The Application of Intelligent Agricultural Big Data Platform on the Internet

Guoning Lv(✉)

Zhengzhou Normal University, Zhengzhou 450000, China
pppp112389@163.com

Abstract. In order to improve the ability of value mining in agricultural information management, control and decision-making, to meet the actual needs of agricultural big data. On this basis, the application of agricultural big data and its architecture are discussed in depth. In order to ensure the high availability, reliability, expansibility and security of the system, this paper proposes a method based on load balancing strategy, file block partitioning, parallel processing optimization and fault detection and recovery. The case analysis shows that big data technology plays a very positive role in improving the development level of intelligent agriculture.

Keywords: intelligent agriculture · Agricultural big data · Big data platform · Distributed cluster · Key technology

1 Introduction

With the rapid development of the Internet of Things, cloud computing, Internet, mobile communication and other technologies, big data technology has become the main driving force for the development of smart agriculture. Agricultural big data is a practical application of big data theory, technology and method in agricultural production. The data sources of intelligent agriculture are very rich, covering data of different fields, industries and professions, which is characterized by large scale, dispersion, diverse types and complex structure. The core work of smart agriculture big data is to extract value from a large number of real-time and complex data and create wisdom through the use of big data technology, so as to make agricultural management, control, prediction and decision-making more "intelligent", so as to improve the development speed and quality of agricultural intelligence.

At present, our application in the field of agricultural big data is still in the initial stage, and the results of practical research are few, and the system is not perfect. There are not many integrated solutions for existing agricultural big data systems, and most of them are designed for large-scale farmers. There are few independent, unified and universal agricultural big data platforms. Secondly, some agriculture-related enterprises have a large amount of real-time and networked data, so it is urgent to use big data technology to solve problems in practical application management. However, due to high technology content, independent development cannot be realized. In addition, agriculture-related

N. Akhtar et al. (Eds.): PMIS 2023, AHIS 8, pp. 412–417, 2023.
https://doi.org/10.2991/978-94-6463-200-2_43

enterprises in the construction of professional IT institutions, often according to their own personal needs for individual customization, but because of the large investment, high cost of customization, it is difficult to be accepted by the majority of customers. Therefore, it is necessary to study the overall solution of agricultural big data, and build an overall solution of agricultural big data that serves the needs of multiple agricultural fields, addresses the customized needs of users, ADAPTS to the expansion and change requirements, protects users' privacy security, information islands, privacy security and other issues. The smart agriculture big data platform shall be built to serve the needs of multiple agricultural fields, meet the needs of users' personalized customization, adapt to the requirements of expansion and change, protect users' privacy security, and share and separate data resources.

2 Classification ID3 Algorithm

The first step in data mining is to prepare the data. In data mining, the data suitable for data mining is selected and pre-processed, such as denoising and weight elimination, among which the data selection and pre-processing is the key of data mining. Finally, converting data into data model and establishing effective data model is the prerequisite of data mining. And then according to the selected algorithm, mining.

A sample set of data with a sample number is defined as sample X, where the sample number is x, the sample belongs to class Y, assuming there are k samples, then the classification set is defined as Y = {y1, y2… Yk-1, yk}, where the data sampling set X is divided into k sampling subsets, then X = {X1, X2… Xk minus 1, Xk}. The information entropy of the sample set can be expressed in the following ways:

$$E(X) = -\sum_{i=1}^{K} u_i \log_2 u_i \tag{1}$$

where, the information entropy of the sample set is also called average uncertainty,

Also known as "prior entropy", the meaning expressed by it is the information contained in data set X, and the selected category attributes are determined by the information contained in it. Suppose a sample set X is a set with L distinct attribute values {h1, h2… Hl-1, hl} properties. Here, xij is used as Xj (j = 1, 2… k) in class Yi (i = 1, 2… k), and express uij as the probability of having class yi. The so-called information gain is to determine the selection of attributes according to the rate of information entropy reduction. On the basis of ID3 classification method, the attribute with the highest information gain is obtained by using the calculated information entropy, and it is taken as the classification attribute of nodes. Selecting this attribute as the classification attribute can obtain the maximum information and the least uncertainty.

