Neural Computing and Applications (2020) 32:5399–5407
https://doi.org/10.1007/s00521-019-04682-z

S.I. : ATCI 2019

Unstructured big data analysis algorithm and simulation of Internet of Things based on machine learning

Rui Hou[1] · YanQiang Kong[2] · Bing Cai[3] · Huan Liu[4]

Received: 15 September 2019 / Accepted: 10 December 2019 / Published online: 2 January 2020

Abstract

Big data values data processing to ensure effective value-added data. With the rapid development of the cloud era, the coverage of big data has gradually expanded, and it has received wide attention from all walks of life. In the process of modern social development, big data analysis is gradually applied to the future development planning, risk evaluation and integration of market development status. With the rapid development of many fields of society, the flow of information has gradually expanded, and the Internet has developed more rapidly, prompting the application of big data in various fields. Machine learning is a multidisciplinary study of how computers use data or past experience. With the ability to independently improve specific algorithms, the computer acquires knowledge through learning and achieves the goal of artificial intelligence. Big data and machine learning are the major technological changes in the modern computer world, and these technologies have had a huge 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 generate large amounts of data every day, and data become the most important information resource of today. Some studies have shown that in many cases, the larger the amount of data, the better the data will be for machine learning. On this basis, this paper proposes an online client algorithm based on machine learning algorithm for IoT unstructured big data analysis and uses it in other big data analysis scenarios. Use the online data entered by the customer to implement background data mining, the parallel way to verify its efficiency through machine learning algorithms such as K-nearest neighbor algorithm.

Keywords Machine learning · Internet of Things · Big data analysis · Unstructured data

1 Introduction

Since the concept of big data was proposed by Toffler in 1980, its development prospects have been expanding and have penetrated into all aspects of life, work and research [1, 2]. Nowadays, with the promotion of information technology, some scattered data are collected and gradually concentrated into large and complex data. The rapid development of big data has brought huge benefits to the advanced technology industry and has changed the habits of users to a certain extent, attracting the attention of many companies with economic strength. In 2017, IT companies

✉ YanQiang Kong
k@ncepu.edu.cn

Rui Hou
90102332@ncepu.edu.cn

Bing Cai
cbycnx@126.com

Huan Liu
418297201@qq.com

[1] School of Economics and Management, North China Electric Power University, Beijing 102206, People's Republic of China

[2] School of Energy Power and Mechanical Engineering, North China Electric Power University, Beijing 102206, People's Republic of China

[3] State Grid Ningxia Electric Power Co., Ltd, Yinchuan 750001, Ningxia, People's Republic of China

[4] State-Owned Assets Supervision and Administration Commission, Yinchuan 750000, Ningxia, People's Republic of China

such as Alibaba, Tencent, and Jingdong invested heavily in big data research to enjoy the financial returns from big data. For example, the business efficiency of drip taxis, shared bicycles, Taobao, etc., is promoted through the promotion of big data [3]. The strategic importance of big data technology is not just about mastering a large amount of data, but rather focusing on the added value of these important data [4, 5]. In other words, if big data are compared with the industry, then the key to profitability in this industry is to perform data processing and enhance data computing capabilities to achieve data appreciation. From a technical point of view, the connection between big data and cloud computing is as inseparable as the front and back of a coin [1, 6]. Big data cannot be processed by one computer and must use a distributed architecture [7, 8]. It provides distributed data mining capabilities for large amounts of data. However, it must rely on cloud computing for distributed processing, distributed data storage, and cloud computing and virtualization technologies. With the advent of the cloud era, big data has attracted more and more attention. The analyst team believes that big data is often used to describe the large amount of unstructured data and semistructured data created by companies. When downloading a relational database for analysis, these data take too much time and money [9, 10]. Big data analytics is often associated with cloud computing. Because real-time large dataset analysis requires a framework like MapReduce to distribute work to dozens, hundreds or even thousands of computers [11, 12]. Therefore, the study of machine learning algorithms in the context of big data plays an important role in promoting the development of the country, enterprises and society. Machine learning plays an indispensable role in today's big data processing. For example, in the AlphaGo game for Kejie in 2017, the game ended with a score of 3–0. This is an important symbol of machine learning. Machine learning overcomes the limitations of human factors, effectively processes data through neural networks, decision trees, and deep learning science, and improves data operations.

