Research on Data Classification Algorithm in Big Data Mining

Liu Weigang
Tianjin Modern Vocational Technology College

Keywords: Big data, Mining, Data classification, Algorithmic technology.

Abstract: With the advancement and development of data mining technology research, along with the concept of big data, it will bring a great opportunity for the development and application of data mining technology. Data mining technology will enter a new period of development. This paper introduces the basic concepts of big data, data mining and analysis methods of data mining. Finally, the application fields and future development directions of mining technologies in the era of big data are presented. If you want to do a good job of data processing, you must apply it to the data classification algorithm. As a key technology in data mining, it can do a good job of data processing.

1. Introduction

In the current era background, many industries have introduced the concept of big data mining, which not only brings development opportunities to the computer industry, but also brings challenges. Because you want to do related work of big data mining, you must master the data classification algorithm, and the data classification algorithm can be regarded as a difficulty in data mining. With the continuous deepening of data analysis, people have developed a variety of classification algorithms to continuously reduce their difficulty.

2. Data Mining in the Age of Big Data

Data Mining In the era of big data, the generation and collection of data is the foundation, and data mining is the key. Data mining is the most critical and valuable work in big data. In general, data mining or knowledge discovery refers broadly to the engineering and systematic process of mining implied, previously unknown but potentially useful information and patterns from large amounts of data. Data mining can be summarized by the following four characteristics.

2.1 Application of data mining

Data mining is a perfect combination of theoretical algorithms and application practices. Data mining originates from the needs of actual production and life applications. Data mining comes from specific applications. At the same time, knowledge discovered through data mining must be applied to practice to assist in actual decision-making. Therefore, data mining comes from application practice and also serves application practice.

2.2 The engineering nature of data mining

Data mining is an engineering process that consists of multiple steps. The application characteristics of data mining determine that data mining is not only an algorithm analysis and application, but a complete process including data preparation and management, data preprocessing and transformation, mining algorithm development and application, result display and verification, and knowledge accumulation and use. And in practical applications, the typical data mining process is an interactive and circular process.

2.3 Collection of data mining

Data mining is a collection of multiple functions. The commonly used data mining functions include data exploration analysis, association rule mining, time series pattern mining, classification
prediction, cluster analysis, anomaly detection, data visualization, and link analysis. A specific application case often involves a number of different functions. Different functions usually have different theoretical and technical foundations, and each function has a different algorithm support.

2.4 The cross-cutting of data mining

Data mining is a cross-cutting discipline that uses research results and academic ideas from many different fields such as statistical analysis, pattern recognition, machine learning, artificial intelligence, information retrieval, and databases. At the same time, some other fields such as stochastic algorithms, information theory, visualization, distributed computing, and optimization also play an important role in the development of data mining.

3. Data mining in the era of big data

3.1 Principles of data mining

Support multiple mining tasks In Pd Miner, not only support the creation and execution of a single task, but also support the simultaneous creation and operation of multiple data mining tasks. These tasks can be different categories of mining tasks, such as parallel association rule tasks, parallel classification, and clustering tasks. When parameters are configured, these tasks can be performed simultaneously in the parallel distributed system Pd Miner. Supporting multiple mining tasks has a very important role. For example, compare all the classification algorithms to select the best algorithm for the existing data set. The general practice is to serially test all the algorithms and then make the selection based on the effect of the algorithm. In Pd Miner, this problem can be solved in parallel. All algorithms are oriented to the same data (reading the same header file information). The final result is displayed through the system to select the most suitable algorithm. From this comparison mechanism, we see that all parallel algorithms are executed in a parallel system and can therefore handle large-scale data; in addition, the execution of these algorithms is parallel and the evaluation process is automatic, so it can reduce the algorithm execution time and User intervention.

To create complex mining tasks, parallel data prepossessing operations and parallel data mining algorithms can be cascaded. The system provides parallel attribute deletion operations, parallel data normalization, and parallel classification algorithms. After all the algorithm parameters are configured, the execution process is as follows: the property deletion operation is performed, the property deletion operation is performed on the data set, and the header file is modified to generate a new header file information. The updated header file is received after the attribute is deleted, and the data normalization operation is performed. Perform classification algorithm tasks. Receive the header file information passed from the second step and start the classification algorithm task. After the task is completed, the classification results are displayed.