3 Intelligent Agriculture Under Classification Algorithm

A data set consists of data objects. For example, in a bank database, the object could be a customer, a wealth management product or a sales channel; In a medical database, objects can be patients, patients, cases, and so on. The following table gives a group of data to determine whether it is suitable for planting, in which the weather is an object, and its characteristics show attributes (Table 1).

Table 1. Agricultural data sets

Serial number	weather	temperature	humidity	Strong wind	Whether it is suitable for planting
1	sunny	high	strong	There is no	no
2	sunny	high	strong	There are	no
3	cloudy	high	strong	There is no	is
4	cloudy	suitable	strong	There is no	is
5	cloudy	low	moderate	There is no	is
6	cloudy	low	moderate	There are	is
7	cloudy	low	moderate	There are	no

For the data set given in the table, classification attributes are selected, and then a decision tree is established to determine whether it is suitable for planting. Firstly, information entropy can be calculated according to the formula.

$$E(X) = -\sum_{i=1}^{k} u_i \log_2 u_i = 0.94 \tag{2}$$

The object, the information entropy of the weather, is.

$$E(X) = -\sum_{i=1}^{k} u_i \log_2 u_i = 0.694 \tag{3}$$

Then the information gain of the weather can be written as

$$G\left(\frac{X}{H}\right) = E(X) - E(H) = 0.246 \tag{4}$$

Under the same available to temperature, humidity, wind information gain value G/temperature (X) = 0.029, (X/humidity) = 0.151 G, G (X/wind).

To sum up, according to the principle of ID3 method, we compare the information gain values and choose weather as the classification attribute. The basic steps of algorithm implementation are given below:

1) Create an initial root node. Then the decision is made. If the samples are in the same class, the algorithm ends and the node is marked as a leaf node. Otherwise, the attribute with the maximum information gain is selected according to the algorithm, and the classification attribute is selected for the next classification.
2) Divide samples, compare each value in the classification attribute according to the attribute value, and extend a corresponding branch.
3) Repeat the loop command, calling the above steps from top to bottom until the conditions are met, then jump out of the loop. The resulting pseudocode is as follows:

ID3(samples, attributes)

//samples the training set, attributes the set of candidate attributes
{ifsamples are nullthen
Returnnull;
Establish node N;
The same class of elssIf sample node is YthenReturnN as leaf node and marked as class Y. The Ifattributes are nullthen
ReturnN as leaf nodes, and marked as samples often class Elssforeachattributeinattributesdo {attribute information gain, and choose the maximum attribute of information gain, as the classification properties as the test - the attribute;} bi,do{attribute value in Foreachtest-attribute = bi, generates corresponding branch from node N according to test-attribute = bi, indicating test conditions; Let Xj be the node returned by sample subset Then when test-attribute = bi (Xj, attributes-test-attribute)}}.

4 Construction of Smart Agriculture Big Data Platform

4.1 System Architecture

According to the business characteristics of intelligent agricultural informatization, in order to meet the requirements of multi-source heterogeneous data storage and processing under various scenarios, and solve the defects of the existing data storage, computing and processing system, the unified centralized storage and management mode of data center is adopted, and a large number of cluster mode is adopted for data storage and processing. The problems of data redundancy, resource utilization, sharing and maintenance cost are solved. This paper proposes a mixed data structure from the bottom layer to the top layer of data exchange layer, data storage layer, data processing layer and resource management layer. Multiple distributed storage technologies are integrated to build a large-scale, multi-level, consistent and transparent data storage and management mode. It integrates computing engines such as parallel computing, memory computing, batch processing, and stream processing to optimize big data processing and ensure high availability, scalability, and reliability of the system by utilizing integrated centralized resources.