The concept of "Internet of Things" began in 1990 and originally originated on the Internet [13, 14]. However, in the later development, it gradually developed into two types: One is an Internet-based extension, and the other is to extend users between projects for information transfer, exchange, and communication. Mainly refers to the use of sensors, two-dimensional code and other technologies to achieve information and access products, and the use of Internet of Things and communication networks for transmission and storage [15, 16]. With the rapid development of animal network technology, the background of big data can provide greater information data resources for the Internet of Things [17, 18]. On the other hand, the development of Internet of Things technology will also promote the rapid arrival of the era of big data. It can be seen that the relationship between the Internet of Things and big data has always been complementary and inseparable. Only in this way can China quickly enter a smart society [19, 20]. Big data includes structured, semistructured, and unstructured data, and unstructured data are increasingly becoming a major part of data. According to IDC's survey report, 80% of the data in the enterprise are unstructured data, with an exponential growth of 60% per year. Big data analysis through machine learning algorithms opens up a new era of big data analysis [21, 22]. In the context of technological innovation represented by cloud computing, no myths or awes are needed. These seemingly difficult to collect and use data are beginning to be easily developed. Big data will gradually create more value for human beings due to continuous innovation in all areas of life.

This research is based on the current new situation of social development and plays an important role in promoting the better development of society. Through a comprehensive analysis of the research background of big data and the status quo of machine learning research, it is found that the effective use of research results in the field of machine learning can better solve the big data problem. In order to improve the value density of massive unstructured data and remove redundant and noisy garbage data, this paper takes unstructured data as a sample and uses related machine learning algorithms to perform preprocessing, dimensionality reduction processing and predictive model training. And in the traditional database for data analysis efficiency comparison, and achieved good results.

2 Proposed method

2.1 Machine learning

Machine learning is a hot research area in current computer science and artificial intelligence disciplines. The industry does not uniformly define the standard for "machine learning," but machine learning is generally a model of human cognitive processes and learning processes that combines the computational power of computers to perform human behavior simulations and to get new knowledge or skill algorithms. It uses prior knowledge and training data to guide learning and continually adopts existing knowledge structures to improve their performance. In recent years, many machine learning algorithms have been widely used in engineering practice and scientific research such as data clustering , support vector machine (SVM), nonlinear regression, neural networks, genetic algorithm, and so on. Whether it is speech recognition, credit monitoring, risk prediction, etc., or data mining of big datasets, machine learning algorithms play an irreplaceable practical guiding role. Machine learning plays a big role in the research of big data. For example, Google's success in text processing is due to machine learning, and when building big data storage warehouses, a lot of knowledge in the fields of neural networks, supervision and unsupervised learning is required to use Hadoop clusters. At the same time, Amazon's product recommendation system is also a combination of big data and machine learning. Deep analytics for big data analysis is also based on statistical analysis and machine learning.

The development of machine learning mainly includes two research directions: first, studying the learning mechanism. The main research focus of the learning mechanism is the study of machine learning techniques. With the development and changes in the big data environment, data analysis has high application requirements in the development of many fields of society. Through machine learning, it can quickly acquire corresponding knowledge and promote the development of machine technology. In the big data development environment, machine learning should highlight the important role of learning, gradually expand the actual scope of machine learning, and carry out data analysis on the basis of machine learning, efficiently process different pieces of data information, and clarify the basic goals of machine learning. The second research direction is studying the rational application of information. The focus is on finding more valuable information from a vastly populated data management repository. In the big data development environment, the data generation efficiency has gradually increased, and the overall number and types of data have undergone major changes. In addition to in-depth analysis of various types of important new rows of data, such as text data analysis, content searching images and image data processing, so that the machine learning research toward the diversification of comprehensive development. At present, the rational selection of semisupervised learning methods to strengthen the quality of training data and enhance learning ability is a key issue of concern to relevant departments. Big data is fundamental to artificial intelligence, and turning big data into knowledge or productivity is inextricably linked to machine learning. We can say that machine learning is the core of artificial intelligence and the fundamental way to ensure that machines have human intelligence. The task of machine learning is to discover information that is contained and useful based on large data volumes. The more data it processes, the more machine learning can show its advantages. This problem can be solved by providing big data or greatly improving performance, such as language recognition, image design and weather forecasting. K-nearest neighbor learning methods according to certain rules will be similar to the data sample is divided into a category, which is similar to real life idiom, "things gathered together, people were divided into several groups." In the machine learning algorithms, the basic idea of the K-nearest neighbor learning method is to first extract the characteristics of the new data to be classified or tested and compare it with the characteristics of each datum in the original sample. Then, select the K closest sample data from the comparison results and calculate which K sample data appear in the number of times. Then what kind of data is to be classified, c class w_1, w_2,…, w_c pattern recognition problem, each type has a sample of category N_i (i \\ u003d 1, 2,…, c). The discriminant function that can specify w_i is:

$$g_i(\mathbf{x}) = \min\left\|\mathbf{x} - \mathbf{x}_i^k\right\| \quad k = 1, 2, \ldots, N_i. \tag{1}$$

