Improved ID3 Algorithm

Liu Yuxun  
College of Information Science and Engineering  
Henan University of Technology  
Zhengzhou, China  
LYX6250@126.com

Xie Niuniu  
College of Information Science and Engineering  
Henan University of Technology  
Zhengzhou, China  
Xieniuniu219@163.com

Abstract—as the classical algorithm of the decision tree classification algorithm, ID3 is famous for the merits of high classifying speed easy, strong learning ability and easy construction. But when use it to classify, there does exist the problem of inclining to choose attributions which has many values, which affects its practicality. This paper for solving the problem a decision tree algorithm based on attribute-importance is proposed. The improved algorithm uses attribute-importance to increase information gain of attribution which has fewer attributions and compares ID3 with improved ID3 by an example. The experimental analysis of the data show that the improved ID3 algorithm can get more reasonable and more effective rules.

Keywords—decision tree; ID3 algorithm; attribute-importance

I. INTRODUCTION

With the rapid development of information technology and network technology, different trades produce large amounts of data every year. The data itself can not bring direct benefits to people, so what people need is the information hiding in the data. How to effectively mine this information from large data has become the hotspot. Therefore, the data mining technology comes into being. Data mining is to discover the relationship and rules exiting in the data, to predict future trends based on the existing data, finally to fully explore and use these wealth knowledge hiding in the database[1,2]. With the rising of data mining, decision tree plays an important role in the process of data mining and data analysis.

As the classical algorithm of the decision tree classification algorithm, ID3 algorithm has the merits of high classifying speed, strong learning ability and simple construction [3]. However, D3 algorithm is also unsatisfactory in practical application. When using it to classify, there does exist the problem of inclining to choose attributions which have more values, and overlooking attributions which have less values. For solving the problem this paper proposes an improved decision tree algorithm based on attribute-importance. The analysis of the experimental data show that the improved ID3 algorithm gets more reasonable and more effective classification rules.

II. ID3 ALGORITHM

ID3 algorithm [4] is a decision tree learning algorithm based on information entropy proposed by Quinlan in 1986 and its predecessor is CLS algorithm [5]. Quinlan introduces Shannon's information theory into the decision tree algorithm. The core of ID3 algorithm is: selecting attributes from all levels of decision tree nodes; using information gain as attribute selection criteria; each selecting an attribute with the largest information gain to make decision tree nodes; establishing branches by the different values of the node; building the decision tree nodes and branches recursively according to the instances of various branches; until a certain subset of the instances belonging to the same category.

Set \( S \) is a collection of \( S \) data samples. Assume that class label attribute has \( m \) different values, define different classes \( \{ C_1, C_2, \ldots, C_m \} \). Let \( s \) represent the sample number of class \( C_i \). For a given sample the exception information for the classification is calculated as follows:

\[
I(s_1, s_2, \ldots, s_m) = -\sum_{i=1}^{m} p_i \log p_i
\]

(1)

Where, \( p_i \) is the probability of any sample belonging to class \( C_i \). Assumed attribute \( A \) has \( v \) different values: \( \{ a_1, a_2, \ldots, a_v \} \), divide \( S \) into \( v \) sub-sets \( \{ S_{a_1}, S_{a_2}, \ldots, S_{a_v} \} \) by attribute \( A \). If make \( A \) as a test attribute (that is the best split attribute), this subset is the branch getting from the nodes including set \( S \). Assumed \( s_i \) is the sample number of class \( C_i \) in subset \( S_i \). The entropy divided by \( A \) is given as follows:

\[
E(A) = -\sum_{j=1}^{v} \frac{s_j + s_{j+1} + \cdots + s_m}{s} I(s_1, s_2, \ldots, s_m)
\]

(2)

Here, \( \frac{s_j + s_{j+1} + \cdots + s_m}{s} \) is the weight of the \( j \) subset, and equals to the sample number of a subset (that is the value of \( A \) is \( a_j \) ) dividing the total number of \( S \). The smaller entropy is, the higher purity of the divided subset is. For the given subset \( S_i \), its expectation \( o \) information is given by the following formula:

\[
I(s_1, s_2, \ldots, s_m) = -\sum_{i=1}^{m} p_i \log p_i
\]

(3)

Where \( p_i = \frac{s_i}{s} \) is the probability that the sample \( s_i \) belongs to the class \( C_i \). Get information gain according to the expectations information and entropy. Information Gain is calculated by the following formula:

\[
Gain(A) = I(s_1, s_2, \ldots, s_m) - E(A)
\]

(4)
Calculate the information gain of each attribute according to ID3 algorithm, and select the attribute with the highest information gain (that is \( \text{Gain}(A) \) gets the maximum value) as the test attribute for a given set. Create a node of the selected test attribute; mark according to this attribute; create a branch for each value of this attribute; and accordingly divide the sample.