4.2 Data Exchange System

The data exchange layer is composed of the data acquisition layer and the value representation layer. The data exchange interface and protocol such as Webservice are used to realize the interconnection between the external system and the big data platform, so as to realize the interactive access of data. In the acquisition layer, the distributed data acquisition layer is adopted to carry out data extraction, conversion and loading through the streamline and parallel way, so that the multi-source heterogeneous data can be quickly guided. Support to extract data from data sources, such as text, table, image, XML file, and active push and passive pull two data transfer modes for data storage and processing, can be defined according to the need for cleaning, encoding, distribution and conversion preprocessing, and can be dynamically expanded to increase the data throughput rate, and can set the reliability level according to the performance requirements.

4.3 Data Storage

Big data storage layer adopts distributed external storage, structured/semi-structured/unstructured database, distributed storage three-tier storage structure, including distributed file system, relational database, NoSQL highly parallel database, memory database, etc. Distributed file system can store text, audio and video files directly in intelligent agricultural system, providing highly reliable and extensible file storage. Associated databases store consistent, structured business data; NoSQL database is mainly used to store data, such as historical log data, meteorological data, social and economic data. The access performance and scalability of these data are mainly considered. Storage databases can store large amounts of data that need to be processed quickly, such as indexes, intermediate results, dimension tables, etc.

4.4 Data Processing System

At the level of big data processing, distributed computing, batch processing, graph computing, interactive analysis, stream processing and other modules are used. Memory operation provides a distributed memory abstraction machine for heterogeneous memory, realizes data caching and improves the performance of I/O. Batch processing technology is mainly aimed at data-intensive offline parallel computing, such as classification, clustering, association rules, etc. Graphic operation to process the structure of the graph, such as agricultural product traceability system, e-commerce logistics platform, etc. Interactive procedures are used for quick responses to SQL requests, such as queries, aggregations, associations, and so on. Stream processing engine is mainly used for real-time and continuous stream data query statistics, cleaning conversion, abnormal alarm and so on. The algorithm is based on overloaded types of specific program modules, such as MPI and OpenMP, to meet the requirements of tightly coupled, iterative computation.

5 Summary

In data mining, classification algorithm is a very important work, it is more suitable for the classification of mass data. Its operation method is simple and easy to understand, so it is widely used. This paper takes big data in agriculture as an example to study data mining based on classification, but on this basis, this paper only classifies the data theoretically, and in the early data selection process still needs manual operation, which needs to be tested in practical application.

Acknowledgement. Fund Project: Key Scientific Research Project of Colleges and Universities of Henan Province, Project Number: 23B520025.

References

1. PORTERBW, BAREISSR, HOLTERC. Concept learning and heuristic classification inweak-theory domains [J]. Artificial Intelligence, 199 Do (1): 229–263.
2. QUINLANJR.Discoveringrulesbyinductionformlargecollectionsofexamples[J]. InExpertSysteminthe Micro Electronc Age, 1979, (1): 26 to 37.
3. Li Zhigang, Ma Gang. Principle and Application of Data Warehouse and Data Mining [M]. Beijing: Higher Education Press, 2008, 2.
4. Fan Ming, Meng Xiaofeng. Concept and Technology of Data Mining [M]. Beijing: China Machine Press, 2012.
5. Zhang Lin, Chen Yan, Li Tao-ying. Research on Decision Tree Classification Algorithm [J]. Computer Engineering,2011,13(13):26-27. (in Chinese)
6. VAESE, MANCHANDAR, NIRR.etal. Mathematicalmodelstodiiscriminatebetweenbenignandmalignantadnexalmasses[J].InternationalJo UrnalofGynecologicalCancer2011, 21 (1): 35–43.
7. ZHU Huan-dong. Improvement and Simplification of ID3 Algorithm [J]. Journal of Shanghai Jiaotong University,2010,44(7):883-886.