![Fig. 1 Data Mining Mode in Big Data Era](image-url)
3.2 Application scope of data mining

From the perspective of data mining applications, big data big data is a phenomenon. The core is to mine the value of data. Combining various features of data mining, especially its application, we propose the following two points of understanding of big data from the perspective of application services. Big data is in an enterprise. Big data usually involves multiple business departments and the business logic is complex. On the one hand, to collect and integrate big data requires the cooperation and communication of business departments and the vigorous participation of business personnel. These require the attention and recognition of corporate decision makers and provide the necessary resources for deployment and support. On the other hand, verifying and using the results of data mining can not be separated from the decisions of relevant personnel. The results of data mining are mostly related, not causal. These results may also have uncertainties. In addition, sometimes the results of data mining are inconsistent with the common sense of business operations, and even the opposite. Therefore, how to look at the analysis results of these possible uncertainties and abnormal knowledge and make full use of the data mining results will inevitably depend on the vision and insight of the decision makers. Big data requires data import, integration, and preprocessing. When faced with a large amount of complex data from different data sources, the trivial relationship between the specific business logic complexity and the data directly leads to difficult understanding of the business process and data flow of the enterprise. When companies implement big data, they may not be clear about what they want to discover and discover. There is no intuitive and clear understanding of what data mining can do to help companies do what they do. In many cases, it is impossible to plan and prepare data in advance. In this way, specific data mining requires great flexibility in data import, integration, and preprocessing, and only through business personnel and data mining engineers. The cooperation and continuous trials can effectively link the business needs of the enterprise with the functions of data mining.

Fig. 2 Mining Data Analysis in the Age of Big Data

3.3 Creating a complex mining process

Through the work flow subsystem, the system also supports the creation of complex mining tasks, and parallel data preprocessing operations and parallel data mining algorithms can be cascaded. The system provides parallel attribute deletion operations, parallel data normalization, and parallel classification algorithm naïve Bayes concatenation. After all the algorithm parameters are configured, the attribute deletion operation is performed, the attribute deletion operation is performed on the data set, and the header file is modified to generate a new header file information. The updated header file is received after the attribute is deleted, and the data normalization operation is performed. Perform classification algorithm tasks. Receive the header file information passed from the second step and
start the classification algorithm task. After the task is completed, the classification results are displayed. Parallel distributed data mining platform Pd Miner is an efficient data processing and analysis tool, mainly for the processing of massive data sets. To ensure the accuracy of the algorithm, construct large data sets to examine the performance of the algorithm. The parallel algorithm developed in the system has been applied in the actual data mining in the communication field. The performance test results of some algorithms on the constructed large data set are given below. Due to privacy and other reasons, no specific parallel algorithm names are given here. The value chain view Big data system is a complex system that provides data processing functions at different stages of the data life cycle (from data generation to extinction). At the same time, big data systems often involve multiple different stages for different applications.

3.4 Big data mining algorithm research

The widely accepted system engineering method in the industry has broken down a typical big data system into four consecutive phases, including data generation, data acquisition, data storage, and data analysis. The data generation phase is concerned with how data is generated. At this point "big data" means a large, diverse, and complex set of data generated from a variety of vertical or distributed data sources (sensors, video, click streams, and other digital sources). Usually, these data sets are related to The different levels of value are tied together. It will focus on three important areas of business, the Internet, and scientific research, because the value of data in these areas is relatively easy to understand. However, there are huge technical challenges in collecting, processing, and analyzing these data sets. New solutions need to be developed using the latest research technologies in the field of information and communication technology (ICT). Data acquisition refers to the process of obtaining information, which can be divided into Data collection, data transmission and data preprocessing. The data comes from different data sources, such as website data that contains formatted text, images, and video. Data acquisition refers to dedicated data collection technology that obtains raw data from a specific data production environment. After data collection is completed, a high-speed data transmission mechanism is required to transfer the data to a suitable storage system for use by different types of analysis applications. There may be some meaningless data in the data that will increase data storage space and affect subsequent data analysis. The data set obtained from sensors monitoring the environment is usually redundant, and data compression techniques can be used to reduce the amount of data transmission. Therefore, data must be predecessor to achieve efficient data storage and mining. Data storage addresses the long-term storage and management of large-scale data. The data storage system can be divided into two parts, hardware infrastructure and data management software. The hardware infrastructure consists of shared ICT resource pools that are organized in an elastic manner based on the immediate needs of different applications. The hardware infrastructure should be able to scale up and out and be dynamically reconfigured to accommodate different types of application environments.

