Computing and Informatics, Vol. 45, 2026, 365–381, doi: 10.31577/cai_2026_2_365

BIG DATA INFORMATION MANAGEMENT AND DATABASE CONSTRUCTION BASED ON MACHINE LEARNING ALGORITHM

Rongguo Fu, Zhu Qi*, Yao Zhang, Nan Chen

Information Center, Jiangsu Earthquake Agency
Nanjing 210014, Jiangsu, China
e-mail: quick_911@163.com, qizhuxmu@126.com, {zy13526866161, chennan19911107}@163.com

Abstract. The database management system based on Machine Learning (ML) algorithm can achieve effective classification by acquiring and analyzing massive data. Therefore, this paper analyzed and studied the Big Data (BD) information management and database construction based on ML algorithm. Artificial intelligence provided a feasibility analysis for this article. The research and analysis results showed that the BD information management efficiency based on ML algorithm was 10.17 % higher than the traditional BD information management efficiency, and the comprehensive evaluation was as high as 88.5. This showed that the BD information management and database construction based on ML algorithm is feasible.

Keywords: Information management, database construction, machine learning, artificial intelligence

1 INTRODUCTION

The development of BD in the Internet industry is faster and faster, and people pay more attention to its storage and use, but there are many deficiencies in the traditional management model. In order to improve database utilization and better solve these problems, BD information management and database construction based on ML algorithms have become the current research focus.

* Corresponding author

With the growing interest in BD management, database construction is also imminent, so many scholars have conducted research on it. Gokulkumari G found that the data corresponds to the data with insufficient scale. BD is represented by different algorithms, technologies, tools and architectures. It needs to improve ways to solve new and old problems, which provides a reference point for BD management [1]. According to Raguseo's research, BD was considered to be one of the most important fields of technology in the future, and it was quickly drawing the notice of many industries because of the high-value services it can offer [2]. Wamba et al. believed that Big Data Analysis (BDA) has received great attention from scholars and practitioners, but there is still a big gap in understanding how organizations effectively manage and use BDA resources [3]. The study of Akhtar et al. found that previous research on BDA has emphasized the significance of selected BD skills and competencies for a successful organization, but has largely ignored research on the relationship between the use of skills in cross-functional teams and the related role of that particular data-driven action [4]. Liu's study found that the database based on BD and deep learning algorithms had higher adaptability of results, good fitting for large-scale data, and greatly improved the precision and efficacy of retrieval [5]. According to the research of Stadler et al., one of the most restrictive bottlenecks in developing useful AI models was the lack of training data. A large number of cases were required, and high-quality ground annotation was required. Therefore, a database needs to be established and described [6]. The research of Ben Mahria et al. found that the purpose of converting relational databases into ontologies was to offer applications based on the semantic representation of data. The use of ontologies to represent data has proven to be a useful facility for the management and interchange of data, which is the reason why it is important to build a bridge between relational databases and ontologies [7]. The above is about BD information management and database construction.

Due to the rapid development of artificial intelligence, many researchers have conducted in-depth research on ML algorithms. Sarker believed that the digital world has rich data. To intelligently analyze this data and develop correspondingly intelligent and automated applications, Machine Learning (ML) is key, and ML algorithms can be used to increase the smartness and capabilities of applications [8]. The research of Flah et al. found that ML algorithms have been extensively employed for structural health monitoring because of their excellent capability in detecting damage and defects in residential civil engineering structures. Along with the advent of IoT, BD and a huge and sophisticated backlog of ageing residential infrastructure properties, this application would grow very rapidly [9]. Zhong and Enke believed that BDA techniques related to ML algorithms were playing an incremental key to a variety of fields of application [10]. The research of Baraneetharan found that ML can enhance the lifetime of a network by providing real-time solutions that take maximum advantage of the resources in the network [11]. Huang et al. believed that the rapid development of ML technology can be used as a regression problem to help people find rules from massive data and achieve prediction results. Currently, this technology is applied in many fields, such as weather forecasting, medical diagnosis

and financial prediction [12]. The research of Roh et al. found that data collection was a major bottleneck of ML and an active research topic. Data collection has become a key problem. The main reason was that, unlike traditional ML, deep learning technology would automatically generate features, which saves the cost of feature engineering. However, as a reward, more labeled data may be required [13]. The above is the relevant research on ML algorithms. It can be seen that ML algorithms have been applied to many fields.

BD is characterized by sparse attributes, super-high dimensions, complex relationships, etc. Traditional algorithms are powerless in the face of such data scale. Therefore, this paper proposes the application of ML algorithms in BD information management and database construction.

