C4.5 Crossover Algorithm Based on MapReduce in University Big Data Analysis

Yang Liu*

Henan University of Radio and Television Finance and Finance College, Zhengzhou, China

*Corresponding author e-mail: statsyangliu@163.com

Abstract: With the continuous development of digitization and informationization, students' data generated by various kinds of application systems in colleges and universities is also accelerating. A big data environment of colleges and universities has been formed. The data contains a wealth of information that requires new storage and analysis tools for processing. Therefore, in this paper, big data processing platform Hadoop is used to analyze college big data, and C4.5 cross-improvement algorithm is proposed based on the low efficiency and over-fitting of decision tree C4.5 algorithm, and then verified in student grants data. The feasibility of the algorithm. Finally, the validity of the improved algorithm is verified from different evaluation criteria. The experimental results show the effectiveness and feasibility of the improved algorithm proposed in this paper, which provides a new idea and method for big data analysis in colleges and universities.

1 Preface

With the continuous advancement of information technology construction in colleges and universities, the amount of data accumulated by colleges and universities is constantly expanding. The campus big data environment has been formed initially. The analysis and mining of campus big data can provide direction and basis for decision-making for school management departments. It is conducive to the improvement of university management and service level. However, due to the large volume and unstructured nature of campus big data, traditional data mining algorithms generally have problems such as poor scalability and low utilization rate. Therefore, more and more scholars are now trying to use the Mapreduce parallel computing model to process large-scale data sets. Qing [1] and other scholars use Mapreduce parallel computing model in the data processing and data cleaning phase to greatly improve the efficiency of operation, reduce resource consumption, and provide a new idea of data processing. Zhang Chunyan [2] introduced the parallel SPRINT algorithm based on the decision tree algorithm. The results show that the algorithm has better acceleration performance. Korhi [3] and others introduced parallel algorithms in data preprocessing, and preprocessed, trained and predicted the text, and the prediction accuracy was greatly improved. Zhao Weizhong [4] proposed an improved K-means algorithm based on the Mapreduce parallel model and applied it to the data of different quantity sets. The results show that the parallel model can greatly improve the scalability and operating efficiency of the model. This paper combines the characteristics of university big data data diversification and large data size, and gives a decision tree model based on Hadoop parallel processing framework, and proposes an improved decision tree cross-blocking algorithm. The result also proves this The effectiveness of the algorithm.
2 C4.5 decision tree model based on Hadoop parallel processing framework

2.1 Decision Tree Correlation Algorithm

The Decision Tree algorithm is essentially an inductive learning algorithm that derives classification rules from a set of irregular data sets. On the internal nodes of the decision tree is a branch attribute that represents the data instance on this attribute. The test is used, and the leaf node is used to indicate that the output of the test represents a classification conclusion. The decision tree first trains the data set to get the tree structure, and finally uses the model to classify and predict the sample.

2.1.1. Information Entropy Related Concepts

Information entropy: reflects the uncertainty of the data set. The larger the information entropy, the more uncertain the data set is.

\[
H(X) = - \sum_{i=1}^{n} p(a_i) \log p(a_i)
\]

Where: \( N \) is the number of data sets, which is the possible data value, which is the probability value obtained.

1) Conditional entropy: Conditional entropy is used to measure the uncertainty of random variable \( X \) under the condition that \( Y \) is determined to be independent of random variable \( X \) and random variable \( Y \).

\[
H(X|Y) = \sum_{j=1}^{k} p(b_j) H(X|b_j)
\]

Where: the data value of the random variable \( X \), the data value of the random variable \( Y \), and the probability that \( Y \) is \( X \).

2) Average mutual information: indicates the size of the information amount of the information variable \( X \) that the random variable \( Y \) can provide.

\[
I(X|Y) = H(X) - H(X|Y)
\]

3) Information Disciplinary: Indicates the cost of obtaining the data value of a random variable.

\[
H(X,a) = - \sum_{i=1}^{n} p(a_i) \log p(a_i)
\]

4) Information gain rate: is a relative measure of information uncertainty

\[
E(X,\text{consume}) = \frac{I(X,\text{consume})}{H(X,\text{consume})}
\]

2.1.2. C4.5 related algorithm

The C4.5 algorithm is based on the traditional decision tree algorithm. The information gain rate is used to replace the information gain to filter the data set. In the process of selecting features, the attribute with the highest information gain rate is always used as the best choice, and the algorithm is added. For continuous data, the processing of missing data, and finally pruning the decision tree [5].