For unknown samples x, simply compare the Mahalanobis distance between the x and N samples of the known category:

$$d = \sqrt{(\mathbf{x}_\mathrm{u} - \mathbf{m})^\mathrm{T}\mathbf{C}^{-1}(\mathbf{x}_\mathrm{u} - \mathbf{m})} \tag{2}$$

where m and C are the mean and covariance matrix of S, respectively. It is determined that x is the same as the sample closest to it. The algorithm has the following advantages: it is simple and easy to understand; there is no need for modeling and training; and it is easy to implement, suitable for classification of rare events, and suitable for multiclassification problems. However, the algorithm also has shortcomings. The algorithm is a lazy algorithm with large memory overhead. When the test sample is classified, the calculation amount is large and the performance is low. The interpretability is poor, and the decision tree and other rules cannot be given. The support vector machine algorithm is one of the classic machine learning algorithms and has achieved good results in both theoretical analysis and practical applications. A straight line is used to divide the data into two categories. This line is used as a linear discriminant function and is recorded as:

$$g(x) = \omega^\mathrm{T} x + b. \tag{3}$$

This line is equivalent to a hyperplane, and the optimal classification hyperplane equation is:

$$\omega^\mathrm{T} x + b = 0. \tag{4}$$

The sample is spatially transformed by nonlinear mapping, and the sample data are transformed from the low-dimensional sample space to the linear dimension of the feature dimension, and then the linear classification purpose is obtained. After mapping, the classification function can be expressed as:

$$f(x) = \sum_{i=1}^{n} \omega_i \varphi(x_i) + b. \tag{5}$$

Reference perceptron idea, classification function to obtain a sample represented by the product of the form:

$$f(x) = \sum_{i=1}^{n} \alpha_i y_i < \varphi(x_i), \quad \varphi(x) > +b \tag{6}$$

In addition to the SVM algorithm, the classic K-means clustering algorithm for machine learning is a partition-based clustering algorithm that can also be used for data analysis. The algorithm calculates the distance between each object and the defined center point and optimizes the coordinates of the center point according to the algorithm strategy to obtain the best clustering result:

$$J_c = \sum_{i=1}^{k} \sum_{p \in C_i} \|p - M_i\|^2. \tag{7}$$

Artificial neural network is also a classic machine learning algorithm. The neural network has a good fitting effect on training data. It has applications in many fields such as medicine, physiology, philosophy, informatics and computer science. While artificial neural network made a good effect in some areas, artificial neural networks in support of big data are still in their early stages, and there are still many issues to be resolved. For example, how to determine the number of layers of artificial neural networks, the number of nodes, how to improve the training speed of the network, especially in the massive data environment, data presentation of various high-dimensional attributes and data big data technology is only a key technology to solve these problems. Machine learning mainly includes the following steps: (1) selecting the type of training experience to provide direct or indirect feedback for system decision making. For example, learning maze problem, every step, whether the current position can walk in a certain direction provides the most direct feedback for the system, but the ultimate destination of walking provides indirect feedback for the system, which makes walking not deviate from the correct direction. In addition, the extent to which training sample sequences are controlled, including negative learning, active inquiry learning and completely autonomous learning, all of which completely imitate human learning styles. The last problem is how close the training samples are to the true distribution of the samples, which is of great significance to the final evaluation of learning results. If the training samples are too different from the actual distribution, they may still perform poorly in the test, although they perform very well in the training samples. (2) Choosing objective function to improve learning performance is to learn a specific objective function. In some cases, the optimal objective function is not operable and can only take the second place. For any classification, the optimal objective function is the smallest error rate, but it is not easy for simple Gauss density distribution or just solving the error rate. In a word, the process of learning is the process of searching in the hypothesis space for the knowledge and constraints that most conform to the existing training examples and some prior knowledge.