2 EVALUATION ON BD INFORMATIZATION MANAGEMENT

1) Relevant meanings of BD

The definition of BD has not been standardized [14]. In a narrow sense, the data that cannot be loaded into the computer's internal memory is BD. Although this is not a formal definition, it is easy to understand and accept. Broadly speaking, BD refers to the data that cannot be loaded into the internal memory of the computer within a certain period of time under the conditions of traditional computer technology and software and hardware tools. Effective data mining is to extract hidden knowledge and value from large and complex data, which is an important direction of modern industrial development.

BD needs a new processing model to develop and optimize processes to enrich information assets [15]. It also has good data collection and storage capabilities. From a technical point of view, data cannot be processed on a single computer, but must be organized and analyzed through a structure that is distributed to guarantee good distribution and data collection capabilities. One of the characteristics of BD is the large-scale processing of multiple data and the improvement of information acquisition speed. The ability of BD to process a large amount of information, conduct market calculations, and analyze the risk factors of enterprise transformation and growth is highly evaluated, as shown in Figure 1.

However, with the rapid progress of modern technology, the following problems are encountered in BD management:

1. Data sources are scattered, including statistical micro data, comprehensive data and macroeconomic data from various sources. Macroeconomic data is scattered in various data management departments and business departments, and the list of data resources is not clear, which makes the management and the large-scale interdisciplinary use of data resources difficult.
2. Different types of research history data are stored on fixed hardware devices. It is inefficient to use the corresponding system to search historical data.

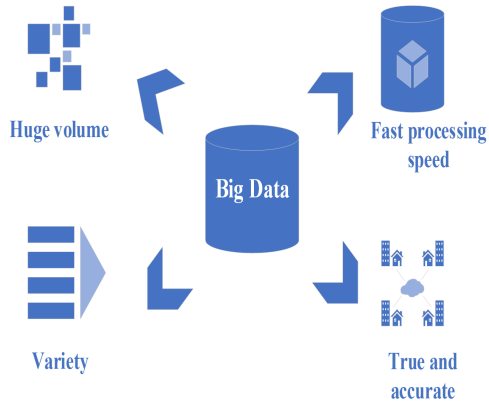


Figure 1. Characterization of the meaning associated with BD

3. The terminology of statistical survey metadata is not uniform among organization management, specialty, data management and different software developers, and modeling methods are not uniform, resulting in poor comparability and continuity of metadata at all stages of the data production process.
4. It is difficult to create time series data for indicators, and it is not easy to quickly view the data changes over time.
5. The data exchange between the database system and the analysis system is weak, and the metadata exchange quality is poor. The main data carriers cannot be directly used in the system, as shown in Figure 2.

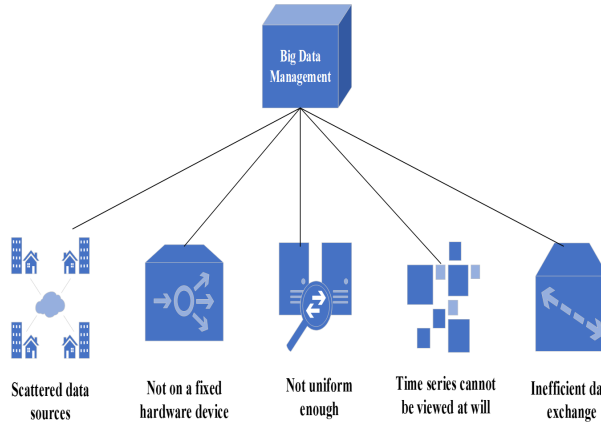


Figure 2. Problems encountered when managing BD

2) Advantages of BD management

According to the 2020 Digital Transformation Trend Report, the proportion of BD information management applied in the Internet industry was 49.6 %, which was also the industry with the highest proportion. The proportion of other industries is shown in Figure 3.