2.1.3. Student Scholarship C4.5 Algorithm under MapReduce

Table 1 shows some of the student information for the 2016 bursary of an undergraduate college in Zhengzhou, Henan Province. The student data has been desensitized. Each data sample has 4 attribute variables, consumption, part-time off-campus, loan, and grade. The data sets are ultimately divided into two categories according to whether they receive bursaries: yes and no.
Table 1. Partial student bursary information training set

| Attributes | consume | external | whether | grades | grant |
|------------|---------|----------|---------|--------|-------|
| 1          | High    | N        | N       | B      | N     |
| 2          | Medium  | N        | N       | B      | N     |
| 3          | Medium  | N        | Y       | A      | Y     |
| 4          | Low     | N        | Y       | B      | Y     |
| 5          | High    | N        | N       | C      | N     |
| 6          | Medium  | N        | N       | B      | N     |
| 7          | Medium  | N        | N       | C      | N     |
| 8          | Low     | Y        | Y       | A      | Y     |
| 9          | Medium  | N        | N       | C      | N     |
| 10         | Medium  | N        | N       | A      | N     |
| 11         | High    | N        | N       | A      | N     |
| 12         | Low     | Y        | Y       | B      | Y     |
| 13         | Low     | Y        | Y       | C      | Y     |
| 14         | Medium  | N        | N       | B      | N     |
| 15         | High    | N        | N       | A      | N     |
| 16         | Medium  | N        | N       | B      | N     |
| 17         | Low     | Y        | Y       | B      | Y     |
| 18         | Medium  | N        | N       | B      | N     |
| 19         | Medium  | N        | N       | A      | N     |
| 20         | High    | N        | N       | B      | N     |
| 21         | Low     | N        | Y       | C      | N     |
| 22         | Medium  | N        | N       | A      | N     |
| 23         | Medium  | N        | N       | C      | N     |
| 24         | High    | N        | N       | B      | N     |

First calculate the information entropy $H(X)$. The information entropy value obtained from the previous discussion is:

$$H(X) = -\frac{5}{24} \log \frac{5}{24} - \frac{19}{24} \log \frac{19}{24} = 0.738$$

Then calculate the information gain ratio of each attribute.

The entropy of the consumption attribute is:

$$H(X_{\text{Consume}}) = -\frac{6}{24} \log \frac{6}{24} - \frac{12}{24} \log \frac{12}{24} = 1.5$$

The conditional entropy is:

$$H(X | \text{Consume}) = -\frac{6}{24} \log \frac{6}{24}$$

Therefore, the information gain ratio of the consumption attribute is:

$$\frac{I(X_{\text{Consume}})}{H(X_{\text{Consume}})} = 0.20$$

Similarly, the off-campus part-time job can be calculated, and the information gain ratios of the scores are 0.56 and 0.03, respectively. Since the loan information gain ratio is the largest, this attribute is selected as the division attribute. Recursive decision tree according to the above method.
From the constructed decision tree model, the following conclusions can be drawn: students who receive the bursary need to apply for a loan at the time of enrollment, and the application for the loan needs certain conditions, such as submitting a poverty certificate and a family situation questionnaire, so that the student loan can be obtained to some extent, it reflects the poverty situation of the family. In applying for student loans, the better the score, the easier it is to get the bursary, which means that the score is also a big reference factor. Finally, according to the consumption data of the canteen card, the students are divided into poverty. In summary, the C4.5 classification algorithm can provide a theoretical basis for the assessment of grants.

3. **MapReduce C4.5 cross-blocking algorithm based on Hadoop framework**

In view of the characteristics of college big data sets discussed above, it can be seen that traditional algorithms have the disadvantages of low operating efficiency and excessive memory usage in large-scale data sets and unstructured data sets. This paper proposes a C4.5 cross-blocking algorithm, which divides the data set from the sample and attribute, respectively, to reduce the frequent scanning and pre-reading of the data set in the calculation process, improve the operation efficiency and reduce the memory occupation. [6].

3.1. **C4.5 cross-blocking algorithm flow**

**Input:** training sample data set  
**Output:** Decision Tree Model

**Step1:** Perform horizontal segmentation on the training sample data set, and divide into m identical data blocks and send them to the map operation with m nodes;  
**Step2:** Let's talk about the vertical division of each attribute on each node to form a vertical data block;  
**Step3:** Perform secondary level division for each vertical data block  
**Step4:** The divided data set is used as input to participate in the operation;  
**Step 5:** Build a root node, and form an attribute division queue according to the division attribute;  
**Step6:** Calculate the woman's gain in the reduce, and select the attribute with the largest information gain as the candidate attribute;  
**Step7:** Insert the test attribute into the queue;  
**Step8:** Delete the test sentence to form a new queue;  
**Step9:** Recursively proceed to Step 6 to Step 8;  
**Step10:** Over.

