Pisco Med Publishing

Original Research Article

Design and Implementation of Machine Learning Algorithm Support Platform Based on Big Data

Ning Yang

Weifang Special Steel Group Co,.Ltd Weifang City, Shandong 261000

Abstract: Machine learning is an interdisciplinary subject in many fields. Its initial research motivation is to give computing machine system learning ability. The early machine learning data source was single and relatively simple to process, but the business scenarios that could be handled were limited. However, with the development of big data technology, machine learning has made major breakthroughs in language, sound, picture and text. More and more applications hope to optimize their experience by adding machine learning algorithms. However, the development process of machine learning applications is still relatively complex, which often requires equipment developers and algorithm developers to work together. Algorithm developers cannot directly optimize the computing efficiency of the platform, and developers cannot understand the specific workflow of the algorithm, It makes the development of machine learning applications more difficult.

Keywords: Big data; Machine learning algorithm; Support platform; Design; Realization

1. Introduction

Big data and machine learning are major technological changes in the field of modern computer, which have had a great impact on all walks of life. At present, with the rapid development of the Internet, mobile communications, social networks and the Internet of things, these networks will produce a large amount of data every day. Data has become the most important information resource today. Some studies have shown that in many cases, the larger the data scale, the better the effect of machine learning using these data. Therefore, machine learning supported by big data has become a hot research field highly concerned by global academia and industry. This paper introduces some classical machine learning algorithms supported by big data.

2. Overview of big data

In recent years, with the rapid development of Internet, mobile communication, social media and Internet of things, various network applications will produce a large amount of data every day, resulting in the explosive growth of the total amount of global data. Data has become the most important basic information resource today, and human society has accelerated the pace of entering informatization. With the explosive growth of data in the industry and the unprecedented accumulation of data, the concept of big data has attracted more and more attention. Big data is bringing huge profits to data intensive enterprises.Big data includes massive structured, semi-structured or unstructured data generated by the Internet, medical devices, video surveillance, mobile devices, intelligent devices, non-traditional it devices and other channels.

What valuable information can human beings obtain in the face of so much data has become the focus of human society. In 2012, the U.S. government announced that big data will become an important technology development field in the United States in the future, following the highway and the Internet. Now many national and international multinationals have also joined the development of big data, such as Google, IBM, Microsoft, Alibaba and Baidu. The basic definition of big data can be summarized from its various characteristics[1]. The basic model of big data is summarized by the characteristics of big data. The basic definition of big data includes data volume, diversity, velocity, variability, virtual and value. In view of these characteristics, Wang Feiyue believes that in the era of big data, knowledge analysis, coordinated work between machine intelligence and human intelligence and intelligent analysis system will play an important role[2]. People need an intelligent analysis interface to connect human beings with the computer world, otherwise they will be submerged in the flood of big data. With the passage of time, big data technology will be applied to all fields of human society, bring huge technological changes and unprecedented development opportunities in all fields. Big data technology includes data generation, data storage, data analysis, data processing, etc[3]. Data analysis is the core technology of big data technology, which can directly generate valuable information. At present, data analysis technologies include data mining, classification and clustering, association rules, genetic algorithm, regression analysis, neural network, machine learning and so on.

3. Classical machine learning algorithm supported by big data

3.1 Bayesian machine learning

Bayesian method was gradually established after the 1950s. It is the most important part of probability theory and mathematical

doi: 10.18282/l-e.v10i4.2604

statistics. Bayesian analysis is the basis of Bayesian learning method. It provides a method to calculate the hypothesis probability based on the probability of different data observed under a given hypothesis and the observed data itself. Bayesian learning method is to synthesize the prior information about unknown parameters with sample information, obtain the posterior information according to Bayesian formula, and then infer the unknown parameters according to the posterior information[4]. Bayesian model needs less estimated parameters. When the attribute correlation is less, the algorithm of the model is simple, the classification error rate is small, and the overall performance is good. The disadvantage of Bayesian method is that in practice, the probability distribution of category population and the probability distribution of various samples are often unknown. In order to obtain a more accurate overall probability distribution and the probability distribution of various samples, the more we know about the population, the better, and the greater the sample requirements, the better. Bayesian machine learning predicts the frequency of the event in the future by calculating the frequency of the event in the past. The prediction result completely depends on the collected data. The more data collected, the better the prediction result[5]. As the main technical means of generating, storing and processing more and more massive data, big data can provide enough good sample data for Bayesian machine learning. Big data technology and Bayesian machine learning method have achieved good results in some research and application.

3.2 Artificial neural network

Artificial neural network (ANN) is a mathematical model formed by many hidden nodes connected by weights. It has the characteristics of large-scale parallel processing, distributed information storage, good self-organizing learning ability and so on. Back propagation algorithm (BP) is a supervised learning algorithm in artificial neural network. Artificial neural network can approach any function in theory, and its basic structure depends on the hidden nodes in the network, so it has strong nonlinear mapping ability[6]. The number of intermediate layers, the number of nodes in each layer and the initial weight of each node in the artificial neural network can be set according to the specific situation, which is very flexible. Artificial neural network has good fitting effect on training data, and has good application results in many fields such as medicine, physiology, philosophy, informatics, computer science and so on. Although the artificial neural network has achieved good application results in some fields, the artificial neural network supported by big data is still in its infancy, and there are still many problems to be solved. For example, how to determine the number of layers and nodes of artificial neural network, and how to improve the training speed of the network, especially in the environment of massive data, the data presents high-dimensional attributes and the diversity of data types. Big data technology is just the key technology to solve these problems. It can bring more surprising learning effects to artificial neural networks through distributed and parallel computing of big data.

4. Conclusion

Big data has the characteristics of sparse attributes, ultra-high dimensions and complex relationships. Traditional machine learning algorithms are powerless in the face of such data scale. Therefore, this paper mainly discusses the theoretical research of several classical machine learning algorithms in big data environment (1) How to select learning samples and attribute characteristics of samples in the big data environment; (2) How to use the distributed computing and parallel computing of big data to provide the execution efficiency and speed of machine learning algorithm. In short, the machine learning algorithm supported by big data has broad research and application prospects. The two complement each other and will certainly push big data machine learning to a higher level.

References:

[1] Chen Di, Zheng Lianxiang, sun composition. Research on the construction and application of rehabilitation big data platform [J]. China health standard management, 2021,12 (11): 1-5.

[2] Wang F Y. A Big-Data Perspective on AI : Newton, Merton, and Analytics Intelligence. IEEE Intelligence Systems, 2012, 27 (5) : 2-4

[3] Chen Di, Zheng Lianxiang, sun composition. Research on the construction and application of rehabilitation big data platform [J]. China health standard management, 2021,12 (11): 1-5.

[4] Song Hongqing, Du Shuyi, Zhou Yuanchun, et al. Big data intelligent platform and application analysis of oil and gas resources development [J]. Journal of Engineering Science, 2021,43 (2): 179-192.

[5] Ye Lishan, Zhao Fei, Chen Jian, et al. Research and practice of big data application based on intelligent electronic health record platform [J]. Chinese Journal of health information management, 2019,16 (6): 672-676.

[6] Shen Wenyuan, Wu Yezheng, Wei Heng, et al. Application and design of deep learning in data analysis of environmental pollution platform [J]. Electronic production, 2021 (2): 46-47,63.

[7] Xu Chengjie, Xiao Xirong, Zhang Jingyi, et al. Design and application of spark based big data analysis platform [J]. Chinese Journal of health information management, 2019,16 (5): 633-637.