2.2 Internet of Things

The Internet of Things is a new technology application model for quickly acquiring remote information through modern wireless communication technologies. The Internet of Things (IOT) refers to “Internet through all connections”; that is, information is obtained by loading information on an information sensing device such as radio-frequency identification. A network that intelligently identifies ubiquitous information can be obtained and transmitted through ubiquitous Internet connections. The technology application model is based on ubiquitous information-gathering devices, such as radio-frequency identification (RFID) tags, sensors, drives and mobile phones, through a unique solution, which is the mutual object between the objects that achieve the common goal: communication and cooperation. In 2005, the International Telecommunication Union (ITU) released the 2005 ITU Internet Report: the Internet of Things, which shows that the ubiquitous “Internet of Things” communication era is coming. The report’s description of the Internet of Things is: At present, information and communication technology has been connected to anyone from any time and place and gradually evolved to the stage of connecting anything. Various information sensing technologies, such as the Internet of Things, connect real-time information on all projects, including materials/spare parts/work-in process/finished products in the supply chain, to the Internet to achieve intelligent management and identification. The Internet of Things consists of a three-tier architecture. The second is the transport layer, which enables the transmission and sharing of information, i.e., through existing local area networks, wide-area networks, the Internet and communication networks and with the electronic product code (EPC), electronic data interchange (EDI) and other data analysis and exchange technology to achieve data transmission; the third is the application layer, which implements the processing and application of the acquired sensor data information, including applications and display terminals. Applications on mobile phones, computers and other mobile devices’ operating systems are installed and applied based on business logic. Key technologies related to the Internet of Things include radio-frequency identification (RFID), sensor technology, nanotechnology, intelligent embedded technology, network communication technology, etc.

In the development of modern science and technology, the use of cloud computing, networking technology and information resource sharing and other high technology to effectively improve the level of intelligence and utility management level city and make the life of urban residents better. This trend of infiltration into urban government, society, economy and foundation is the trend of smart cities. The best example of using big data and IoT technologies in the healthcare field is to identify patients by RFID technology, which is used for matching, patient positioning, vital sign collection and monitoring management. Specifically, it guides patients to wear an electronic watch when they are admitted to the hospital in order to keep abreast of the patient's identity information. Within the coverage of the frequency identification detection network, doctors can better use frequency identification technology to identify, organize, track and record the patient's identity anytime, anywhere. The Internet of Things and big data have an inseparable relationship from the beginning. (1) The Internet of Things is a new Internet model developed based on Internet technology, enriching the content of big data. (2) The Internet of Things generated big data at the beginning of development, and big data promoted the improvement of the Internet of Things. (3) The mobile intelligent terminal is a multifunctional IoT stage, which is the main application mode of the Internet of Things in the big data environment. (4) The Internet of Things can bring the greatest functionality and value to smart cities and is the primary condition for building smart cities.

The emergence and development of the Internet of Things brings not only the rapid development of social productivity, but also another great innovation to the mode of production, lifestyle and thinking of human society: (1) reforming the mode of human production. The Internet of Things (IOT) is an integrated innovation of human wisdom, such as technology, sensing technology, information technology, intelligent computing technology and wireless communication technology. It is also a space of interconnection between the physical world and the network world. It will greatly promote the integration of industrialization and informationization and promote the adjustment of economic structure and social and economic development, thus promoting the transformation of production mode. (2) changing the way of human life. At present, the Internet of Things covers the fields of smart industry, smart agriculture, smart logistics, smart transportation, smart grid, smart environmental protection, smart security, smart medicine and smart home. The Internet of Things will bring people unprecedented convenience and comfort and will completely change the way of human life. (3) Changing people's way of thinking. As an important part of the new generation of information technology, Internet of Things technology represents the new carrier of information dissemination and the new connotation of scientific and technological innovation. The change of this new tool, new technology and new method has gradually changed and influenced people's life trip and daily behavior from the superficial point of view. Considering from the deep level, the intellectualized society brings not only the change of lifestyle, but also the change of thinking mode.

