ID3 algorithm and its improved algorithm in agricultural planting decision

Guangping Qiang¹, Lin Sun² and Qiang Huang ¹*
College of Information Engineering, Sichuan Agricultural University, Ya'an City, Sichuan Province, 625000, China
*Corresponding author’s e-mail: 19623556@qq.com

Abstract. With the advent of the era of 5G big data, it is not only necessary to mine valid data in mass information, but also to classify it. Mining and classifying information can be used as an important basis for decision making. Making correct and efficient decisions on agricultural planting can not only improve agricultural production efficiency, but also lay the foundation for the realization of smart agriculture. This paper will compare the improved algorithm, which is called MIND, with the algorithms in other literatures by experiments. Through calculations in agricultural big data such as temperature, humidity, wind, and weather, it is demonstrated that the improved algorithm is more suitable for agricultural planting decisions.

1. introduction
With the advent of the era of 5G big data, big data and 5G have attracted worldwide attention and research. China is a big agricultural country, and it is even more important to make good use of big agricultural data. However, agricultural big data not only has strong randomness, but also is accompanied by noise. If these data cannot be effectively processed, it will become garbage data. Therefore, it is necessary to effectively classify and mine agricultural big data. Through the practical application of agricultural planting decisions, not only can it save a lot of manpower, material resources and financial resources, but also lay the foundation for the realization of smart agriculture. The classification ID3 algorithm is the most classic classification algorithm in the decision tree algorithm. This algorithm uses information gain as the criterion for attribute division. The generated decision tree has a simple structure and is easy to understand. [1-2] However, the ID3 algorithm tends to select attributes with many attribute values as split nodes, and it is inconvenient to process continuous data.

In order to solve the problem of ID3 algorithm, a large number of scholars have also made a lot of improvements to the ID3 algorithm. Reference [3] improved the information entropy formula by using the power-level expansion of the function. Aiming at the shortcoming of the multi-value bias of the ID3 algorithm, the Reference [4] multiplies the information entropy of each attribute by the dynamic attribute weight. It makes the result of information entropy less dependent on the number of attribute values, and avoids making decisions that are not suitable for the actual situation due to the use of subjective weight parameters like user interest. Reference [5-6] proposed an improved ID3 algorithm combining fuzzy sets and rough sets. The roughness of the attributes was used to replace the information entropy in the ID3 algorithm, and the roughness was used to divide the attributes, using this method to solve the problem of the multi-value bias. References [8-9] proposed C4.5 and CART (Classification and Regression Tree) algorithms. The C4.5 algorithm splits continuous attributes and minimizes the entropy of the segmented information using the information gain rate as the criterion for...
splitting attributes. In the case of continuous attribute values, the CART algorithm uses the minimum residual variance to determine the optimal partition of the regression tree to generate a regression tree. However, the output results of the C4.5 algorithm and the CART algorithm have a specific range of continuous attribute values, which is not easy to understand.

The above improvement scheme still has some analysis and discussion space in accuracy and application scenarios. In order to solve the above-mentioned ID3 algorithm attribute preference problem, this paper introduces the GINI index as a measurement threshold to let it control the information gain and balance the information gain preference for multiple attribute values. This paper will use the classification ID3 algorithm and the MIND algorithm to mine and process weather, temperature, humidity, wind and other information in agricultural big data to determine whether it is suitable for agricultural planting production and provide computer decision support for smart agriculture. By the application of classification ID3 algorithm and the MIND algorithm, the effectiveness of the MIND algorithm is verified.

2. classification ID3 algorithm

2.1 the fundamental of Classification ID3 algorithm

The classification ID3 algorithm mainly relies on the establishment of a decision tree, which was created by J. Ross. Quinlan in 1986. ID3 uses a top-down greedy strategy to build a decision tree based on the information gain effect. Its gain decision attribute and classification ability are measured according to the information gain, thereby selecting the decision node attributes, and applying the tree-building method to an iterative model. The core of the ID3 algorithm is the property with the largest information gain after splitting. The smaller the expected information is, the greater the information gain is.

If a sample set is $X=\{X_1, X_2, ..., X_n\}$, the number of samples is $|X|$. If $Y=\{Y_1, Y_2, ..., Y_m\}$ is a set of decision attributes, $Y_j (j=1, 2, ..., m)$ represents the j-th attribute, and $|Y_j|$ represents the number of samples in $Y_j$. Then information entropy of $Y$ represents:

$$\text{info}(Y)=-\sum_{j=1}^{m} P_j \times \log_2 P_j$$  \hspace{1cm} (1)

In the formula $P_j$ is the j-th probability. Generally, the ratio of the number of samples in this category to the total number of samples is used for estimation.