Data partitioning in MapReduce is shown below:
Attributes consume external part-time jobs whether loan grades grant
1 High N N B N
2 Medium N N B N
3 Medium N Y A Y
4 Low N Y B Y
5 High N N C N
6 Medium N N B N
7 Medium N N C N
8 Low Y Y A Y
9 Medium N N C N
10 Medium N N A N
11 High N N A N
12 Low Y Y B Y
13 Low Y Y C Y
14 Medium N N B N
15 High N N A N
16 Medium N N B N
17 Low Y Y B Y
18 Medium N N B N
19 Medium N N A N
20 High N N B N
21 Low Y Y C Y
22 Medium N N A N
23 Medium N N C N
24 High N N B N

Attributes consume external part-time jobs whether loan grades grant
1 High N N B N
2 Medium N N B N
3 Medium N Y A Y
4 Low N Y B Y
5 High N N C N
6 Medium N N B N
7 Medium N N C N
8 Low Y Y A Y
9 Medium N N C N
10 Medium N N A N
11 High N N A N
12 Low Y Y B Y
13 Low Y Y C Y
14 Medium N N B N
15 High N N A N
16 Medium N N B N
17 Low Y Y B Y
18 Medium N N B N
19 Medium N N A N
20 High N N B N
21 Low N Y C N
22 Medium N N A N
23 Medium N N C N
24 High N N B N

Figure 2. The process of crossover block algorithm
First divided by the graph, the algorithm is to block processing samples, and bring them as the original input graphs, then on the second vertical division, will be divided into data subset as a decision tree node, on the basis of the third level, this step is the same sample values are divided into a set of attributes. Through this step, large data sets can be divided horizontally and vertically, which can effectively reduce the number of data scans and improve the efficiency of parallel computing [7-9].

### 3.2. Experimental results and analysis

In order to further analyze and verify the effectiveness and feasibility of the proposed C4.5 cross-partitioning algorithm, different levels of university grants data sets were selected as experimental objects. The experimental results are shown as follows.

![Figure 3. Acceleration ratio of MapReduceC4.5 cross-blocking algorithm based on Hadoop platform under different cluster numbers](image)

| The data set | 1node | 2node | 3node | 4node | 5node | 6node | 7node |
|--------------|-------|-------|-------|-------|-------|-------|-------|
| 1G data set  | 1.0   | 1.7   | 2.7   | 3.5   | 4.0   | 4.9   | 5.4   |
| 2G data set  | 1.9   | 3.0   | 3.8   | 4.3   | 4.7   | 5.4   | 6.0   |
| 4G data set  | 2.1   | 3.4   | 4.2   | 4.7   | 5.4   | 5.9   | 6.0   |
| 8G data set  | 2.3   | 3.4   | 4.3   | 4.9   | 5.6   | 6.2   |

Table 2 shows the acceleration ratio of MapReduce C4.5 parallel algorithm based on Hadoop platform on different cluster nodes. It can be seen from the table that the acceleration ratio of the algorithm increases linearly with the increase of the number of cluster nodes. Figure 2-3 chooses data sets of different sizes to run in different numbers of clusters. The running data sets are 1G, 2G, 3G, 4G, 5G, 6G, 7G, 8G, and the cluster number is 1, 2, 3 and 4. The results show that the running time is effectively shortened. Number of clusters. Therefore, the improved algorithm proposed in this paper can effectively improve the high time complexity caused by large data sets and unstructured data, and verify the high scalability and good parallelism of the algorithm.

### References

[1] Qing. Performance evaluation of MapReduce using full virtualisation on a departmental cloud[J]. International Journal of Applied Mathematics and Computer Science, 2011, (10): 275-284.

[2] Korhi. Fitting the Problem to the Paradigm: Algorithm Characteristics Required for Effective Use of MapReduce[J]. Procedia Computer Science, 2012, (7): 212-217.

[3] Wulf-Peter, Schmidt. Life Cycle costing as part of design for environment environmental business cases[J]. The International Journal of Life Cycle Assessment, 2013, (3): 167-174.

[4] Zhang chunyan. Big data analysis -- competition and symbiosis between RDBMS and MapReduce [J]. Journal of software, 2012, (1): 14-19.

[5] automatic modeling of binary targets in data mining (algorithm) [J]. Ding xueping. Popular science and technology. 2013(10).

[6] application of decision tree analysis in the analysis of factors influencing life pressure of college students [J]. Chen xinlin, bao sheng geng, yan weihong, wang xiaoguang, wan jiancheng, wu dangui. Health soft science, 2013(02).
[7] Public: A decision tree classifier that integrates building and pruning, Rajeev Rastogi, Kyuseok Shim. Proceedings of the 24th International Conference on Very Large Databases (VLDB). 1998.
[8] World Health Report 2003-Shaping the future. World Health Organization.
[9] Application of data mining method based on decision tree in fault diagnosis [J]. Shi jinyan, li minchen, haiyan. Water conservancy and electric machinery. 2006(04).