Advances in Intelligent Systems Research, volume 147

International Conference on Network, Communication, Computer Engineering (NCCE 2018)

The Application of Machine Learning in Data Mining under Big Data Environment

Weini Chen

Hubei Engineering Institute; Huangshi Hubei 435003 China.

Abstract. With the development of economic globalization, the rapid development of industries in various fields, big data technology has attracted more and more attention. Network data is constantly being generated at an unprecedented rate, and it is necessary to intelligently process huge data, and then to make full use of the value in the data, you need to use machine learning methods. This paper describes the related theory of machine learning in detail. Based on data mining, it discusses the flow chart of neural network training algorithm in detail and studies the application and prospect of machine learning in big data.

Key words: big data; machine learning; data mining; application.

INTRODUCTION

In the era of big data [1], people generally appreciate the value of big data [2], and big data analysis its intelligent application has become a hotspot of research and application in recent years [3], in these hot spots, machine learning and data analysis [4] is the key point of research on data analysis. Traditional machine learning and data analysis algorithm [5] cannot be directly used for the analysis of the data processing, in addition, still need to solve the technology problems such as large-scale distributed storage and parallel computing support and so on. Therefore, big data machine learning is a comprehensive research topic, because it involves the design of large-scale system and machine learning algorithm. The basic processing flow of big data consists of three parts: data analysis, data extraction and integration and data interpretation. The value of big data lies in the improvement of decision making after analyzing the data, so data mining is of great significance in the process of big data processing.

MACHINE LEARNING

In the big data environment, machine learning requires computers to quickly acquire valuable information from a wide variety of data. Traditional machine learning focuses on using pre-set statistical methods to discover the value of data in data analysis. But the goal of machine learning in large data environment is to search out specific rules that hidden behind dynamic, changeful, multi-source heterogeneous data and finally maximize the data value. It is necessary to combine big data technology with machine learning algorithm to find some relevant links from the complex and changeable data, and then excavate the research value of data through the computer.

In this era of big data, the new challenge faced by traditional machine learning is how to deal with massive amounts of data, while the problems of traditional machine learning mainly include the following aspects:

(1) study the language differences between computer systems and users.

(2) understand and simulate human learning.

(3) realize the inference requirement of incomplete information.

There are two important research directions in the development of machine learning: one is the study of human learning mechanism, focusing on simulation and even realizing human learning behavior; On the other hand, it studies how to obtain valuable and effective knowledge from huge data. This has a higher requirement for machine learning algorithm, and it even requires that the corresponding algorithm must have the ability to process massive data and

Copyright © 2018, the Authors. Published by Atlantis Press.
This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

high dimensional data. It requires that the training model be less complex and less time-consuming to calculate. The requirement of the large-scale data processing is widespread, but because the existing learning algorithms cannot meet the requirements, and even exist defects, the existing algorithms cannot deal with important data very well. Therefore, new machine learning algorithm is one of the research directions of machine learning. The goal of machine learning algorithm is to use the training method to make computer intelligent based on historical data set. The machine learning system consists of four parts, as shown in figure 1, including learning elements, environment, execution and knowledge base.

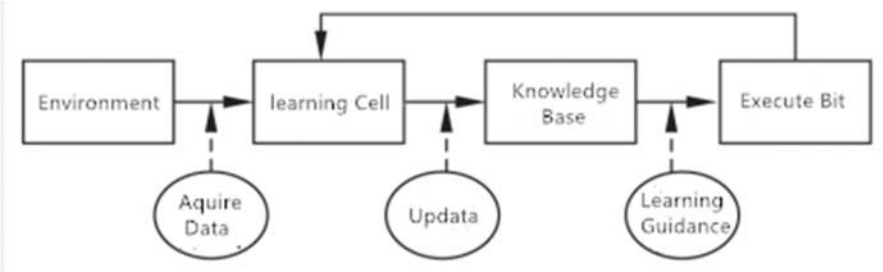

FIG. 1. structure diagram of machine learning system.

DATA MINING

Data Mining is also known as knowledge discovery. Data Mining is generally composed of four kinds of methods, such as classification method, clustering method, association rule and regression method. The association rule is mainly to find the relation between variables; In this case, clustering usually identifies a similar set of processes; The regression process is to try to use the least error rate function to fit the data model easily.

When choosing a training data set for learning, first we need to determine a classification model, which can then be automatically divided into many categories and finally completed the classification. There are many algorithms for machine learning classification, such as decision tree, naive Bayesian classification algorithm, support vector SVM and artificial neural network. The typical classification is shown in figure 2.

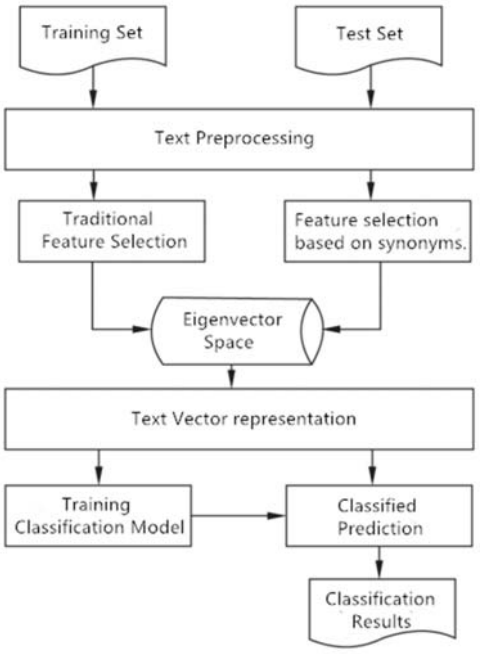

FIG. 2. classification task flow chart in machine learning.

