DOI: 10.23977/jaip.2023.060107
ISSN 2371-8412 Vol. 6 Num. 1

Research on Artificial Intelligence Applications Based on Data Mining Algorithms in the Era of Big Data

Lihua Luo

College of Software Engineering, Guangdong University of Science and Technology, Dongguan, Guangdong, 523000, China

Keywords: Big data; data mining algorithms; artificial intelligence applications

Abstract: At present, data has penetrated into every industry field and has become an important production factor. The scale of data is also expanding at an alarming rate. Big data is becoming the most significant label of this era. The concept of big data has subverted our understanding of traditional data, and has also caused technological changes in data acquisition, storage, analysis, mining, and visualization. The development of big data and related technologies is changing the important foundation of human production and lifestyles. Based on the understanding of the connotation of big data and data mining algorithms, this paper discusses the application of artificial intelligence based on data mining algorithms on its technical system and analyzes its future development trends, hoping to provide some reference for related research.

1. Introduction

Big data is profoundly changing the way human society survives. Big data and its technological development not only show great value in business, but also in the industrial, medical, agricultural, and aviation fields. Based on the collection, analysis and processing of massive data, it provides unprecedented scientific decision-making. Information support, the generation and effective analysis of big data will become an important force in the change of human survival. The real-world information discovery process includes more knowledge preprocessing, visual steps, machine learning, and evaluation. Therefore, big data and data mining need to consider complex experimental styles, optimization, clear processing and parameter processing to solve such an environment. The era of big data makes people more eager to acquire valuable knowledge from rapidly growing information. People tend to rely on the era of big data because they find it more efficient to use mass storage of data. Traditional data processing technology and artificial intelligence are very different in terms of extracting relevant information from existing data. Traditional data processing technology can only realize basic functions such as statistics and data query, but cannot extract knowledge from existing data. Compared with human intelligence, artificial intelligence is more effective and more complex. In this way, artificial intelligence plays a vital role in processing data mining, especially big data. Humans tend to consider data close to them and data accumulated in the past to make wise decisions. However, artificial intelligence does not rely on such instructions, but on a large amount of data dump to obtain a clear goal.

Therefore, people are more willing to use artificial intelligence to extract information from massive amounts of existing data. Artificial intelligence has a profound impact on data mining and big data. It affirms that artificial intelligence relies on massive amounts of data to obtain sufficient and effective information. [1].

2. The impact of artificial intelligence on big data and data mining

The essential core of big data is a data set, which is a complex data set that is quite different from traditional data in terms of acquisition, storage, analysis and processing, and requires special technical support. The strategic significance or value manifestation of big data does not lie in the mastery of massive data, but in the analysis, processing and processing capabilities of these data. Data mining technology has been used in various fields to assist in predicting challenges, segmentation, classification, correlation and diagnosis. People like artificial intelligence in data mining to obtain practical information from massive amounts of data [2].

According to research, artificial intelligence of big data and data mining is widely used in various fields to solve key issues such as classification, diagnosis, analysis of customer data, enhanced planning, collection and promotion of rapid calculations. The core goal of introducing artificial intelligence is to identify and reflect on recent developments by introducing various algorithms, and to adopt advanced information by introducing various algorithms. Because they are essential for preserving large amounts of generated data. By promoting security to prevent unauthorized personnel from reaching unauthorized personnel, data mining and big data are also essential [3]. Artificial intelligence improves efficiency by processing complex analysis tasks beyond human imagination, confirming that learning rules are the standard for identifying artificial neutral networks. It clarifies that output and input information is critical to these networks because they are responsible for communicating the data for the best purpose [4].

3. The connotation of data mining algorithms

Data mining is a process of using minute data analysis methods or tools to build various models from a large amount of data, discover the relationship between the data, and then obtain unknown valuable information. Data mining is an interdisciplinary subject, involving many fields such as statistics, machine learning, and high-performance computing. Complete data mining usually includes several key steps such as data cleaning, data integration, data specification, data change, knowledge discovery, pattern evaluation, and knowledge representation [5]. The purpose of data cleaning is to eliminate data noise and data inconsistency, and make it Comply with data mining algorithm specifications; data integration is the combination of different data sources, and sometimes it is necessary to eliminate the data redundancy in them; data protocol is to extract relevant data from the collected data as needed to reduce the consumption of data mining as much as possible The purpose of data transformation is to use techniques such as smoothing and aggregation processing to transform data into a form suitable for mining; knowledge discovery is the use of various data mining algorithms to mine useful new information from the data; pattern evaluation is the use of measurement methods The secondary results of knowledge discovery are evaluated to verify whether the data mining results are correct; knowledge identification is to display the mining knowledge in a visual manner. The process of data mining usually needs to be repeated many times, or one of the steps needs to be adjusted or re-executed [6].