2.3 Big data analysis

The theoretical basis of big data analysis technology is a large amount of sample data, that is, data with accurate sources, rich data and intrinsic connections. Big data analysis theory mainly includes two analysis strategies: cluster analysis and association analysis, and predictive analysis methods are based on this. At present, big data processing technologies mainly include distributed computing technology, memory computing technology and stream processing technology. The fields in which these three technologies are applicable are different. In-memory computing technologies are developed to address issues such as efficient data reading and online real-time processing. Streaming technology solves real-time, continuous, uncontrolled data streams. Distributed computing technology can be used to break down problems into many small tasks that are assigned to multiple computer processes. Open-source Hadoop has become the mainstream distributed computing technology, of which distributed file system (HDFS) and parallel distributed programming framework (MapReduce) are two core technologies. It has good scalability, efficient equipment utilization and high reliability. Distributed computing technology is applicable to distributed data sources in power enterprise collections. In-memory computing technology puts large-scale data into memory for query and analysis operations. The memory computing technology avoids a lot of time overhead when reading and writing disks, which greatly increases the calculation speed. As an emerging engine of in-memory computing technology, Spark's main advantage is cluster-based distributed memory abstraction (RDD). Spark reads the required data into memory. As the name implies, stream processing techniques treat continuous datasets as data streams and return processing results as soon as the data appear. The results are calculated, analyzed and presented in the latest data as soon as possible. Storm is a representative technology for streaming media technology, which is mainly used for real-time computing, online machine learning and other aspects. With the rapid development of smart substations, the real-time requirements of grid monitoring data are getting higher and higher, and the organic combination of streaming media processing technology and intelligent substation will inevitably become the mainstream trend in the future.

Cluster analysis is based on big data analysis, defining a large number of complex categories with attribute data such as quantity, speed and diversity. In addition, a large amount of basic data is quantized by aggregating phase categories or similar categories of data. Therefore, it is possible to extract, estimate, and predict valid information from the data of the same type of attribute. Combined with analytical methods such as cross-category correlation analysis, data can be refined to a high level, making full use of discrete, unordered and complex basic data information. After collecting, analyzing and collating a large amount of basic data, a relatively stable data filling resource is obtained through cluster analysis. And how to identify the intrinsic links between these well-defined and well-defined data so that the data can be fully analyzed and utilized from varying degrees. This is a problem that needs to be addressed in big data analytics. The actual meaning of the so-called association analysis refers to the seemingly unrelated data or information, trying to start the correlation analysis from different angles and the data correlation analysis method obtained from the comprehensive judgment. By associating different types and levels of information, we can make the data after clustering more closely connected with the data between different categories. Providing data analysts with a reliable source of reference data information and saving time in complex data analysis processes is easier. Mining data so that the data analyst to better understand the data, and the prediction analysis allows to make some analysts predict the results of determination of visual analysis and data mining.

Computer technology can support the collation and filtering of large amounts of irrelevant data when a large amount of basic data is collected and accessed through the database. However, for mobile communication optimization with strong awareness, it is necessary to provide initial and established decision support for the optimizer. The mobile communication network in the Bihuang area has hundreds of millions of users, and the amount of data is huge. Here, the decision tree method is used to process a large amount of data analysis processing and problem location in daily optimization work. The decision tree algorithm finds some deep information with important value by purposefully classifying the underlying data. The biggest advantage is that you can use simple language fast classification and description, very suitable for large-scale data analysis and processing.

2.4 Unstructured data

The processing and computing power of the existing traditional analytical system architecture is facing the impact of the rapid growth of big data scale and complexity. According to the research report, the volume of data in various fields is expanding, and the scale of data collection has been measured. It has risen from GB and TB to EB and ZB, and there are many types of data. In addition to a wide range of data sources, data types are diverse, and data structures are not only traditional structured data, but also unstructured data. This makes traditional data storage solutions more and more unsuitable for current data structures, and their requirements for data processing capabilities are increasing. Unstructured data usually cannot directly understand its content and must be opened by the corresponding software. It brings a lot of trouble for future data retrieval. Moreover, the data are not easy to understand, and the meaning of their expression cannot be directly obtained from themselves. Unstructured data have no defined structure, cannot be standardized, and are not easy to manage, so querying, storing, updating, and using these unstructured data require a smarter system.