If $D=\{D_1, D_2, ..., D_k\}$ is a set of conditional attributes, $D_i (i=1, 2, ..., k)$ is the value of the i-th conditional attribute, and $|D_i|$ represents the number of samples in $D_i$, and $D_{ij}$ represents the j-th $D_i$, and $|D_{ij}|$ represents the number of samples in $D_{ij}$. Then the information entropy divided by the conditional attribute $D$ is:

$$\text{info}(D)=-\sum_{i=1}^{k} w_i \times \text{info}(D_i)$$  \hspace{1cm} (2)

In the formula: $\text{info}(D_i)=-\sum_{j=1}^{m} P_{ij} \times \log_2 P_{ij}, w_i=\frac{|D_i|}{|X|}$.

The difference between the information entropy before and after the attribute $D$ is divided is the information gain:

$$\text{Gain}(D)=\text{info}(Y)\text{-info}(D)$$  \hspace{1cm} (3)

The ID3 algorithm is a typical decision tree learning algorithm. The core of the ID3 algorithm is to use the information gain method as the criterion of the attribute selection at each level of the decision tree to help determine the appropriate attributes to use when generating each node. In this way, the attribute with the highest information gain can be selected as the test attribute of the current node, so that the information required for classification using the subset of training samples obtained by the attribute division is the smallest.

2.2 the deficiencies of the classification ID3 algorithm

The classification ID3 algorithm has a clear theory, simple method, and strong learning ability, and it is suitable for dealing with large-scale learning problems. However, the classification ID3 algorithm is
not suitable for all data mining and machine learning problems because the classification ID3 algorithm is inefficient. Following are the disadvantages of the ID3 algorithm:

1. The ID3 algorithm does not consider continuous values, and cannot distinguish between features that are continuous values.
2. The ID3 algorithm tends to choose attributes with more categories (more branches will lead to greater information gain).
3. The ID3 algorithm only generates trees and is prone to overfitting. Because it is more sensitive to noise data generated by samples with wrong decision attribute values and wrong categories.
4. The ID3 algorithm uses a greedy algorithm. Each partition is considered a local optimization, and the local optimization is not a global optimization. Usually, pruning is required, and decision tree pruning is the overall optimization of the model.

3. Experimental verification and result analysis
To verify that the improved algorithm, MIND algorithm, has certain effectiveness, this paper uses 1000 agricultural data sets in Table 1 for experimental verification and result analysis.

Table 1 agricultural data sets

| serial number | weather | temperature | humidity | gale | suitable for planting |
|---------------|---------|-------------|----------|------|-----------------------|
| 1             | sunny   | high        | high     | no   | no                    |
| 2             | sunny   | high        | high     | yes  | no                    |
| 3             | cloudy  | high        | high     | no   | yes                   |
| 4             | rainy   | suitable    | high     | no   | yes                   |
| 5             | rainy   | low         | suitable | no   | yes                   |
| 6             | rainy   | low         | suitable | yes  | no                    |
| 7             | cloudy  | low         | suitable | yes  | yes                   |
| 8             | sunny   | suitable    | high     | no   | no                    |
| 9             | sunny   | low         | suitable | no   | yes                   |
| 10            | rainy   | suitable    | suitable | no   | yes                   |
| 11            | sunny   | suitable    | suitable | yes  | yes                   |
| …            | …      | …            | …        | …    | …                     |
| 999          | cloudy  | high        | suitable | no   | yes                   |
| 1000         | rainy   | suitable    | high     | yes  | no                    |

This experiment was conducted in 2 times. Firstly, the effectiveness of the classification ID3 algorithm and the MIND algorithm in agricultural planting decisions was verified. Secondly, it is compared with other improved algorithms, such as CART algorithm, C4.5 algorithm, N-value attribute, fuzzy entropy and other algorithms. The parameters for the comparison experiment are divided into 10 groups of 1000 data sets, and each group has 100 data sets. The data set is divided into a training set and a test set, the ratio of which is 7: 3, and the maximum allowable error of leaf nodes is 0.001 to build the model. The model evaluation methods are the accuracy of the training set, the accuracy of the test set, and the accuracy of cross validation for ten times.