Therefore, the principle of selecting attributes as the test attribute is: select the attribute with the highest information gain. Selecting information gain as the attribute selection criteria by ID3 algorithm has the problem of attribute bias. That is this method does have the problem of inclining to choose attributions which have more values. While it is not reasonable to select the attributions having more values which are not often the best attributes.

### III. IMPROVED ID3 ALGORITHM

But using it to classify, there does exist the problem of inclining to choose attributions which have more values, and overlooking attributions which have less values. For solving the problem this paper proposes an improved decision tree algorithm based on attribute-importance. The algorithm through introducing attribute-importance emphasizes the attributes with less values and higher importance, dilute the attributes with more values and lower importance, and solve the classification defect of inclining to choose attributions with more values. The analysis of the experimental data show that the improved ID3 algorithm gets more reasonable and more effective classification rules. In order to increase the attributes which have fewer values and high importance, and reduce the attributes which have more values and have low importance, improved ID3 algorithm based on attribute importance is proposed in this paper. A new attribute selection criterion of ID3 algorithm has improved, through introducing attribute-importance emphasizes the attributes with less values and higher importance, dilute the attributes with more values and lower importance, and solve the classification defect of inclining to choose attributions with more values. The analysis of the experimental data show that the improved ID3 algorithm gets more reasonable and more effective classification rules. In order to increase the attributes which have fewer values and high importance, and reduce the attributes which have more values and have low importance, improved ID3 algorithm based on attribute importance is proposed in this paper. A new attribute selection criterion of ID3 algorithm has improved, through introducing attribute-importance emphasizes the attributes with less values and higher importance, dilute the attributes with more values and lower importance, and solve the classification defect of inclining to choose attributions with more values. The analysis of the experimental data show that the improved ID3 algorithm gets more reasonable and more effective classification rules.

The corresponding formula (4) becomes:

\[
\text{Gain}(A) = I(s_1, s_2, \ldots, s_v) - E'(A)
\]

Thus, when choose the attributes, people can use \( \text{Gain}(A) \) instead of \( \text{Gain}(A) \). That is people can choose the attribute with highest \( \text{Gain}(A) \) as splitting attribute. Theoretically the algorithm through introducing attribute-importance emphasizes the attributes with less values and higher importance, dilute the attributes with more values and lower importance, and solve the classification defect of inclining to choose attributions with more values.

### IV. ALGORITHM VALIDATION

This paper uses the same test data in the literature [6] as showing in table 1.

<table>
<thead>
<tr>
<th>D</th>
<th>dressing index</th>
<th>temperature</th>
<th>humidity</th>
<th>wind power</th>
<th>weather comfort</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>more</td>
<td>high</td>
<td>great</td>
<td>no</td>
<td>uncomfortably</td>
</tr>
<tr>
<td>2</td>
<td>more</td>
<td>high</td>
<td>great</td>
<td>no</td>
<td>uncomfortably</td>
</tr>
<tr>
<td>3</td>
<td>more</td>
<td>high</td>
<td>moderate</td>
<td>no</td>
<td>uncomfortably</td>
</tr>
<tr>
<td>4</td>
<td>normal</td>
<td>high</td>
<td>moderate</td>
<td>no</td>
<td>comfortable</td>
</tr>
<tr>
<td>5</td>
<td>normal</td>
<td>high</td>
<td>great</td>
<td>no</td>
<td>comfortable</td>
</tr>
<tr>
<td>6</td>
<td>many</td>
<td>moderate</td>
<td>great</td>
<td>no</td>
<td>comfortable</td>
</tr>
<tr>
<td>7</td>
<td>many</td>
<td>moderate</td>
<td>great</td>
<td>no</td>
<td>comfortable</td>
</tr>
<tr>
<td>8</td>
<td>many</td>
<td>high</td>
<td>normal</td>
<td>no</td>
<td>comfortable</td>
</tr>
<tr>
<td>9</td>
<td>many</td>
<td>high</td>
<td>normal</td>
<td>great</td>
<td>uncomfortably</td>
</tr>
<tr>
<td>10</td>
<td>more</td>
<td>moderate</td>
<td>great</td>
<td>no</td>
<td>uncomfortably</td>
</tr>
<tr>
<td>11</td>
<td>more</td>
<td>moderate</td>
<td>great</td>
<td>no</td>
<td>uncomfortably</td>
</tr>
<tr>
<td>12</td>
<td>many</td>
<td>moderate</td>
<td>normal</td>
<td>no</td>
<td>uncomfortably</td>
</tr>
<tr>
<td>13</td>
<td>many</td>
<td>moderate</td>
<td>normal</td>
<td>great</td>
<td>uncomfortably</td>
</tr>
<tr>
<td>14</td>
<td>more</td>
<td>moderate</td>
<td>normal</td>
<td>great</td>
<td>comfortable</td>
</tr>
<tr>
<td>15</td>
<td>more</td>
<td>moderate</td>
<td>normal</td>
<td>great</td>
<td>comfortable</td>
</tr>
<tr>
<td>16</td>
<td>normal</td>
<td>moderate</td>
<td>great</td>
<td>no</td>
<td>comfortable</td>
</tr>
<tr>
<td>17</td>
<td>normal</td>
<td>moderate</td>
<td>great</td>
<td>no</td>
<td>comfortable</td>
</tr>
<tr>
<td>18</td>
<td>normal</td>
<td>high</td>
<td>normal</td>
<td>no</td>
<td>comfortable</td>
</tr>
<tr>
<td>19</td>
<td>many</td>
<td>moderate</td>
<td>great</td>
<td>no</td>
<td>uncomfortably</td>
</tr>
<tr>
<td>20</td>
<td>normal</td>
<td>high</td>
<td>normal</td>
<td>mediu</td>
<td>comfortable</td>
</tr>
</tbody>
</table>