Data mining algorithms are essentially statistical methods based on statistics, but data mining algorithms are different from general statistical methods, that is, data mining is for non-random samples, while statistical methods are for random sampling samples, which results in the results of data mining algorithms. The conclusion is more scientific [7]. At present, there are many kinds of data mining algorithms, and practical applications include shopping basket, MBR, decision tree, cluster analysis, etc. Algorithms are the key link in computer information data processing. Using algorithms can create data mining techniques, such as association rule data mining. For example, a shopping basket, its manifestation is that when a user enters the platform again after browsing a certain product on an e-commerce platform or the Internet, the platform will recommend similar products. This is a data mining algorithm under association rules [8].

4. Application of data mining algorithms

From the perspective of big data, the data collected and stored by computers that can be indexed cannot meet the needs of big data analysis. Because invalid data may exist, it must be preprocessed to eliminate invalid data. Typically, this step is based on the data index. After invalid data is deleted, the useful data formed can be stored in the database in a unified format. The core of big data technology is algorithm. Scientific and reasonable algorithm will improve the efficiency and accuracy of data processing. Data mining algorithms are used to refine data and conduct accurate data analysis [9]. Data mining algorithm is a set of heuristic methods and calculations for creating data mining models based on data. As a classical algorithm of data mining algorithm, decision tree algorithm is taken as an example to introduce the application of data mining algorithm.

Decision tree is a tree structure used for classification, where each internal node represents the test of attributes, each edge represents the test results, and leaf nodes represent classes or class distribution. The decision process of the decision tree needs to start from the root node of the decision tree, compare the data to be tested with the characteristic nodes in the decision tree, and select the next comparison branch according to the comparison result until the leaf node is used as the final decision result. The decision tree algorithm has the advantages of simple implementation, small calculation, transparent decision-making process, repeatable, and more obvious advantages [10].

In 1986, Ross Quinlan proposed to use "information entropy" and "information gain" as the criteria for selecting nodes to build a decision tree. This algorithm is ID3 (Iterative Dichotomiser 3) algorithm. In ID3 algorithm, "information gain" is taken as the measure of purity, that is, the measure of feature selection. The calculation formula of information gain is: information entropy - conditional entropy, with which the information gain of each feature can be calculated. Each time a node is created, only the feature with the largest information gain needs to be selected as the current node. The more orderly a system is, the lower the information entropy is; On the contrary, the more chaotic a system is, the higher the information entropy is.

The decision tree algorithm uses the principle of information entropy to classify data. The value of information entropy can characterize the degree of confusion of the data. Information entropy is defined as formula (1):

$$\mathrm{H} = -\sum_{i=1}^{n} P_i \ln(P_i) \tag{1}$$

In formula (1), H is the information entropy of the data set, P_i is the probability of the corresponding data i occurring in the entire data set, and n is the number of data classes in the data set.

The implementation process of the decision tree algorithm is as follows:

1) Calculate the information entropy of the entire data set, as in formula (2):

$$H_1 = -\sum_{i=1}^{n} \frac{n_i}{N} \ln(\frac{n_i}{N}) \tag{2}$$

Equation (2) is the calculation method of the entropy of the entire data set. Among them, n_i is the number of each type of data in the data set, and N is the total amount of data in the data set;

2) Calculate the gain of information entropy, the calculation method of gain of information entropy is shown in formula (3):

$$\mathrm{C(T)} = \sum_{t\in leaf} N_t \cdot H(t)$$

$$\Delta C = C_{ij} - C_i \tag{3}$$

In formula (3), N_t is the probability of the corresponding branch, $C_i$is the information entropy of the data set, and $C_{ij}$is the information entropy after adding the segmentation attribute j. Then, the attribute with the largest entropy gain is selected as the root node, and the segmentation is completed. Re-execute steps 1) and 2) of the above decision tree algorithm. Thus, the classification model of the decision tree can be established [11].

The ID3 algorithm decision tree model is shown in Figure 1. The circle and box represent the internal node and leaf node respectively.

