Machine Learning Research in Big Data Environment

Shi Jiang

Jinan University, Guangzhou, China

jiangshi@jnu.edu.cn

Keywords: big data; machine learning; environment; algorithm

Abstract: With the explosive growth of data in the industry, the concept of big data has received more and more attention. Due to the massive, complex and diverse nature of big data, the characteristics of rapid changes, how to effectively use information in big data, and use this information to increase productivity has become an urgent problem to be solved. Machine learning is one of the effective ways to solve such problems. Therefore, the topics of common concern of the machine learning academic community and industry under the big data environment are studied. This paper aims to study some basic methods and problems faced by machine learning.

1. Introduction

Big data and machine learning are two hot areas in the rapid growth of the information industry. From the past information obstruction to the current data explosion, the data volume and data scale growth rate in all fields have increased at an alarming rate. According to the National Security Agency's statistics, the Internet processes 1826 PB bytes per day [1]. As of 2011, digital information has increased by a factor of nine in the past five years, and by 2020 this figure will reach 35 trillion gigabytes [2]. The scale of this digital data brings tremendous opportunities and potential for change. It can take advantage of the completeness of these data to help us make better decisions in all walks of life, and it provides a great opportunity for data-driven research in scientific research. Good example. This allows us to use search engine big data prediction capabilities in the field of medicine, astronomy, and other fields when conducting scientific research on data.

It is generally believed that traditional machine learning is a shallow learning structure. In contrast, deep learning refers to machine learning techniques that automatically learn deeper structures under supervision or non-supervision and use them for classification or data mining. Inspired by the human brain's signal processing model in nature, the concept of deep learning has been proposed and has received increasing attention due to its superior processing performance in many fields. Many companies now make full use of the advantages of big data to widely apply them to commercial products and have achieved great success [1]. These companies and organizations collect massive amounts of information and analyze them on a daily basis based on a large amount of data, and then use the results of the analysis for deep learning related projects. For example, a virtual personal assistant for iPhones, offers a wide range of services such as weather forecasting, sports news, answering user questions, and reminding services [1]. Google will apply massive chaotic data to deep learning algorithms. These bits and pieces of data come from Google translate, Android's speech recognition, Google Street View, and search engines. Other industry giants are not far behind.

Compared with traditional machine learning, machine learning under big data greatly expands the number of samples, and many types of problems are supported by rich samples. This is an advantage of big data, but it also causes many problems. Now with the constant optimization of hardware technology and programming algorithms, the collection and magnitude of data are no longer the main problems that hinder the research of big data [2]. The relationship between data, that is, data which data is useful, which is redundant, and even interferes with other data, how these data are sometimes used to work is the major challenge facing big data. Big data has enormous potential value in all aspects of our society. Obtaining valuable information from big data is not a simple task. It is a core goal of big data technology to dig out the laws hidden in the data and the information we need from

the data with huge volumes and various structures so that the data can maximize its value.

2. The Concept of Big Data and Machine Learning

2.1 The concept of big data

There is currently no unified regulation for big data. Usually, the understanding of big data is that data cannot be loaded into the computer's internal memory. This is an informal definition because each computer has a size that cannot be loaded into memory. The industry extends the features of big data from the very first 3V models to the current 4V models, which mainly include: volume, variety of data types, low data value density, and a lot of real-time data requirements [2]. In response to these characteristics, knowledge analysis in the era of big data, coordination of machine intelligence and human intelligence, and intelligent analysis systems will play an important role. People need an intelligent analysis interface to connect humans with the computer world. Otherwise, they will lose the torrential data flow.

The issue of big data is a challenging issue that is currently a common concern of the academic community and industry. With the maturity of related technologies for the collection, transportation, processing, and application of big data, non-traditional tools can be used to process a large amount of structured, semi-structured, and unstructured data to obtain a series of analysis and prediction results.

2.2 The concept of machine learning

Since the computer was invented, people wanted to know if it could learn. Machine learning is essentially a multidisciplinary field. It draws on achievements in artificial intelligence, probability statistics, computational complexity theory, cybernetics, information theory, philosophy, physiology, and neurobiology.

The main purpose of machine learning research is to use computer to simulate human learning activities. It is a method to study computer recognition of existing knowledge, acquisition of new knowledge, continuous improvement of performance and self-improvement. Learning here means learning from data [3]. It includes Supervised Learning, Unsupervised Learning, and Semi-Supervised Learning. Supervised learning requires training of known samples to obtain the algorithm model, and then predicts the measurement results (or labels) of unknown samples; unsupervised learning is the direct prediction of the results of unknown samples, and the training process is not realized. And semi-supervised learning is a machine learning method that falls somewhere between the two.