Office documents, text, images, images, and audio and video information in all formats are unstructured data. (1) In terms of text, the traditional full-text search technology is based on keyword matching and the results are difficult to meet the demand. Intelligent search uses word segmentation dictionary, synonym dictionary, and homonym dictionary to improve the retrieval effect and combines user retrieval context analysis and user-related feedback technology to assist the query. It provides users with intelligent knowledge prompts and finally returns valid information to the user accurately. Prerequisite for realizing the functions is required to use a text segment, word frequency, text analysis to analyze text, text clustering, semantic analysis, text mining, and other text feature extraction techniques, a preprocessing operation on text library, as a result the input of the next layer of modules to achieve similarity text search. (2) Image, image feature extraction is based on image analysis technology. Image feature extraction is the use of computer extraction capabilities. Image feature extraction includes the following three levels: The main visualization function extracts the original features of the image, such as color, edge, shape, texture, layout, and so on. Intermediate object features are local features that extract images from external knowledge and logical reasoning (such as specific objects or characters). Advanced abstraction requires more external support to perform feature extraction on abstract attributes of an image, including specific events, specific content, or style image features. (3) Audio and audio analysis techniques include audio feature extraction, audio classification, and more. In the audio feature extraction, information about frequency domain energy, subband energy ratio, zero crossing rate, bandwidth, and the like in the audio is included. And the audio clip ratio in the audio clip, the subband energy ratio average, the spectrum traffic, and other content undergo corresponding feature extraction. Extracted features can be used for audio matching and recognition. (4) Video, video is currently the most complex type, and common video data may contain rich information such as audio, images and text. At the same time, because each video file is much larger than the other data, the problem is complex and variable. Video analytics techniques can rely on the analysis techniques of the above categories of unstructured data. For example, image recognition techniques can be used to extract key frames from a video, and the results obtained can be used as an image summary of the video, or an image index can be built for these key points to implement a video indexing service. In the analysis of unstructured data technology, the key method is to extract features from unstructured data, and the features obtained are usually high-dimensional data. High-dimensional feature extraction involves "distance" and "dimension reduction" problems. Desirable feature extraction algorithm has lower measured values for the degree of distance keeping as follows:

$$\text{Stress} = \sqrt{\frac{\sum_{i,j}\left(d'_{ij} - d_{ij}\right)^2}{\sum_{i,j} d_{ij}^2}} \tag{8}$$

3 Experiments

The machine-structure-based IoT unstructured big data analysis algorithm proposed in this paper belongs to the online terminal analysis algorithm. The specific design is as follows: The online terminal analysis algorithm infers the model and input data form to obtain the final result of the data, mainly for unstructured data design. For direct application-oriented scenarios, the OTA-selected training set instance consists of unstructured data. The OTA uses the adjacent node distance as a weighting parameter to evaluate the correlation. Figure 1 shows the file reading process of the online terminal analysis algorithm.

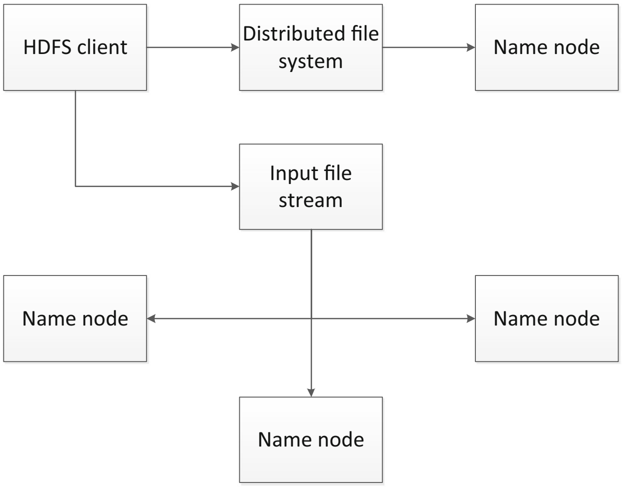

Fig. 1 Schematic diagram of OTA internal file reading process under HDFS

In order to deeply analyze the performance of the online terminal analysis algorithm studied in this paper, this paper analyzes the performance of the original data based on the Internet of Things sensor for big data analysis. Due to the large amount of user data information, a big data platform was created to test the data and then configure the platform. Building a test big data platform uses Ubuntu Linux 10.04, Hadoop 1.03, and SunJava6. Hadoop needs to enable SSH access, and SSH can manage remote nodes and local nodes. After the configuration is complete, the operational data will be fully analyzed. Table 1 shows the time and number of nodes used for each analysis.

4 Discussion

In the four data analysis experiments, the number of sensor nodes and the time used in the experiment are shown in Fig. 2.