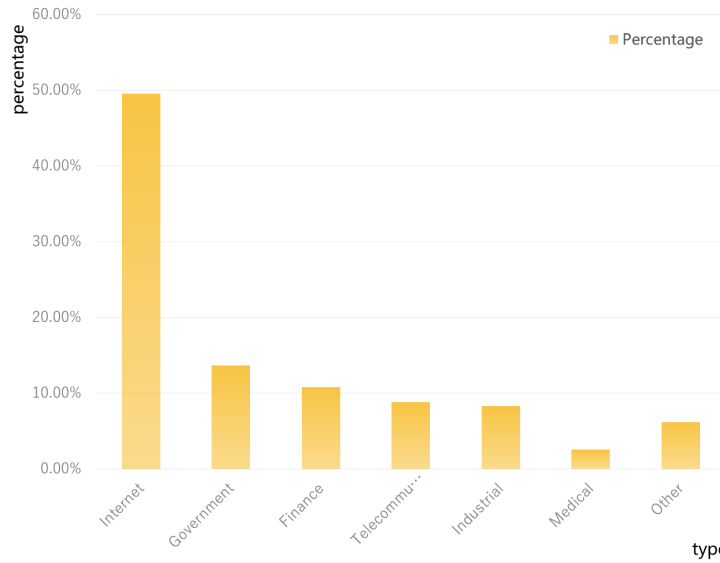


Figure 3. Share of BD management in different industries

It is evident from Figure 3 that the industry with the highest proportion of BD information management was the Internet industry, accounting for 49.6 %. The second was the government and the financial industry, accounting for 13.7 % and 10.8 %, respectively. The industry with the lowest proportion was the medical industry, accounting for only 2.6 %. The application of BD in different industries is shown in Figure 4.

The era of BD has arrived. BD runs through the entire product life cycle, from capturing market demand, product research and development, production and manufacturing, equipment operation, and market service to scrapping and recycling, and even the intelligent link of the product itself.

In order to more rigorously study the market scale of BD information management, this paper now conducts a statistical analysis of the BD market from 2016 to 2021, mainly focusing on the total income and annual growth rate of BD, as shown in Figures 5 and 6.

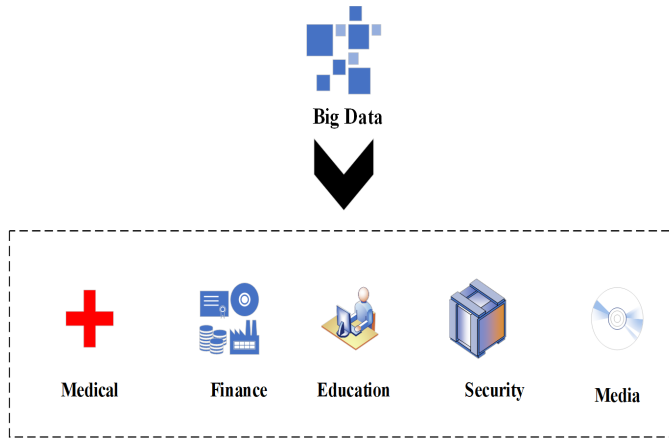


Figure 4. Applications of BD

As can be seen in Figures 5 and 6, the BD market was also on the rise over time. The revenue of BD management would increase from 9.2 billion dollars in 2016 to 20.6 billion dollars in 2021, and the growth rate of BD market would also increase with the growth of the year.

Figure 5. BD market revenues 2016–2021

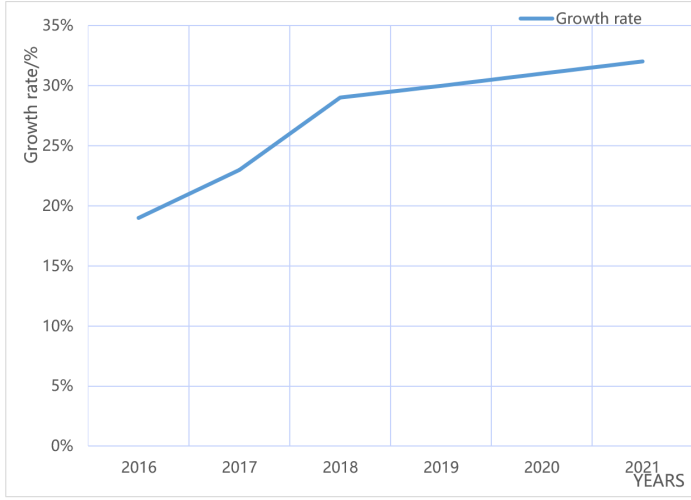


Figure 6. BD market growth rate scenario 2016–2021

The ownership attribute of BD focuses on helping to clarify the list of data assets and the allocation of data resources, and defining the ownership boundary of data assets. BD is the frontier technology of data analysis. Its main value lies in storing and analyzing massive data to achieve the goal of end-to-end cost optimization. In the future, data may become the largest commercial commodity. The characteristics of BD are summarized as four "V": the first is large data volume (Volume); the second, there are many data types (Variety); the third is fast processing speed (Velocity); the fourth is low value density (Value). The value of BD lies in maximizing the use of data through sharing and cross-use. The advantages of applying BD information management are as follows:

1. The use of BD technology for information management is an innovative breakthrough with modern characteristics, which effectively promotes the development of BD computing and makes information management more efficient and reliable. The application of BD has brought great changes to information management. It not only increases the reliability of information management, improves the efficiency of information collection and storage, but also reduces the overall cost of information management and maintenance.
2. Using BD for information management can avoid a series of problems caused by traditional information management methods. Traditional information management uses paper materials for information management. As time goes on, the amount of information becomes more and more, and the space occupied by information materials becomes larger and larger. The maintenance and updating of information materials has also become more and more time-consuming and labor-intensive. In addition, printed materials are easy to be damaged and

lost, which is not conducive to managing a large amount of information. Using BD to manage information data can improve information security, and make information management and updates fast and efficient. It also greatly reduces the data error rate and the cost of project information management.

3. The development of BD and information security is inseparable, so people must have a sense of security. With the continuous growth of data, the pace of information security should also increase. Any security risk would pose a serious threat to the project data. Therefore, people must start from the data itself, strengthen network security monitoring and assertion, and implement various security protection policies in a timely manner to prevent malicious damage or information theft. The development of BD and data security are two inseparable aspects. In the age of BD, it is necessary to do a good job in data collection and storage, and security protection is essential. Doing a good job in data security can ensure the management of BD information.

In the BD environment, information technology is becoming more and more developed, and the network has gradually penetrated into all aspects of people's work and life. In this case, information management must keep pace with the times, and use database technology to improve data records, improve work efficiency, and develop modern information management methods. While making use of large-scale data collection and information data storage, people should pay attention to data security issues and improve the security management system, strengthen the implementation of modern information management methods, and achieve continuous practice and innovation.

3 EVALUATION ON BD INFORMATION MANAGEMENT AND DATABASE CONSTRUCTION OF ML ALGORITHM

In AI, ML brings new opportunities and challenges for managing BD information. There is no doubt that new ideas, concepts and functions have also been introduced in establishing and reconfiguring database intelligence. At present, ML research is characterized by the integration of various disciplines, such as statistics, mathematics, information theory and biological neural networks. Its most famous application fields include face recognition. Database construction is an ideal part of modern construction, which includes the classification, mining and reuse of existing information resources to provide more intelligent and convenient services for everyone. Therefore, ML plays an irreplaceable role in database construction, which would help it develop in an intelligent direction.

At present, it is more important to pay attention to rationalizing the basic relationship between ML speed and speed training, and eliminating the contradiction between them. With the overall development of network technology, data collection and analysis technology have been improved. However, due to the unlabeled data in some fields, the pressure of ML has gradually increased.

In the BD development environment, the efficiency of data production has gradually improved, and the number and types have changed greatly. In addition, important new data lines must be analyzed, such as text data analysis, image content search and image data processing, so that ML research can develop in different directions.

The construction of the database should have the following two main goals. The first is to integrate different raw data resources. Standardized metadata is used to describe the original data from different surveys, so that the original data in various formats can be easily investigated. The second is to provide support for data analysis, so that various applications and analysis tools can easily access the data resources in the basic data, improve the data resources in the database, and increase the value of data resources.

The overall architecture of the basic database is divided into data application layer, data management service layer, application support layer, data resource layer and the infrastructure layer from top to bottom. The basic database framework is shown in Figure 7.

1) Data application layer

The data application layer is directly oriented to users of data analysis and provides basic business functions, such as metadata data search, data analysis, data calculation and other basic business functions of the system. The metadata portal can search and display various metadata, metadata association, metadata attribute information, etc. in the core database.

2) Data management service layer

As the back-end support of data analysis and various data applications, the core of data management service layer is the data resource service platform.

3) Application support layer

The application support layer includes the core development platform, high-performance data loading engine, mixed data access engine, reporting tool, news engine, chart engine, etc.

4) Data resource layer

The data resource layer is composed of four main elements: metabase, integrated library, data warehouse and data mart.

5) Infrastructure layer

The basic database is based on the BDA architecture, integrating the repository, data and data analysis. It includes a distributed database, a distributed file storage system and an application service engine, which jointly serve the data warehouse and analysis applications. It also has the ability to process a large amount of structured and unstructured data, and can meet the requirements of fast search, high-performance data computing and complex data mining.