A new challenge facing traditional machine learning is how to handle big data. At present, machine learning problems that involve large-scale data are ubiquitous [3]. However, because many existing machine learning algorithms are based on memory, big data cannot be loaded into computer memory. Therefore, many existing algorithms cannot handle big data. How to propose new machine learning algorithms to meet the needs of big data processing is one of the research hotspots in the era of big data.

3. Mechanical Learning in Big Data Environment

The combination of machine learning and big data has generated tremendous value. Based on the development of machine learning technology, data can be "predicted." For humans, the richer the accumulated experience, the extensive experience, the more accurate the judgment of the future. For example, people who often say "experienced people" have more work advantages than young men who "get out of the limelight", because they are more accurate than others [3]. In the field of machine learning, based on a well-known experiment, it effectively validates a theory in the machine learning world: The more data the machine learning model has, the better the prediction efficiency of machine learning.

3.1 Machine learning classification

Supervised learning learns a function from a given set of training data. When new data arrives the

result can be predicted based on this function. The training set requirements for supervised learning include input and output as well as features and goals. The goal of the training set is marked by people [3]. Common supervised learning algorithms include regression analysis and statistical classification. The difference between supervised learning and unsupervised learning is whether the training set target is marked. They all have training sets and both input and output unsupervised learning compared to supervised learning. The training set has no artificially labeled results. Common unsupervised learning algorithms have clustering. Semi-supervised learning is between supervised learning and unsupervised learning. Enhanced learning learns how to make actions through observation. Every action will have an impact on the environment. Learning objects make judgments based on the feedback of the observed surrounding environment.

3.2 The scope of machine learning

Machine learning has a deep connection with pattern recognition, statistical learning, data mining, computer vision, speech recognition, and natural language processing.

In terms of scope, machine learning is similar to pattern recognition, statistical learning, and data mining. At the same time, the combination of machine learning and processing techniques in other fields has formed interdisciplinary subjects such as computer vision, speech recognition, and natural language processing [3]. Therefore, generally speaking, data mining can be equivalent to machine learning. At the same time, what we usually call machine learning applications should be universal, not only limited to structured data, but also applications such as images and audio.

3.3 Machine learning methods

Machine learning methods include regression algorithm, neural network, SCM (Support Vector Machine), clustering algorithm, dimension reduction algorithm, recommendation algorithm, gradient descent method, Newton's method, BP algorithm, SMO algorithm, etc. The algorithm is not explained here in detail [4]. Interested in Relevant data or attention to the micro signal data, there will be an irregular algorithm to explain.

In addition to these algorithms, there are some algorithmic names that often appear in the field of machine learning. But they are not themselves a machine learning algorithm, but are born to solve a sub-problem. You can understand them as the sub-algorithms of the above algorithms, which are used to greatly improve the training process. The representatives include: gradient descent method, which mainly uses online regression, logistic regression, neural network, and recommendation algorithm; Newton method, which mainly uses online regression, BP algorithm, which is mainly used in neural networks, SMO algorithm, which is mainly used SVM.

3.4 Mechine learning in big data environment

Under the big data environment, machine learning has two main research directions in the development process. The first is to study the learning mechanism and focus on the exploration of the human learning mechanism; the second is to research and use information effectively. Put in valuable knowledge that can be learned from large databases. The study of the learning mechanism mainly comes from machine learning techniques. Under the current big data environment, analyzing the data has become the focus of attention in different industry fields [4]. Machine learning can absorb knowledge from it and enable machine learning to effectively promote the development of machine technology. In the current big data environment, how to use effective learning means is the significance of machine learning, and machine learning will become a well-respected and popular learning and service technology. Based on machine learning data analysis, how to quickly and efficiently process a large amount of data information is the key research direction of machine learning [5].

Under the current big data environment, the number and types of data have greatly changed and improved, and the speed of data generation is also increasing. In addition, the innovation of data types also makes the analysis more difficult, such as text sentiment analysis, image search and understanding, image data analysis and so on. In this way, the research directions and learning methods of machine learning have been further extended, presenting a variety of characteristics [4].

For example, the rational use of semi-supervised learning methods to improve the quality of training data and the transfer of learning in different knowledge backgrounds are all current priorities.

In addition to the above, in order to further promote the efficiency of machine learning, we should also solve a series of problems that can be extended, that is, to solve the problem of big data. At this time, parallel methods should be adopted [5]. From these aspects big data analysis: visualization analysis, data mining algorithms, predictive analysis capabilities, semantic engines, and data quality and management.

