics can be crucial in inferential analysis, dimensional approximation, causal identification, and result interpretation, machine learning can enhance out-of-sample prediction and pattern recognition, such as facial recognition. Therefore, machine learning and statistics are complementary, and their cross-fertilization can lead to the development of new techniques and tools for data science and statistical science (Figure 1).

Machine learning is a branch of artificial intelligence. Using algorithms and statistical models, machine learning enables computer systems to automatically "learn" data, identify patterns, and make predictions or judgments without explicit human programming. Learning machines can be categorized as reinforced, supervised, and unsupervised.

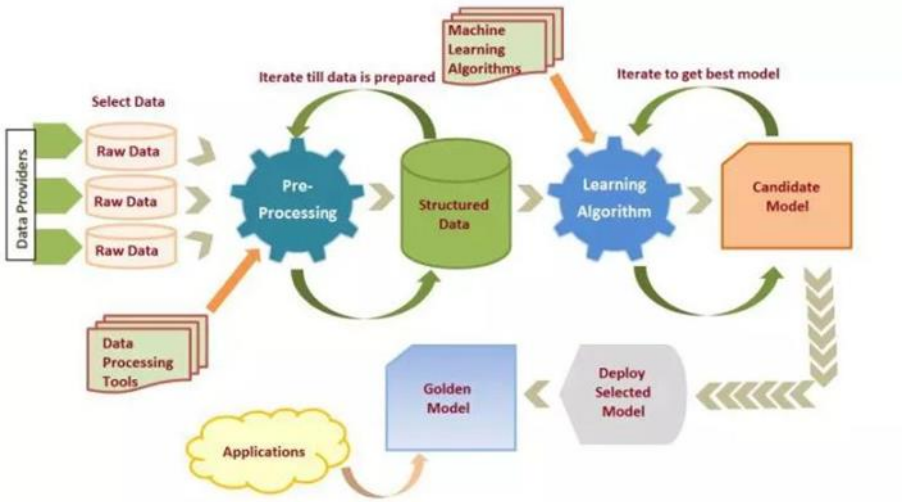

Fig. 1 Machine Learning Process

Reinforcement learning is the study of how algorithms execute tasks in dynamic contexts in order to maximize the cumulative reward. Due to the universality of reinforcement learning, several fields have investigated this topic, including operations research, cybernetics, simulation optimization, and multi-intelligent systems. The dynamic environment is typically described as a Markov decision process in machine learning. Many algorithms for reinforcement learning employ dynamic planning strategies. Reinforcement learning algorithms can be utilized in autonomous driving and human-versus-human gaming events.

The development of algorithms is reliant on training data in supervised learning. The training data consists of a collection of training samples, each with one or more outputs and inputs referred to as supervised signals. The supervised learning approach explores a function that may be used to anticipate the output corresponding to fresh test data by optimizing the goal function iteratively. By optimizing the goal function, the algorithm can properly forecast the output by direct mapping to the new input. The algorithms for supervised learning include classification and regression. Classification algorithms are utilized when the output is limited to a finite set of possible values. Whenever the output can take on any value within a given range, regression procedures are utilized.

Unsupervised learning seeks structure in input-only training data, such as the grouping or clustering of data points. Instead of reacting to feedback, unsupervised learning algorithms discover standard aspects of the training data and make decisions based on the presence or absence of standard features in each new test data. Unsupervised learning is mostly used to estimate the statistical probability density function, but it may also be utilized in various contexts requiring the summary and interpretation of data characteristics. Cluster analysis is a crucial tool for unsupervised learning. It categorizes data items according to the information present in the data that characterizes the objects and their connections. The goal is for things within a group to be similar (correlated), but objects in

separate groups would be dissimilar (uncorrelated). The better the clustering, the larger the within-group similarity and the between-group disparity. Different clustering methods result in different calculated distances between data points and, accordingly, different closeness of Cluster, which depends on the best algorithm. Figure 2 shows the difference between supervised and unsupervised learning