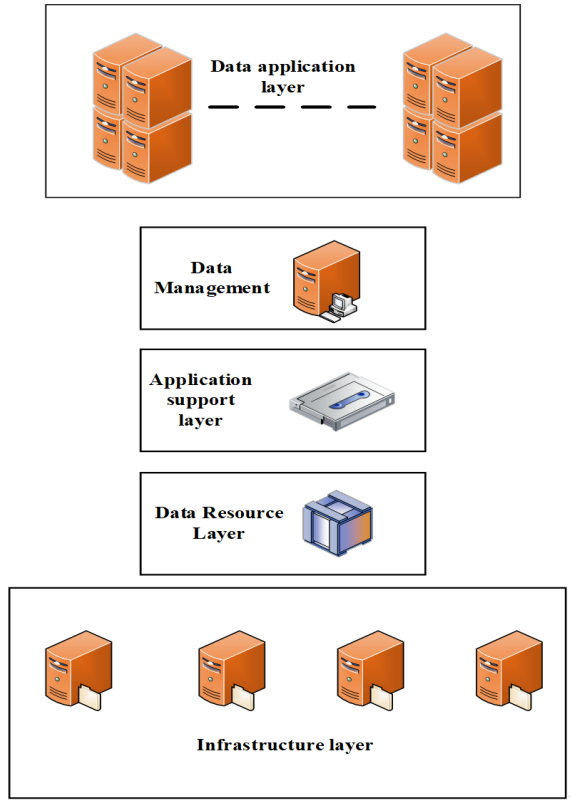


Figure 7. Basic database framework diagram

4 APPLICATION OF NEURAL NETWORK ALGORITHM IN DATABASE INFORMATION MANAGEMENT

ML is an important branch of artificial intelligence. Its purpose is to create or design mathematical models on the basis of existing data by referencing and simulating the intelligent behavior of human brain or mathematical tools, so as to achieve an efficient prediction process for future new data. ML has become one of the main technologies in the age of BD. Traditional ML has two main data requirements. First, the supervised algorithm must use a large number of labeled data and have enough training samples; second, training data and test data must meet independent and identical distribution. The use of ML algorithms can improve the accuracy of data classification, make better use of data and build databases. Generally speaking, ML methods are mainly decision trees and Support Vector Machines (SVM), and ML algorithms based on neural networks can improve the resolution accuracy of databases.

1) BP neural network

BP neural network model is mainly composed of input layer, hidden layer, and output layer. Generally speaking, the hidden layer can have many layers. If $B = \{(x^i, y^i) \in X \times Y\}, i = 1, 2, \ldots, n$, where j is a data sample, and $X \in T^o$ is the data feature space. Among them, o is the number of data features and (x^i, y^i) is the corresponding data sample. The input of BP neural network is expressed by the formulas:

$$\hat{y}_j^i = f\left(\sum_{l=1}^{n} W_{jl} s_l\right), \quad l = 1, 2, \ldots, n, \tag{1}$$

$$s_l = f\left(\sum_{m=1}^{k} V_{ml} x_m^i\right), \tag{2}$$

$$E^i = \frac{1}{2}\sum_{j=1}^{n}\left(y_j^i - \hat{y}_j^i\right)^2. \tag{3}$$

Among them, $\hat{y}_j^i$ is the j^{th} output layer neuron of the i^{th} sample, and E^i is the loss function of the i^{th} castration version. According to the BP neural network diagram, the BP neural network including the hidden layer can be divided into two parts. The first part is the connection between the input layer and the hidden layer, and the second part is the connection between the output layer and the hidden layer. The better BP network classifiers are as follows:

$$\Delta V_{ml} = -\eta \frac{\partial O^i}{\partial V_{ml}}, \tag{4}$$

$$V_{ml} = V_{ml} + \Delta V_{ml}, \tag{5}$$

$$\Delta W_{jl} = -\eta \frac{\partial O^i}{\partial W_{jl}}, \tag{6}$$

$$W_{jl} = W_{jl} + \Delta W_{jl}. \tag{7}$$

2) Machine learning

ML is another important research field of computer science and artificial intelligence applications after expert systems, and it is one of the key research topics of artificial intelligence and neural computing. In the past decade, ML has been applied to many fields of computer science, and has also provided important support for many related disciplines. Learning is a basic feature of human intelligent behavior. It is to strengthen or improve the ability of a system to repeatedly perform the same or similar tasks, so that the system would perform the same or similar tasks better or more efficiently the next time. In a general sense, ML is the ability of computer to simulate human learning behavior and

learn human skills; in a practical sense, ML is a method of training models from existing data to predict unknown data. The application of ML algorithms in BD management has the advantages of high efficiency and high security.

5 EVALUATION ON BD INFORMATION MANAGEMENT AND DATABASE CONSTRUCTION OF ML ALGORITHM

1) Management efficiency based on ML algorithm

This paper compares traditional BD information management with BD information management based on ML algorithm, and analyzes the management efficiency of different data groups. The observation period is 6 cycles, and the detailed analysis is shown in Figure 8.