As it can be seen from the figure, as the number of nodes increases, the time spent on the data processing is also increased, especially when the number of nodes increased

Table 1 Time and number of nodes used for data analysis

Number of experiments	Transport node	Analysis time
1	2000	200,000
2	5000	420,000
3	11,000	810,000
4	240,000	7,400,000

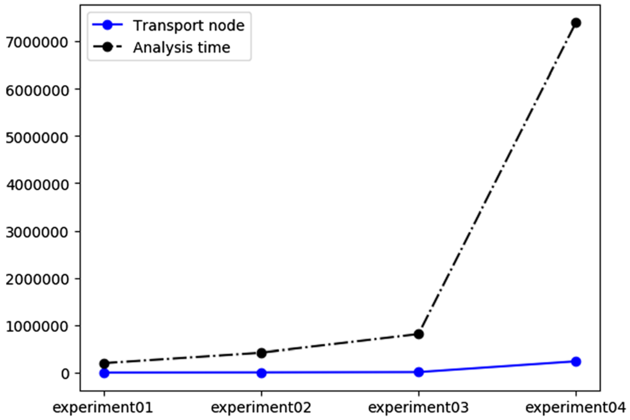

Fig. 2 Time and number of nodes used in the experiment

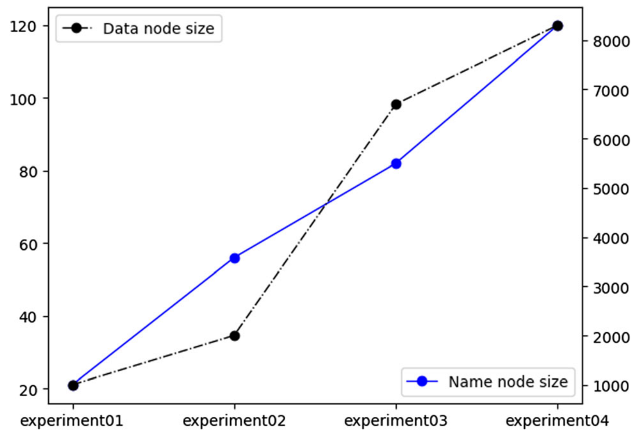

Fig. 3 Node name, data size, and data node

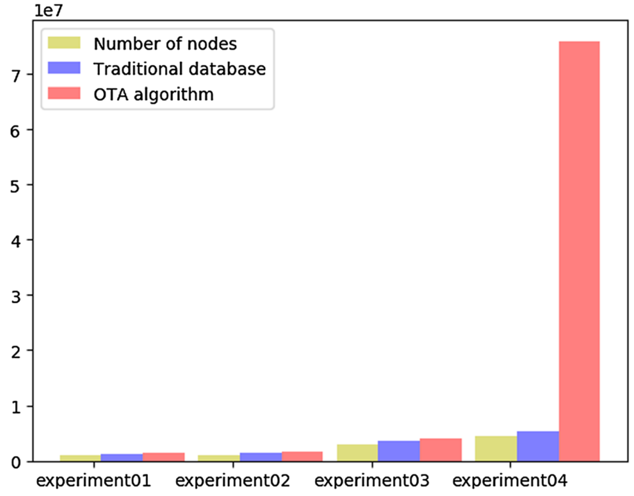

Fig. 4 Comparison of results between traditional database and online terminal algorithm analysis

to 240,000. The processing time has a very large increment so that the data analysis results of the online terminal analysis algorithm during the running process can be comprehensively evaluated, and the experiment continues to analyze the name node and the data node. The results are shown in Fig. 3.

It can be seen in Fig. 3 that the name node size is positively correlated with the data node size, and the amount of data analysis results obtained by the OTA execution can be evaluated. In terms of the number of processing operations per second, the number of nodes in four runs is the same as the number of nodes in Table 1.

As can be seen from Fig. 4, the efficiency of data analysis in traditional databases is lower than that of online terminal algorithms. In particular, when the number of nodes in sensor networks is large, the efficiency of online terminal algorithms is much higher than that of traditional databases, leading to good results.

5 Conclusions

Relational databases have evolved decades of structured data management technology and are now mature. However, unstructured data account for approximately the total amount of information. Its complexity is much higher than structured data. Therefore, how to effectively manage unstructured data becomes the top priority of data management. The management of unstructured data through structured data obtained by heterogeneous data transformation is a very effective method. All in all, in the process of big data development, the amount of data information is increasing rapidly, and the traditional single machine learning algorithm cannot meet the basic requirements of social development. The massively parallel machine learning algorithm can meet the needs of the development of the big data era and can effectively adapt to the basic development requirements of artificial intelligence. Promoting social modernization is the focus of future development of machine learning. This paper explores the application of machine learning algorithms in unstructured data analysis through the support of big data storage technology, big data analysis technology and big data processing technology.