Fig. 3 Data Analysis Mode in the Big Data Era

Data management software is deployed on the hardware infrastructure to maintain large-scale data
sets. In addition, in order to analyze the stored data and its data interactions, the storage system should provide functional interfaces, quick queries, and other programming models. Data analysis uses analysis methods or tools to examine, transform, and model data and extract value from it. Many application areas use domain-relevant data analysis methods to achieve the desired results. Although different areas have different requirements and data characteristics, they can use some similar underlying technologies. The current data analysis technology research can be divided into six important directions: structured data analysis, text data analysis, multimedia data analysis, web data analysis, network data analysis, and mobile data analysis. Corresponding open source or proprietary technologies are linked together. Reflects the development trend of big data. In the data generation phase, the structure of big data is gradually complicated, from structured or unstructured data to different types of mixed data. At the data acquisition stage, research on data acquisition, data prepossessing, and data transmission occurs at different times.

4. Data Mining Applications in the Big Data Era

4.1 Big data mining applications in marketing

Marketing is the earliest and most extensive application of data mining technology. Through the analysis and mining of user's consumption habits and consumption characteristics, to improve the sales of goods. At present, data mining has not only been limited to supermarket shopping in terms of marketing, it has been widely used in various financial fields such as insurance, e-commerce, banking, and telecommunication retail. The use of data mining technology to analyze customer's consumer behavior brings potential customers and benefits to the industry.

4.2 Scientific research application of big data mining in scientific research

Often need to analyze a variety of experimental and observational data, and to find out the relevant laws and knowledge. These data analysis and mining all require certain algorithms. Using data mining techniques can scientifically find out the laws between data and find out what we have not found. For example, exploration of outer space objects, analysis of DNA data, and so on.

4.3 Manufacturing big data mining applications in the manufacturing industry

Through the analysis of various data of parts and components production, to improve production efficiency and increase the yield rate, such as analyzing the links and data that cause product defects, find out the factors that affect the productivity of the production process, expose the changes in the manufacturing and assembly operations, etc. These factors, in order to improve the company's production efficiency and interests, by focusing on improving related links and processes.

4.4 Big data mining applications in the telecommunications industry

With the rapid development of the telecommunications industry, the telecommunications industry's technology and services have become a very large hybrid carrier. It not only relates to the market, but also relates to technology and services. In the process of integrating enterprise's effective resources and achieving seamless connection, a large number of complex data sets will be generated. How to find out the laws and knowledge from the complicated data has become an important issue in the telecommunications industry. Through the use of data mining, it can solve the above problems well and play a good role in the development of the industry, such as system load, data communication rate and capacity, user behavior, profit rate, etc., using isolated point analysis and clustering methods. , Find out the misappropriation patterns and abnormal status; use correlation and sequence patterns to identify relevant factors that affect the development of telecommunications and promote the development of telecommunication services.
4.5 Big data mining application in educational management in education

At present, there are many applications in this area, such as search engines and e-commerce. Through data mining technology, we can find information that meets the needs of users in massive data, such as predicting classification algorithms to predict information that users may need to search. How to fully grasp the student's learning status, psychological status, course selection and teaching evaluation and other information and the optimal allocation of teaching resources, can use data mining technology to solve. With the data management of big data and the progress of retrieval technology research, data mining technology will usher in huge development opportunities. The application of data mining technology will also be more extensive, and the tools for data mining will also be more powerful.

5. Conclusion

Big data mining is the trend of the times, so the importance of data classification algorithms will also appear. By analyzing several different applications, you can compare data analysis speed, capability, and accuracy of results to select the most appropriate area for big data applications. They all have their own advantages and disadvantages in varying degrees, so we must continue to further study to develop a better classification algorithm.

References