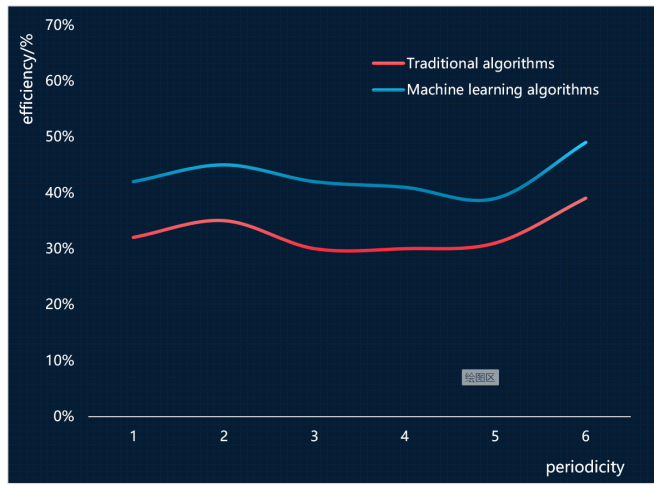


Figure 8. Efficiency of BD information management with different algorithms

It can be seen from the broken line chart in Figure 8 that the efficiency of BD informatization management based on ML algorithm was better than that of traditional BD informatization management. The efficiency of BD informatization management based on ML lgorithm was between 38 % and 50 %, with an average management efficiency of about 43 %. The average management efficiency of traditional BD informatization was 32.83 %, and the management efficiency under ML algorithm was 10.17 % higher than that of traditional management, which showed that ML algorithm can improve the efficiency of BD informatization management.

2) BD informatization management speed

This paper compared traditional BD information management with BD information management based on ML algorithm, and analyzed the management speed

of different data groups. The observation period is 6 cycles, and the detailed analysis is shown in Figure 9.

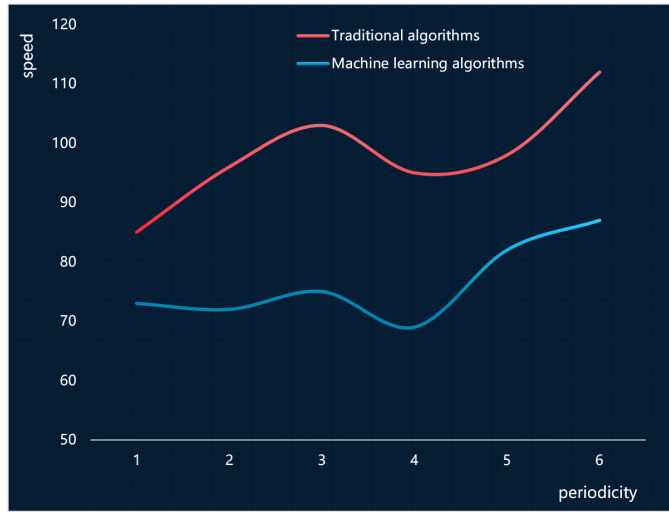


Figure 9. BD information management speed comparison

It can be seen from the broken line chart in Figure 9 that the BD informatization management speed based on ML algorithm was faster than the traditional BD informatization management speed. The BD informatization management speed based on ML algorithm was about 76 on average; the average speed of traditional BD information management was about 98; the speed of BD information management based on ML algorithm was 22 faster than the traditional speed, which showed that the BD information management based on ML algorithm is feasible.

3) Database security

The ML algorithm has superior security performance and can protect data. Now, the security performance under the traditional algorithm is compared with the security performance based on the ML algorithm. The experiment is divided into six times, as shown in Figure 10.

It can be seen from the broken line chart in Figure 10 that the data security performance under the ML algorithm was higher than that under the traditional data security performance, which showed that the ML algorithm can protect data more efficiently.

