Software Design of Machine Learning Model Based on Big Data Analysis

Feng Zou\textsuperscript{1,}\textsuperscript{*}, Lizhu Ye\textsuperscript{1}, Wenhai He\textsuperscript{1}, Yunxiang Liu\textsuperscript{1}

\textsuperscript{1}Guangzhou College of Commerce, Guangzhou, China, 511363

*Corresponding author e-mail: zouyudg@gcc.edu.cn

Abstract. With the advent of the information age, big data analysis technology has been popularized in various fields. Using big data analysis technology has also effectively promoted the development of various fields. This article mainly analyzes the trend of machine learning algorithms in the big data environment, it also designed and studied machine learning models.

Keywords: Big Data, Machine Learning, Design Method

1. Introduction

Traditional data analysis is mainly to explore data and filter data information through artificial methods. With the advent of the information age, a large amount of complex data is generated every day. Obviously, the traditional data analysis method can no longer complete data analysis, so it must be done with the help of machine learning models, which is based on big data analysis technology [1].

2. Overview of the connotation of big data

Big data needs a new processing model to have good development capabilities and process optimization capabilities, this can enrich information assets. Big data has good data acquisition and storage capabilities, and it can build a data processor with management and analysis capabilities. From a technical point of view, big data cannot carry out information processing on a single computer. It needs to organize and analyze data information through a distributed structure, so as to ensure that it has good distributed data collection capabilities. The characteristics of big data mainly include processing a variety of data and improving the speed of information data acquisition. Big data can quickly process a large number of large-capacity information. It can not only perform market calculations, but also analyze the risk elements of enterprise transformation and development, so it has been highly valued by people [2].

3. Trend analysis of machine learning algorithms in the big data environment

3.1. Improve the generalization ability and comprehensibility of machine learning

In the following decades of development, machine learning algorithms still have great challenges, and it is also an important trend in technological development. In the process of machine learning, it is very important to comprehensively improve the generalization ability of machine learning. At present,
many fields of society expect to comprehensively improve the generalization ability of machine learning. From the current actual development status, support vector machine generalization technology can integrate theoretical and practical results, which is a comprehensive learning method [3].

3.2. Improve machine learning speed and data application capabilities
Machine learning is integrated into the development process of many fields of society. We must fully optimize the speed of machine learning, which is also the basic goal pursued by people in the long-term development. At present, people pay more attention to the rational adjustment of the basic relationship between machine learning speed and speed training, and eliminate the basic contradiction between the two. For example, the K-nearest neighbor algorithm test runs slower, but the method used in the actual training process is faster. Traditional machine learning methods mainly learn from labeled data. With the development of network technology, data analysis and collection technology has been gradually improved. Due to the unmarked data in some fields, the pressure on machine learning has gradually increased, such as clearing all kinds of spam and image data. In addition, the main areas will be disturbed by junk data such as attribute changes and large amount of noise information. Such imbalanced data will have a greater impact on the rational application of image data. Therefore, we must fully utilize the application value of unlabeled data information, and then reasonably handle the imbalance between data information and spam, so as to improve Basic application efficiency of data [4].

3.3. Improve the ability to deal with sensitive cost issues
In the era of big data, the focus of machine learning algorithms is focused on controlling the error rate. However, the tolerance for error rates of different industries in the development process is quite different, and the price paid by the same learning and the same industry for differentiated judgments is also quite different. For example, in the medical field, different patients' disease diagnosis results have different impacts. In the criminal case judgment process, according to different judgments of criminals' behavior, the costs that need to be paid are quite different. In the application of traditional machine learning algorithms, most of them are comprehensively considered on the basis of the same cost. Therefore, in the long-term development process in the future, we must focus on dealing with sensitive cost issues. In recent years, many technical researchers have begun to introduce signal-related theories and medical diagnosis learning methods in the application of machine learning algorithms. We must strive to further improve the machine algorithms in the development of the big data environment.

4. Machine learning model design
Machine learning is part of the research area of artificial intelligence. Its purpose is to allow computers to learn autonomously like humans, so that they can speed up the processing of data. In 1997, Mitchell TM, Dean and Professor of the School of Machine Learning at Carnegie Mellon University, believed that the machine learning process is the process by which computers improve their own performance based on the experience they have learned [5,6]. The ultimate goal of machine learning is to obtain knowledge from data. The design of a machine learning model generally consists of four parts: environment, learning element, knowledge base and execution element, as shown in Figure 1.
The main purpose of big data is to discover valuable information in the data, and machine learning is an important means among them. This method is an important solution for big data analysis. Big data makes machine learning algorithms more accurate. At the same time, machine learning algorithms require more and more memory computing speeds. Therefore, big data and machine learning restrict each other, promote each other, and rely on each other. This article deals with big data by studying the design methods of machine learning models.

4.1. Support Vector Machine

In 1995, Corinna Cores and Vapnik created Support Vector Machine (SVM). It is a new machine learning model. This model has attracted widespread attention. Supporting the questioning machine learning model is based on the VC dimension theory and the principle of structural risk minimization, with sample information as the base point, to find the best balance between the complexity of the model and the learning ability, in order to achieve the best results. Support The learning model of vector machine is often used in the fields of small sample, non-linear, high-dimensional pattern recognition, face detection, machine translation and so on.

SVM is the most widely used machine learning model. For example, for linearly separable problems, SVM is to find the hyperplane with the largest interval to separate two different samples, and the hyperplane with the largest interval has the best generalization ability. as shown in picture 2 [7].