Funding This study was supported by National Key R&D Plan (Grant No. 2018YFB0605504), Fundamental Research Funds for the Central Universities (Grant No. JB2019078), the Postdoctoral Innovative Talent Support Program of China (Grant No. BX20180098), and the China Postdoctoral Science Foundation (Grant No. 2018M640102).

References

1. Yang C, Huang Q, Li Z et al (2017) Big Data and cloud computing: innovation opportunities and challenges. Int J Digital Earth 10(1):13–53
2. Akter S, Wamba SF (2017) Big data and disaster management: a systematic review and agenda for future research. Ann Oper Res 9:1–21
3. Chaurasia SS, Rosin AF (2017) From Big Data to Big Impact: analytics for teaching and learning in higher education. Ind Commer Train 49(7/8):321–328
4. Zhang Y, Qiu M, Tsai CW et al (2017) Health-CPS: healthcare cyber-physical system assisted by cloud and big data. IEEE Syst J 11(1):88–95
5. Zhao L, Chen Z, Hu Y et al (2018) Distributed feature selection for efficient economic big data analysis. IEEE Trans Big Data PP(99):1–1
6. Kim JH (2017) A review of cyber-physical system research relevant to the emerging IT trends: industry 4.0, IoT, Big Data, and Cloud Computing. J Ind Integr Manag 2(3):1750011
7. Sharma PK, Chen M-Y, Park JH (2017) A software defined fog node based distributed blockchain cloud architecture for IoT. IEEE Access PP(99):1–1

8. Pérez H, Gutiérrez JJ, Peiró S et al (2016) Distributed architecture for developing mixed-criticality systems in multi-core platforms. J Syst Softw 123:145–159
9. Yuan J, Holtz C, Smith T et al (2017) Autism spectrum disorder detection from semi-structured and unstructured medical data. EURASIP J Bioinf Syst Biol 1:3–12
10. Banowosari LY, Purnamasari D (2016) Approach for unwrapping the unstructured to structured data the case of classified ads in HTML format. Adv Sci Lett 22(8):1909–1913
11. Lai L, Lu Q, Lin X et al (2017) Scalable subgraph enumeration in MapReduce: a cost-oriented approach. VLDB J Int J Very Large Data Bases 26(3):421–446
12. Cheng D, Jia R, Guo Y et al (2017) Improving performance of heterogeneous MapReduce clusters with adaptive task tuning. IEEE Trans Parallel Distrib Syst 28(3):774–786
13. Zhang Y, Wen J (2017) The IoT electric business model: using blockchain technology for the internet of things. Peer-to-Peer Netw Appl 10(4):983–994
14. Schulz P, Matthe M, Klessig H et al (2017) Latency critical IoT applications in 5G: perspective on the design of radio interface and network architecture. IEEE Commun Mag 55(2):70–78
15. Mehdi M, Al-Fuqaha A, Sorour S et al (2018) Deep learning for IoT big data and streaming analytics: a survey. IEEE Commun Surv Tutor 20(4):2923–2960
16. Zhu D (2018) Deep learning over IoT big data-based ubiquitous parking guidance robot for parking near destination especially hospital. Pers Ubiquit Comput 4:1–8
17. Liu C, Yang C, Zhang X et al (2015) External integrity verification for outsourced big data in cloud and IoT: a big picture. Future Gener Comput Syst 49(C):58–67
18. Zhang Q, Yang LT, Chen Z et al (2018) High-order possibilistic c-means algorithms based on tensor decompositions for big data in IoT. Inf Fusion 39:72–80
19. Vijaykumar S, Saravanakumar SG, Balamurugan M (2015) Unique Sense: smart computing prototype for industry 4.0 revolution with IOT and Bigdata implementation model. Indian J Sci Technol 8(35):1–4
20. Bi Z (2017) Embracing internet of things (iot) and big data for industrial informatics. Enterp Inf Syst 123:145–159
21. Yeu CWT, Lim M-H, Huang G-B et al (2012) A new machine learning paradigm for terrain reconstruction. IEEE Geosci Remote Sens Lett 3(3):382–386
22. Hassan MK, El-Desouky AI, Badawy MM, Sarhan AM, Elhoseny M, Manogaran G (2019) EoT-driven hybrid ambient assisted living framework with naïve Bayes-firefly algorithm. Neural Comput Appl 31(5):1275–1300

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.