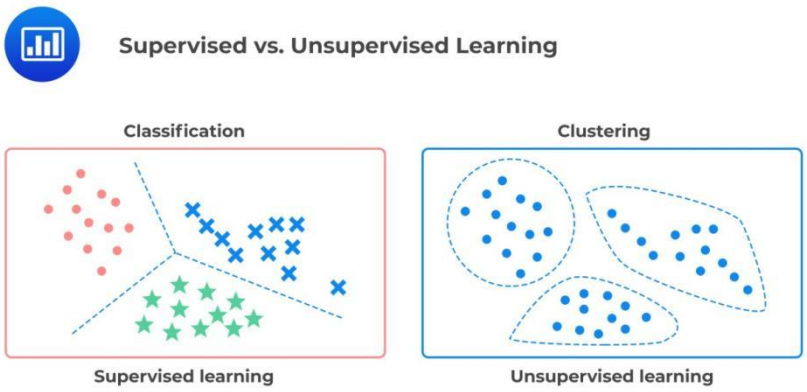

Fig. 2 The difference between supervised and unsupervised learning

Machine learning is fundamentally a mathematical and computational optimization issue. Mathematical optimization, which offers the field's theory, techniques, and applications, is closely connected to machine learning. At the same time, machine learning is closely related to computational statistics, often overlapping and focusing on using fast and efficient computer algorithms for prediction. In machine learning, many learning problems can be formulated as minimizing some predetermined loss function. In order to avoid the phenomenon of overfitting, the ultimate goal usually translates into the problem of minimizing the prediction error based on anonymous data. Specifically, ML (machine learning) is based on training data, learning and mining the systematic features and statistical relationships between variables of the training data in order to predict new unknown data. In order to obtain an accurate prediction algorithm, the available data are generally divided into two subsets: training\testing data. The training data are used to learn and mine the systematic features of the data and the statistical relationships between variables, and then use these systematic features and statistical relationships to predict the behavior of the unknown data. To ensure accurate prediction, overfitting of the training data must be avoided [8]. The phenomenon of "overfitting" refers to mining features and statistical relationships that exist only in the training data but not in the anonymous data, which can improve the in-sample fit of the training data but do not contribute to out-of-sample prediction. Therefore, the prediction effect's evaluation needs to be based on another part of the data, i.e., the test data. In addition, to avoid overfitting further, a penalty term is usually introduced to give a corresponding penalty to the algorithm's complexity, i.e., the higher the algorithm's complexity, the heavier the penalty. To achieve optimal out-of-sample prediction, machine learning requires the identification of an optimization strategy from the training data that minimizes the loss function and penalty term for predicting the test data.

Standard methods include decision trees, the k-nearest neighbor, SVM, random forests, artificial neural networks, and deep learning.

Decision tree. Decision trees as a method of prediction embody the complete process of prediction, from the observed values of some feature variables to the desired values of the target variables. Decision tree learning is a statistical and machine learning prediction approach. Suppose the variable of interest has discrete values. In this scenario, the decision tree is referred to as a classification tree, where the leaves represent class labels and the branches represent the functional conjunctions that produce these class labels. If the target variable has continuous values, the decision tree is known as a regression tree. A decision tree gives a physical representation of choice and the decision-making process in decision analysis. In data mining, a decision tree reflects the data, while the outputs of a classification tree may be used as input for decision-making.

Random forest. Breiman proposed the random forest method. Based on the original observation data, a series of new random data is generated by repeated sampling, and a decision tree is grown for each data. The predicted values of the ensuing series of fast trees are then averaged; this type of prediction is known as random forest.

Support vector machine (SVM). This is a supervised classification and regression learning approach. The SVM training method predicts which of two classes a fresh sample belongs to, given a series of training examples labeled as belonging to one of two classes. Support vector machine training is a probabilistic binary linear classifier. SVM may conduct effective nonlinear classification in addition to linear classification by implicitly mapping its input to a high-dimensional feature space.

The k-nearest neighbor approach chooses the k observations of a character variable that are closest to a specified value. The dependent variable is then predicted by averaging the observations of the dependent variable closest to the intended value. This technique is known as the k-nearest neighbor technique [8].

The artificial neural network is the computer algorithm system that "learns" how to accomplish tasks by analyzing training data samples (Figure 3). Artificial neural networks consist of nodes called "neurons" that are linked and resemble the system of neurons in a human brain. Like synapses in a normal brain, each connection between artificial neurons may transfer "signals." The signal-receiving artificial neuron may analyze the information and then transmit it to other artificial neurons linked with it. Artificial neurons often have a learning-adjusted weight that can increase or decrease the signal strength in a connection's transmission [9]. The threshold value of the artificial neuron may be delivered only when the aggregated weighted signal surpasses that value, such that each artificial neuron's output is governed by an activation function that combines the input weights. Artificial neurons are often aggregated into one or more hidden layers. Different hidden layers are capable of performing distinct changes on their respective inputs. After traversing many layers, signals may be transported from the original input layer to the final output layer. The initial objective of the artificial neural network strategy was to solve issues in the same or a comparable manner to the human brain. Over time, however, the focus turned to accomplishing certain tasks, diverging from biology. Computer vision, voice recognition, social network filtering, and chess games are all uses of artificial neural networks.