BP NEURAL NETWORK

BP Neural Network (BP) has the following characteristics: It mainly selects distributed information storage mode, which is convenient for large-scale parallel processing. It has strong self-adaptability, good robustness and fault tolerance. BP neural network, because of its characteristic advantages, is favored by many industries in pattern recognition, signal processing, process control, function approximation, image processing, market analysis, industry estimates, the application of fault detection and other fields has a very significant.

BP neural network generally has three layers of network topology, such as hidden layer, input layer and output layer. The layers are interconnected, and the nodes on each layer are disconnected. BP's design includes forward signal transmission and the error back propagation. The basic principle of the two stages of learning process is to use some implicit layer transformation of the input vector to calculate output vector, thus it is concluded that the mapping relationship between input data and output data. The BP network information loop is composed of the forward propagation of input information and the reverse propagation of the output error. Of course, to control the output error in a certain predetermined range, this algorithm needs to iterate iteratively, which can change the connection weight coefficient between each layer of neurons. The flow chart of neural network training is shown in figure 3.

APPLICATION AND PROSPECT.

Big data, as a new hotspot industry, needs to be equipped with a set of relatively scientific reasonable machine learning algorithms which can classify data effectively, decrease the difficulty of data processing analysis to further improve the ability of machine learning, to constantly adapt to the needs of society, Big data technology has been integrated into many fields, such as telecommunication, medical, financial, and many other industries, and has been widely applied. But with the progress of the society, if we want to make a breakthrough in the big data, we must optimize the traditional machine learning algorithms, make it have strong vitality in the era of big data, it will need more in-depth studies of machine learning to deal with huge data information and get the useful information in large data. In this way, we can promote the development of the civilization of human society.

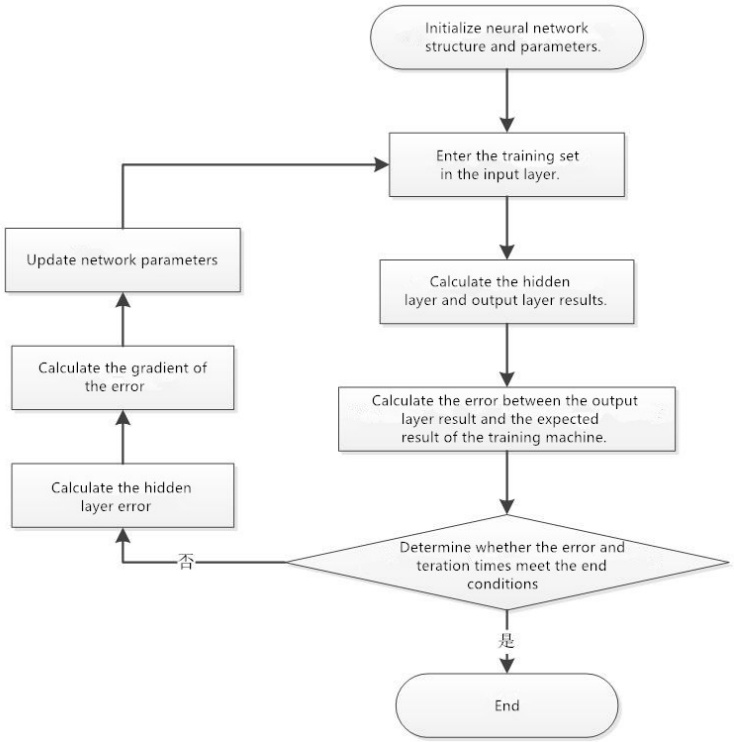

FIG. 3. flow chart of neural network training.

REFERENCES

1. Sun Cunyi, Gong Liutang. Research on interest rate pricing under big data thinking -- an empirical analysis based on machine learning [J]. Financial theory and practice, 2011,67(7):1-5.
2. Xu Qianyi, qi fang. Research on non-structured big data analysis algorithms based on machine learning [J]. Laser journal,2016,37(10):125-128.
3. Mohammed Ali al-gardai, Kasturi Dewi Varathan, Sri Devi Ravana. Cybercrime detection in online communications: The experimental case of cyberbullying detection in The Twitter network[J]. Computers in Human Behavior,2016,63.
4. He Qing, Li Ning, Luo Wenjuan, Shi Zhongzhi. A review of machine learning algorithms under big data [J]. Pattern recognition and artificial intelligence,2014,27(04):327-336.
5. Zhang Shaocheng, Sun Shiguang, Qu Yang, Dong Yu. Application of machine learning in data mining in big data environment [J]. Journal of Liaoning university (natural science edition), 2014,44(01):15-17.