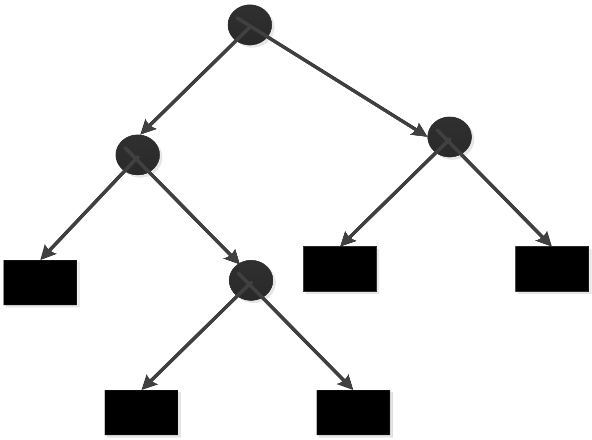

Figure 1: Decision tree model

ID3 algorithm uses the information gain as the evaluation criterion, and selects the feature that maximizes the information gain as the current node each time. The main problem of ID3 algorithm is that it cannot handle continuous features, that is, continuous values cannot be used in ID3 algorithm. ID3 algorithm has not considered the problem of fitting. In 1993, Ross Quinlan improved the ID3 algorithm and proposed a new C4.5 algorithm. Decision tree C4.5 algorithm uses the information gain ratio to select attributes, that is, to select the information gain ratio to select the best feature. 2. The information gain ratio metric information gain ratio index is jointly defined by the gains (D, X) and segmentation information (D, X) in ID3 algorithm. The segmentation information measures that the segmentation information (D, X) is equal to the entropy of feature X (the values are x1, x2,..., xn, and their respective probabilities are P1, P2,..., Pn, and Pk are the number of feature X values in the sample space Xk divided by the total number of sample spaces).

The processing of continuous distribution features C4.5 converts continuous attributes into discrete attributes before processing. If there are N samples, there are N-1 discretization methods:<=v_j to the left subtree,>v_j to the right subtree. Calculate the maximum information gain rate under N-1 conditions.

(1) Sort the feature values in ascending order.

(2) The midpoint between the two characteristic values is taken as the possible split point, the dataset is divided into two parts, and the information gain (InforGain) of each possible split point is calculated. The optimization algorithm is to calculate only those feature values whose classification attributes have changed.

(3) The split point with the maximum information gain (InforGain) after correction is selected as the best split point of the feature.

(4) Calculate the Gain Ratio of the best split point as the Gain Ratio of the feature. Calculation process of maximum information gain rate as shown in Figure 2.

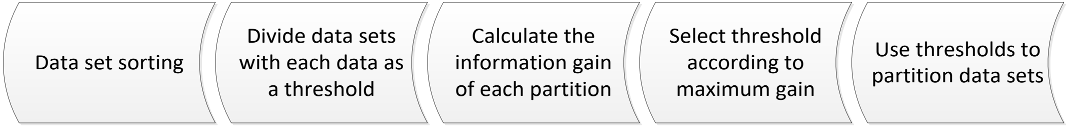
Data set sorting
Divide data sets with each data as a threshold
Calculate the information gain of each partition
Select threshold according to maximum gain
Use thresholds to partition data sets

Figure 2: Calculation process of maximum information gain rate

In C4.5 algorithm, a method for discretization of continuous attributes is designed. The algorithm uses dichotomy to segment continuous attributes. Table 1 shows the comparison between ID3 and C4.5 decision tree algorithms.

Table 1: Comparison of two decision tree algorithms

Algorithm / **Attribute**	**ID3**	**C4.5**
The most characteristic selection method	Information gain	Information gain rate
Branching mode	Multi branch point	Multi branch point
Variable Type	Discrete variable	Discrete and continuous variables
Whether missing value processing	False	True
Scope of application	Classification	Classification, regression
Sample Properties	Single use	Single use

Therefore, C4.5 algorithm cleverly chose a compromise approach. The algorithm does not directly select the feature with the highest information gain rate as the optimal feature, but first finds the feature with higher information gain than the average among all candidate features to ensure that the selected feature probability is good, and then selects the feature with the highest information gain rate from them to ensure that the extreme feature will not be selected finally, which can be said to maximize the advantages of the two indicators.