3.1 Classification ID3 algorithm application experiment
For the data set given in the table, the classification attributes are selected, and then a decision tree is established to determine whether it is suitable for planting.

Calculate the information entropy according to the classification ID3 algorithm:

\[ \text{info}(Y) = - \sum_{j=1}^{m} P_j \times \log_2 P_j = 0.94 \]

Information entropy of weather:

\[ \text{Info}(\text{weather}) = - \sum_{j=1}^{m} P_{ij} \times \log_2 P_{ij} = 0.694 \]

According to the calculation, the information gain of the weather:
Gain(D)=\text{info}(Y)-\text{info}(D)=0.246

Similarly, the information gain of temperature, humidity, and wind can be calculated according to the above methods and formulas.

G(temperature)=0.029, G(humidity)=0.151, G(wind)=0.048

It can be known that the weather information gain is the largest, so weather is selected as the root node of the decision tree, as shown in Figure 1.

After the root node is found, the search process of other leaf nodes is the same as that of the root node.

3.2 MIND algorithm application experiment

According to the ID3 algorithm, the key to generating a decision tree is to find the root node. The ID3 algorithm calculates the information entropy and information gain of each attribute to find the root node. The following introduces an improved algorithm of the ID3 algorithm, the MIND algorithm, which finds the root node and the leaf node by calculating the Gini coefficient of each attribute. The Gini coefficient is a proportional value, between 0 and 1, and it is an important indicator used internationally to comprehensively examine the difference in income distribution among residents. We can use the idea of the Gini coefficient to calculate the degree of attribute distribution to select the root node of the decision tree.

In fact, the GINI index was first used in economics and was mainly used to measure the fairness of income distribution. In computing CART algorithm with decision tree, the GINI index is used to measure the purity or uncertainty of the data, and the GINI index is used to determine the optimal dichotomous of the categorical variables.

In classification problems, if there are K classes, probability that the sample points belong to k is $P_k$, then definition of GINI index for probability distribution is shown:

$$\text{GINI}(p)=\sum_{k=1}^{K} p_k (1 - p_k)=1 - \sum_{k=1}^{K} p_k^2$$

If the sample set D is divided into two parts D1 and D2 according to a certain feature A, then under the condition of feature A, the GINI index of the set D is defined as:

$$\text{GINI}(D, A) = \frac{D_1}{D} \text{Gini}(D_1) + \frac{D_2}{D} \text{Gini}(D_2)$$

GINI (D, A) represents the uncertainty of data set D with different grouping of feature A. The larger the GINI index is, the greater the uncertainty of the sample set is, which is similar to the concept of entropy.

For the data set given in Table 1, the calculation using the MIND algorithm is as follows:

First, each attribute in Table 1 is divided into different small trees according to categories, as shown below:
Because whether to use planting is a class label attribute, let \( S_1 \) (yes) = 9 and \( S_2 \) (no) = 5. Calculate the Gini coefficient of weather attributes. It is known that sunny weather is suitable for planting as 2 and unsuitable for planting as 3, so it is recorded as sunny (2, 3). Cloudy weather is suitable for planting as 4 and unsuitable for planting as 0, so it is recorded as cloudy (4, 0). Rainy weather is suitable for planting as 3 and unsuitable for planting as 2, so it is recorded as rainy (3, 2).

\[
\text{GINI}(\text{sunny}) = \sum_{k=1}^{2} p_k (1 - p_k) = 0.48 \\
\text{GINI}(\text{cloudy}) = \sum_{k=1}^{4} p_k (1 - p_k) = 0.12 \\
\text{GINI}(\text{rainy}) = \sum_{k=1}^{3} p_k (1 - p_k) = 0.49
\]

From the above formula, the Gini coefficient of the weather attribute is:

\[
\text{GINI(\text{weather})} = \frac{D_1}{D} \text{GINI}(D_1) + \frac{D_2}{D} \text{GINI}(D_2) + \frac{D_3}{D} \text{GINI}(D_3) = 0.343
\]

According to the formula the same can be calculated \( \text{GINI}\) (humidity) = 0.44, \( \text{GINI}\) (temperature) = 0.367, \( \text{GINI}\) (wind) = 0.429.

It can be known from the above calculation that the Gini coefficient of the weather is the smallest, so the weather is selected as the root node of the decision tree. This conclusion is the same as the ID3 algorithm. At the same time, the root node decision tree is consistent with Figure 1. It can be known from the calculation that the classification ID3 algorithm and the MIND algorithm is effective in agricultural planting decision.