4) Comprehensive evaluation

The ultimate goal of database construction and information data management is to meet people's needs. This paper is to evaluate the satisfaction of the internal personnel of an Internet company with the database construction and

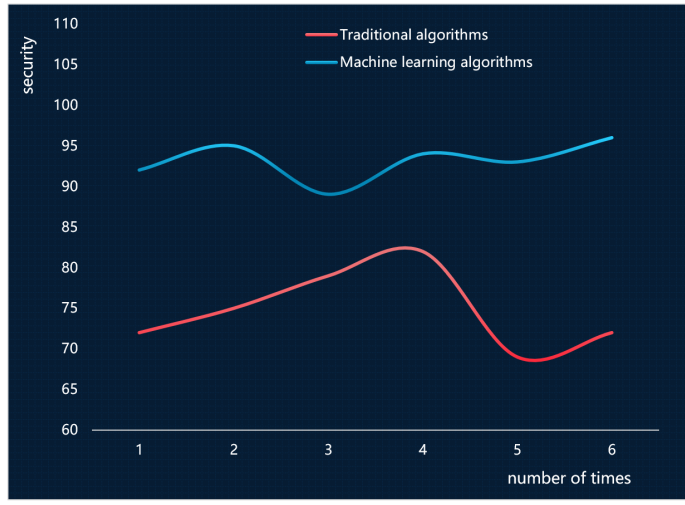


Figure 10. Security analysis study of databases

BD information management under the ML algorithm. The number of people evaluated is 100, and the survey results are four levels: satisfied, good, average, and dissatisfied, as shown in Table 1.

	Databases	Big Data Information Management	Comprehensive Evaluation
Satisfaction	88 %	83 %	84 %
Fine	9 %	13 %	10 %
Commonly	2 %	3 %	6 %
Discontent	1 %	1 %	0 %

Table 1. Satisfaction evaluation of an internet company on database construction and big data information management with ML algorithms

It can be seen from Table 1 that database construction and BD information management based on ML have been unanimously satisfactory. Therefore, this paper has drawn a statistical chart for this, as shown in Figure 11.

It can be seen from Figure 11 that the comprehensive evaluation of database construction and BD information management based on ML algorithm was 88.5.

6 CONCLUSIONS

The construction of BD information management and database is an inevitable trend of the development of the times. With the increasingly extensive application of computer technology, network communication and other information technologies

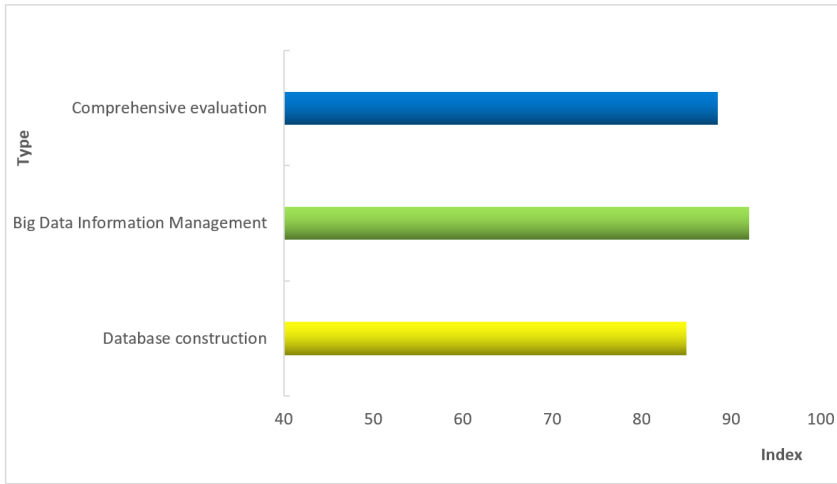


Figure 11. Comprehensive evaluation with ML algorithms

in social life, people also put forward higher requirements for the information industry. This paper built a database based on ML algorithms. The research results showed that database construction based on ML algorithms and BD information management is conducive to improving data management efficiency, and reducing management costs. They also improved data utilization and promoted better operation of BD information management.

Funding

This work was supported by the National Seismic Intensity Rapid Reporting and Early Warning Project.

REFERENCES

[1] Gokulkumari, G.: An Overview of Big Data Management and Its Applications. Journal of Networking and Communication Systems, Vol. 3, 2020, No. 3, pp. 11–20, doi: 10.46253/jnacs.v3i3.a2.

[2] Raguseo, E.: Big Data Technologies: An Empirical Investigation on Their Adoption, Benefits and Risks for Companies. International Journal of Information Management, Vol. 38, 2018, No. 1, pp. 187–195, doi: 10.1016/j.ijinfomgt.2017.07.008.

[3] Wamba, S. F.—Akter, S.—Guthrie, C.: Making Big Data Analytics Perform: The Mediating Effect of Big Data Analytics Dependent Organizational Agility. Systèmes d'Information & Management, Vol. 25, 2020, No. 2, pp. 7–31, doi: 10.3917/sim.202.0007.

[4] Akhtar, P.—Frynas, J. G.—Mellahi, K.—Ullah, S.: Big Data-Savvy Teams' Skills, Big Data-Driven Actions and Business Performance. British Journal of Management, Vol. 30, 2019, No. 2, pp. 252–271, doi: 10.1111/1467-8551.12333.