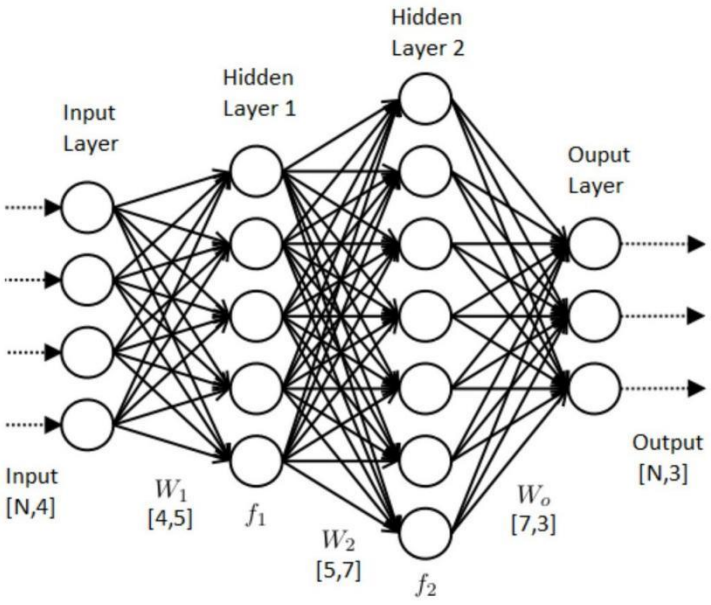

Fig. 3 The Artificial Neural Networks

Deep learning. If an artificial neural network has numerous hidden layers, the process is referred to as deep learning. Deep learning seeks to replicate how the human brain converts light and sound into sight and hearing. Computer vision and voice recognition are two effective uses of deep learning [10].

4. Conclusion

This paper discusses the concepts and methods of Big Data and Machine Learning. Big data has four characteristics: Volume, Velocity, Variety, and Veracity, which lead to the need for more efficient methods to clean and process the data. Furthermore, this article discusses the nature of machine learning, which is a big data analysis method that has emerged widely along with the emergence of big data and cloud computing. It predicts unknown samples by learning the systematic features and statistical relationships of training data, which is consistent with the idea of statistics to infer the total from sampling. In addition, the relationship between machine learning and big data is illustrated. Machine learning shares the same stochastic probability basis as mathematical statistics. However, it does not assume that the structure or probability distribution of DGP satisfies a specific function or model form but learns the systematic features of data and statistical relationships between variables from training data by computer algorithms to achieve out-of-sample classification and prediction. Machine learning algorithms are often known for their accurate out-of-sample predictions, but they are often like black boxes that are difficult or even impossible to interpret. However, many important machine learning methods, such as decision trees, random forests, k-nearest neighbor methods, artificial neural networks, and deep learning, share or are very similar to the basic idea of nonparametric analysis.

References

[1] Bok B, Caratelli D, Giannone D, et al. Macroeconomic nowcasting and forecasting with big data. Staff Reports, 2014, 11(16): 51-59.

[2] Breiman L. Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). Statistical Science, 2001, 16(3):199-215.

[3] Biau G. Analysis of a Random Forests Model. Journal of Machine Learning Research, 2010, 13(2):1063-1095.

[4] Groenvik H. A self-normalizing approach to the specification test of mixed frequency models. Michigan Technological University. 2016.

[5] Khan A, Gul M A, Uddin M I, et al. Corrigendum to "Summarizing Online Movie Reviews: A Machine Learning Approach to Big Data Analytics". Scientific Programming, 2021 12(6):1-10.

[6] Fan M, Gu S, Jin Y, et al. Big data-based Grey Forecast Mathematical Model to Evaluate the Effect of Escherichia Coli Infection on Patients with Lupus Nephritis. Results in Physics, 2021, 26(1):104339.

[7] Wang Z, Zhao X, Han Z, et al. Advanced big-data/machine-learning techniques for optimization and performance enhancement of the heat pipe technology -A review and prospective study. Applied Energy, 2021, 294(15):116969.

[8] Lasheras F S. Predicting the Future-Big Data and Machine Learning. Energies, 2021, 14:129-136.

[9] Balasubramanian K, Donmez P, Lebanon G. Unsupervised Supervised Learning II: Training Margin Based Classifiers without Labels. Eprint Arxiv, 2010, 12(6):3119-3145.

[10] Dang A T, Tsujimura M, Ha N T, et al. Evaluating the predictive power of different machine learning algorithms for groundwater salinity prediction of multi-layer coastal aquifers in the Mekong Delta, Vietnam. Ecological Indicators, 2021, 127:107790.