### 3.3 Comparison of MIND algorithm with algorithms of other literatures

The comparative experimental accuracy results are shown in Table 2.

| Algorithm                  | Accuracy of Training Set | Accuracy of Test Set | Accuracy of Cross Validation for Ten Times |
|----------------------------|--------------------------|----------------------|------------------------------------------|
| MIND Algorithm             | 0.9938                   | 0.9631               | 0.9526                                   |
| ID3 Algorithm              | 0.9159                   | 0.9524               | 0.9372                                   |
| CART Algorithm             | 0.9968                   | 0.9522               | 0.9351                                   |
| C4.5 Algorithm             | 0.9917                   | 0.9714               | 0.9258                                   |
| N-value Attribute Algorithm| 0.9938                   | 0.9936               | 0.9201                                   |
| Fuzzy Entropy Algorithm    | 0.9265                   | 0.9425               | 0.9225                                   |
According to Table 2, it can be concluded that the accuracy of the MIND algorithm is improved compared to the ID3 algorithm. The MIND algorithm can reduce the situation that the information gain is biased to a large number of attribute values to a certain extent. The specific experimental results are analyzed as follows:

1. In this case, the accuracy of the training set of the MIND algorithm and other optimized algorithms from the literatures is significantly improved. And the accuracy of training set used by the MIND algorithm is significantly higher than other algorithms.

2. The accuracy of cross validation for ten times used by fuzzy entropy algorithm is higher than used by N-value attribute algorithm. The accuracy of cross validation for ten times used by C4.5 algorithm is higher than the accuracy used by fuzzy entropy algorithm. The accuracy of cross validation for ten times used by CART algorithm is higher than used by C4.5 algorithm. It can be seen that the improved algorithm has higher accuracy than the optimized algorithm, and the experimental results are more stable.

3. The accuracy of cross validation for ten times used by MIND algorithm is highest. Therefore, the MIND algorithm is not only effectively improved, but also has more scientific meaning in agricultural planting decisions.

4. Conclusion
The classification ID3 algorithm is an important algorithm in data mining, because it is a greedy algorithm, each partition only considers local optimization, not global optimization. Pruning is usually required, and decision tree pruning is the overall optimization of the model. The improved algorithm, MIND algorithm can reduce the amount of information of unimportant attributes to a certain extent. Compared with the classification ID3 algorithm, it makes the entire operation more independent, which depends on more attributes, thereby improving the classification rules and results, and finding the root node of the decision tree easily. This paper improves the problem of attribute value preference in the algorithm, through a large number of experimental verifications and comparative experiments. It is concluded that the MIND algorithm is more suitable for making decisions on agricultural planting, thereby effectively improving agricultural production and farmers' lives.

Acknowledgments
This work was supported by Sichuan Agricultural University education reform project X2013039, X2014025 "Agricultural Information Engineering" Sichuan key laboratory of higher education.

References
[1] Weiming Du, Yu Ran.(2018) Research on Decision Tree ID3 Algorithm[J]. Technology horizon,(11):145-146,103.
[2] Chong Liu, Lei Yang, Na Li. (2016) Research on ID3 Classification and Pruning Algorithm[J]. Software Guide,15(12):33-34.
[3] Mei Li, Zhipin Zhang.(2018) Formula improvement in ID3 algorithm[J]. Computer Knowledge and Technology,14(29):20,28.
[4] Wenting Wu.(2019) An Improved ID3 Algorithm Based on Dynamic Attribute Weights[J]. Computer Knowledge and Technology,15(6):223-225.
[5] Sibo Wu,Zhigang Chen, Rui Huang.(2016) ID3 optimization algorithm based on correlation coefficient[J]. Computer Engineering and Science,38(11):232-234
[6] Junhai Zhai, Huachao Wang, Sufang Zhang.(2010) A Fuzzy Classification Algorithm Based on Fuzzy Entropy[J]. Computer Engineering and Applications,46(20):176-180.
[7] Gu Gong, Junhui Lu, Yongqing Huang, etc. (2009) Methods for effectively improving the C5.0 algorithm[J]. Computer Engineering and Design,30(22):5197-5199.
[8] Liang Zhang, Qian Ning.(2015) Two Improvements and Applications of CART Decision Tree [J]. Computer Engineering and Design,36(5).
[9] Xiaowei Wang, Yuming Jiang. (2011) Analysis and Improvement of Decision Tree ID3 Algorithm [J]. Computer Engineering and Design, 32(9): 3069-3072.