[5] Liu, F.: Language Database Construction Method Based on Big Data and Deep Learning. Alexandria Engineering Journal, Vol. 61, 2022, No. 12, pp. 9437–9446, doi: 10.1016/j.aej.2022.02.069.

[6] Stadler, C. B.—Lindvall, M.—Lundström, C.—Bodén, A.—Lindman, K. et al.: Proactive Construction of an Annotated Imaging Database for Artificial Intelligence Training. Journal of Digital Imaging, Vol. 34, 2021, No. 1, pp. 105–115, doi: 10.1007/s10278-020-00384-4.

[7] Ben Mahria, B.—Chaker, I.—Zahi, A.: A Novel Approach for Learning Ontology from Relational Database: From the Construction to the Evaluation. Journal of Big Data, Vol. 8, 2021, No. 1, Art. No. 25, doi: 10.1186/s40537-021-00412-2.

[8] Sarker, I. H.: Machine Learning: Algorithms, Real-World Applications and Research Directions. SN Computer Science, Vol. 2, 2021, No. 3, Art. No. 160, doi: 10.1007/s42979-021-00592-x.

[9] Flah, M.—Nunez, I.—Ben Chaabene, W.—Nehdi, M. L.: Machine Learning Algorithms in Civil Structural Health Monitoring: A Systematic Review. Archives of Computational Methods in Engineering, Vol. 28, 2021, No. 4, pp. 2621–2643, doi: 10.1007/s11831-020-09471-9.

[10] Zhong, X.—Enke, D.: Predicting the Daily Return Direction of the Stock Market Using Hybrid Machine Learning Algorithms. Financial Innovation, Vol. 5, 2019, No. 1, Art. No. 24, doi: 10.1186/s40854-019-0138-0.

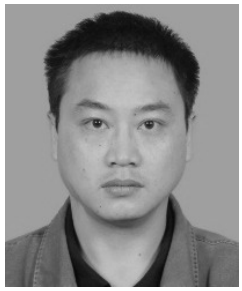
[11] Baraneetharan, E.: Role of Machine Learning Algorithms Intrusion Detection in WSNs: A Survey. Journal of Information Technology and Digital World, Vol. 2, 2020, No. 3, pp. 161–173, doi: 10.36548/jitdw.2020.3.004.

[12] Huang, J. C.—Ko, K. M.—Shu, M. H.—Hsu, B. M.: Application and Comparison of Several Machine Learning Algorithms and Their Integration Models in Regression Problems. Neural Computing and Applications, Vol. 32, 2020, No. 10, pp. 5461–5469, doi: 10.1007/s00521-019-04644-5.

[13] Roh, Y.—Heo, G.—Whang, S. E.: A Survey on Data Collection for Machine Learning: A Big Data – AI Integration Perspective. IEEE Transactions on Knowledge and Data Engineering, Vol. 33, 2019, No. 4, pp. 1328–1347, doi: 10.1109/TKDE.2019.2946162.

[14] Oussous, A.—Benjelloun, F. Z.—Ait Lahcen, A.—Belfkih, S.: Big Data Technologies: A Survey. Journal of King Saud University – Computer and Information Sciences, Vol. 30, 2018, No. 4, pp. 431–448, doi: 10.1016/j.jksuci.2017.06.001.

[15] Niebel, T.—Rasel, F.—Viete, S.: BIG Data – BIG Gains? Understanding the Link Between Big Data Analytics and Innovation. Economics of Innovation and New Technology, Vol. 28, 2019, No. 3, pp. 296–316, doi: 10.1080/10438599.2018.1493075.



Rongguo Fu holds his Master's degree in computer applications technology from the Nanjing Technology University and works at the Information Center of Jiangsu Earthquake Agency. His primary focus is on technical research related to earthquake information network communication, network security, data security, cryptography applications, and information services.

Zhu Qi holds his postgraduate degree and a Master's degree in military communication from the People's Liberation Army University of Science and Technology. He works at the Information Center of Jiangsu Earthquake Agency, primarily engaging in technical work and research related to earthquake information network communication, network security, and data security.

Yao Zhang holds his Bachelor's degree in computer science and computing from the Zhengzhou University. Currently, he works at the Information Center of Jiangsu Earthquake Agency, where he is primarily engaged in technical research on earthquake information network communication, information innovation, and classified information systems.

Nan Chen holds her Master's degree in computer technology from the Northeast Electric Power University. She works at the Information Center of Jiangsu Earthquake Agency, primarily focuses on earthquake information network communication and network security.