Now use formula (1) to explain the definition of hyperplane, $x$ is the weight vector, and $y$ is the optimal hyperplane offset.
The distance from a sample to the optimal hyperplane is: $r = \frac{g(a)}{||x||}$.

$g(a) = WTa + y = 0$ is the discriminant function determined by the hyperplane. SVM can maximize the distance between $y$ and $y$. Then by solving the dual problem, the values of $x$ and $y$ are obtained, and then the kernel function is introduced into the non-linear separable problem. Linear inseparability is a normal phenomenon, it exists in many problems, because the objective function of the dual problem is uncertain, so it is impossible to achieve the optimization. To solve this problem, there are two methods: soft interval optimization, that is, relax the restrictions on the input space, and you can choose to ignore some errors. But when some extremely linear inseparable problems appear and too many classification errors cannot be solved, this method is not applicable. The kernel technique is to find a kernel function and transform the data in the low-dimensional space into the high-dimensional space, so that the data becomes separable and thus can be solved. This method also cannot guarantee to solve all linearly inseparable problems, so for more complex linearly inseparable problems, the two methods should be used in combination [8].

4.2. Artificial Neural Network

Artificial Neural Network (ANN) is a machine learning method that simulates the operation of the brain, referred to as neural network. It is a machine learning method similar to mathematical statistics. According to Figure 3, humans have established an artificial neural network model:

![Figure 3. Schematic diagram of neurons](image)

This model has three main advantages: First, it can learn independently. For example, to exercise the function of artificial neural network for face recognition, we input thousands of facial images and corresponding person information into this model, and the neural network will gradually learn this skill. This skill plays an important role in forecasting and can help people make disaster predictions, risk predictions, etc. Secondly, it has Lenovo storage function. This kind of function needs to be realized through the feedback system of artificial neural network. Finally, it can quickly find the best solution to the problem. But if you want to find the best solution, you have to try all the solutions, which requires a lot of calculations, and the use of neural network machine learning models can simplify this complex problem, so as to quickly find the best method.

Artificial neural network has the advantages that support vector machines do not possess. It has stronger processing capacity for nonlinear problems and better adaptability. It makes up for the shortcomings of traditional machine learning, and it has a wider range of applications. The combination of neural networks and other machine learning models is more effective in processing big data information, which makes artificial intelligence a big step forward, and it also promotes the development of information processing technology. With the continuous development of informatization, artificial neural network has determined a new development direction, making its
operation mode more humane. The combination of information geometry and artificial neural network has opened up a new way for the theoretical research of artificial neural network [9]. People have more in-depth research on artificial neural network machine learning models, so the scope of application of this model will become wider and wider, but there is still a lot of room for improvement. Among them, the combination of neural networks and other technologies and the resulting hybrid methods and hybrid systems have become the main research object. But because they all have their own advantages and disadvantages, to combine neural networks with other technologies, we must use their strengths and avoid weaknesses. In this way, a better application effect can be obtained, which is also one of the biggest difficulties.

5. Classification of machine learning models
There are many machine learning models under big data, and support vector machines and neural networks are the two most important models. According to the algorithm, the model can be divided into three categories. The first type, supervised learning, is that the computer extracts relevant information from big data, and then the big data verifies the information and provides results. The fundamental purpose of this model is to allow computer science to learn experience through this process, and then to solve similar problems. Neural networks and support vector machines are both supervised learning; the second type, unsupervised learning, refers to the computer's autonomous interception of useful information in big data. The goal of this kind of learning is uncertain; the third type of reinforcement learning, it refers to the computer autonomously assessing information without big data verification [10].

6. Conclusion
Although the current machine learning methods can already meet the current needs of enterprises, with the development of the times and technology, we need to conduct more in-depth research on machine learning. Only in this way can we cope with the increasing amount of data information.

References
[1] Li Zhijie, Li Yuanxiang, Wang Feng. Overview of online learning algorithms for big data analysis [J]. Journal of Computer Research and Development, 2015(8):1707-1721.
[2] WANG Wei. Machine Learning Algorithm in Big Data Environment [J]. Information Systems Engineering, 2016(7):133.
[3] XIAO Hong. Discussion on Machine Learning Algorithm Based on Big Data [J]. Communications World, 2017(6):265-266.
[4] Chen Wei. Research and Implementation of Parallel Machine Learning Based on MapReduce [J]. Industry & Technology Forum 2017,16(9):69-70.
[5] Alicia wang. Machine learning model design method based on big data analysis [J]. Computer Knowledge and Technology, 2018:203-204.
[6] LIU Xingjian. Trend Analysis of Machine Learning Based on Big Data [J]. Information and Computer, 2019:121-122.
[7] WANG Shuo. Discussion on Machine Learning Algorithm Based on Big Data Analysis [J]. Information & Computer (Theoretical Edition), 208:64-65.
[8] Wang Yuqiong. Research on Adaptive Learning Analysis Model Based on Big Data [J]. Journal of Jiamusi Vocational College, 2020:139-140.
[9] Zhi-hui liu;Zhi-gang huang;Xie and bright;Is big data risk control effective?-- Comparative analysis based on statistical scoring card and machine learning model [J]. Statistics & Information Forum, 2019:19-27.
[10] Peng Hua, Li Xiang, Liao Hongcun. Intelligent investment "wisdom" -- model construction based on big data and machine learning [J]. China Financial Computer, 2019:30-34.