5. Artificial intelligence research design based on data mining algorithms

Based on the artificial intelligence application of data mining algorithms, research design is a research design that includes a mixed method of qualitative and quantitative information. Qualitative research is less expensive because it considers second-hand resources based on appropriate techniques that facilitate data mining algorithms to obtain information. In the research, qualitative and quantitative research design elements will be combined to show the research results [12]. In addition, it is appropriate to use literature review as another research method to collect and determine the relationship between data mining algorithms and artificial intelligence. Libraries and Internet research on research projects will help gather accurate information [13]. Finally, an evaluation is needed to confirm that artificial intelligence is necessary for extracting large amounts of data. Collect complete information, thereby saving time and costs in the research process. Generally, the above-mentioned research methods are related to obtaining and supplementing data to advance the research plan[14]. After collecting the required information, you need to include a data analysis plan to analyze and understand the various methods used in collecting the information. The data analysis plan focuses on examining artificial intelligence. Various methods of analyzing qualitative data are questionnaire surveys, surveys, observations and interviews. The core reason for analyzing qualitative data is to ensure that the data is useful, transparent, analyzable, and credible to improve the overall quality [15].

6. Conclusion

As an advanced application of computer information data processing technology, big data effectively improves the capability of computer information data processing technology, which is a key point to promote the progress of the big data industry [16]. A large number of data mining algorithms have been widely used in scientific research or industry. As people’s needs become more diversified and companies’ business needs become more complex, artificial intelligence helps to obtain valuable information that is effective for specific processes in data mining. Artificial intelligence is effective in processing existing data mining algorithms. Algorithm performance, algorithm scalability and other aspects still need to be further improved and improved, and a lot of manpower and material resources are still needed to carry out related research [17].

References

[1] Gao Yuan. Thoughts on computer information processing technology based on the "big data" era [J]. Digital Communication World, 2018(09):195.

[2] Shi Pingping. Analysis and research on computer information processing technology in the era of big data [J]. Information and Computers (Theoretical Edition), 2019(15): 18-19.

[3] Wei Haozhi. The application of computer science in the era of big data information [J]. Electronic Technology and Software Engineering, 2018(09): 179-180.

[4] Zhang Yali. Discussion on computer information processing technology in the era of big data [J]. Computer Programming Skills and Maintenance, 2019(07): 86-87+102.

[5] Luo Tianqi. Analysis of computer information processing technology under the background of the big data era [J]. Electronic Components and Information Technology, 2021, 5(01): 64-65.

[6] Chen Yingquan. Analysis of computer information processing technology under the background of the big data era [J]. Information and Computers (Theoretical Edition), 2021, 33(01): 209-210.

[7] Xiong Yong. Analysis of computer information processing technology under the background of the big data era [J]. Computer Knowledge and Technology, 2021, 17(01): 32-33+40.

[8] Zhou Sheng. Give full play to the role of "big data" in scientific research management in universities [J]. China University Science and Technology, 2017(12): 89-91.

[9] Wu Lu. The value logic of learning evaluation supported by big data [J]. Educational Research of Tsinghua University, 2019, 40(1): 15-18.

[10] Zhang Liwei, Yang Jun, Liao Xi, et al. Design and development of a decision support system for high school students' subject selection based on big data [J]. Modern Educational Technology, 2018, 28(8): 5-11.

[11] Li Sa. Research on the analysis method of learning behavior association based on association rules [J]. Microelectronics and Computer, 2018, 35(6): 71-74.

[12] Zhang Guiyuan. Research on the application of association rules based on data mining in score analysis [J]. Computer Programming Skills and Maintenance, 2016, 23(9): 60-61.

[13] Zhang Tian, Yin Changchuan, Pan Lin, et al. Student performance analysis based on improved clustering and association rule mining [J]. Journal of Beijing University of Posts and Telecommunications (Social Science Edition), 2018, 20(2): 91-96.

[14] Zhang Yufeng, Zeng Yitang. Research on logistics information analysis model based on dynamic data mining [J]. Information Science, 2016, 34(01): 15-19+33.

[15] Zhang Shaocheng, Sun Shiguang, Qu Yang, et al. Application research of machine learning in data mining under big data environment [J]. Journal of Liaoning University (Natural Science Edition), 2017, 44(1): 15-17.

[16] Dai Huili. Research on the application of machine learning in data mining under the background of big data [J]. Journal of Luliang Education College, 2019, 36(3): 20-21.

[17] Zhou Xu. Application of machine learning in data mining [J]. Electronic Technology and Software Engineering, 2019(7): 173.