Nowadays, printed materials produced by human activities have reached a data volume of about 200 PB (1 PB = 210 TB). In the long history, the amount of information spoken by humans has reached 5 EB (1 EB = 210 PB). With the development of science and technology, the personal computer's data storage capacity has long reached the TB level, and the data volume of some large-scale enterprises is as high as EB level. Therefore, it is not difficult to see that the era in which we live is an era of big data, and the huge amount of data surrounding our lives.

4. The Trend of Machine Learning in Big Data Environment

In the course of research, many experts agreed that in the coming decades, there will be the following challenges in the field of machine learning algorithms, and it is precisely the trend of its development.

4.1 Improve the generalization ability of machine learning

This is a trend in the development of machine learning. It is also a very common problem. Many industries are eager to further enhance the generalization of machine learning. From the current point of view, support vector machines have the most wanted technology with generalization ability. They combine theory and practice very well, and are a kind of comprehensive learning method. Their origins originate from practice to theory [6].

4.2 Improve the speed of machine learning

In terms of machine learning in different fields, how to effectively improve the speed of machine learning is a focus that we are highly concerned about and is also the goal of continuous football [6]. At present, people are more concerned with how to deal with the relationship between machine learning speed test and speed training, eliminating the conflict between the two. For example, the K nearest neighbor algorithm usually has a slower test speed, but its method of training the speed is very fast.

4.3 Improve the intelligibility of machine learning

There are also many areas where attention is paid to improving the comprehensibility of machine learning. For example, in the clinical field of medical treatment, patients want to be able to understand the reasons for adopting such a treatment plan [5]. At present, machine learning is more powerful in this area, such as ensemble learning, neural networks, and support vector machines.

4.4 Improve data use capability

In the past, machine learning methods mainly focused on marked data. However, with the development of network technology and the gradual improvement of data analysis and collection technologies, many fields have encountered machine learning pressures caused by unmarked data, such as spam and Medical image data, etc. [3]. In addition, there are many areas that suffer from inconsistency, missing attributes, and large amounts of noise, such as garbage data. This part of the unbalanced data is often the normal use of image data, such as medical diagnosis and treatment of breast cancer, the presence of patient samples the number is far greater than the number of healthy samples, which leads to new problems [6]. That is how to make full use of unlabeled data information to correctly handle the impact of spam data and unbalanced data, so as to improve the data use capacity.

4.5 Improve the ability to deal with the issue of sensitivity costs

In the current big data environment, the emphasis of machine learning algorithms is on how to reduce the error rate. However, various industries and disciplines have different tolerances for errors, even if they are treated in the same discipline or in the same industry [6]. The cost of differential judgment is also very different. For example, in the field of medical cancer diagnosis and treatment, patients are misdiagnosed as being healthy and healthy people are misdiagnosed as suffering from cancer, and the cost is different. In the same way, the machine's judgment on the burglary's burglary, misbehavior is misjudged as the return of the owner and the misappropriation of the owner's home misappropriation as a thief. The price paid by both is also very large.

In the past, machine learning algorithms were basically based on the consideration of the same cost. In the future development process, we should focus on improving the ability to deal with this sensitive cost problem. In recent years, there have been many experts in related fields to medical diagnosis. Analytical and signal-related theories were introduced into machine learning algorithms. I believe there will be significant progress in this area of research in the context of big data.

5. Summary

In the era of big data, there are many advantages that make machine learning more applicable. For example, with the development of the Internet of Things and mobile devices, we have more and more data, and the types also include unstructured data such as pictures, texts, videos, etc. This makes the machine learning model obtain more and more data. At the same time, distributed computing Map-Reduce in big data technology make the speed of machine learning faster and easier to use. All kinds of advantages make the advantages of machine learning get the best performance in the era of big data.

References

[1] L.T. Huang, Big data and machine learning, Information Systems Engineering, 2012 vol.3, pp.30-34.

[2] W.W. Hu, Machine learning under big data, Information Systems Engineering, 2011, vol.5, pp.66-69.

[3] Q.T. He and N.B. Li, Machine learning in big data, Pattern recognition and artificial intelligence, 2014, vol.4, pp. 32-34.

[4] Y.M. Zheng, Trends in machine learning algorithms in big data environment, Pattern Recognition and Artificial Intelligence, 2004, vol.6, pp.54-56.

[5] W.W. Wang, Machine learning in big data environment, Information Systems Engineering, 2016, vol.7, pp.131-133.

[6] X.X. Wang, Research on machine learning algorithm trend in big data environment, Journal of Natural Science of Harbin Normal University, 2013, vol.4, pp.48-50.