Table I gives the data set affecting summer weather comfort. There are four attributes: dressing index, temperature, humidity, wind power, weather comfort. These four attributes are divided into two classes: comfortable (positive example expressed by P) and uncomfortable (counter example expressed by N). Decision tree by ID3

\[
E'(A) = \sum_{j=1}^{s} \left( \frac{s_j + q(A)I(s_1, s_2, \ldots, s_v)}{s} \right) + q(A)I(s_1, s_2, \ldots, s_v) \]

where \( p \) is the probability of attribute \( A \) belonging to class \( C \). Obviously, \( 0 \leq q \leq 1 \).

Several areas should be taken to when using attribute importance: First, select realistic attribute importance of the attribute with fewer values and high importance according to experience, knowledge, or the actual situation in this field. It is a vague concept, often referring to prior knowledge about a transaction, including the domain knowledge, expert advice or the actual situation. In the generation of decision trees, it refers to the rules and selection factors which affect the decision tree generation process, and it is a changing value.

\[
0 \leq q = q(A) \leq \min(p_1, p_2, \ldots, p_s)
\]
algorithm and improved algorithm are constructed as following. Obtain the decision tree using ID3 algorithm to classify training set samples is shown in figure 1.

![Diagram of decision tree by ID3 algorithm](image)

Figure 1. decision tree by ID3 algorithm

Get the classification rules as follows according to the path from the root node to the leaf nodes in figure 1:

1) IF(dressing index=more and temperature=normal)THEN(class=positive example)
2) IF(dressing index =more and temperature =great)THEN(class = counter example)
3) IF(dressing index =more and temperature =moderate)THEN(class = counter example)
4) IF(dressing index =more and temperature =high and wind power =great) THEN (class= positive example)
5) IF(Dressing index =more and temperature =high and wind power =no)THEN(class = counter example)

In real life, the amount of clothing of many people is the subjective behavior, and has a great relationship with the individual actual situation, for example, usually the elderly, the sick, infants, pregnant women will dress more than adults. As a result, it may lead to that people are not very concerned about humidity and temperature and are only concerned about dressing index. For such cases, people can introduce attribute importance to reduce the conditional entropy of humidity, temperature, and to increase the classification importance of humidity and temperature.

So to some extent dress index can reflect weather comfort, but it can not be the critical condition, that is it is not a very important class attribute. Finally it needs to reduce the importance of dressing index, relatively raises the importance of temperature and humidity in classification.

Regenerate the decision tree using improved ID3 algorithm according to table 1.

Through test, set the attribute importance of dressing index and Wind power is: \(q(dressing index) = q(wind power) = 0\), the attribute importance of temperature and humidity is \(q(temperature) = q(humidity) = 0.25\). First classify according to root node. From table 1, the number of positive example is 9, and the number of counter example is 11. So entropy value is:

\[ I = -\frac{9}{20}\log_{2}\frac{9}{20} -\frac{11}{20}\log_{2}\frac{11}{20} = 0.99277 \]

Choose the dressing index as the test attribute, then get the information entropy of it as followings:

\[ E(dressing index) = 7\left(-\frac{5}{7}\log_{2}\frac{5}{7} - \frac{2}{7}\log_{2}\frac{2}{7}\right) + 11\left(-\frac{6}{11}\log_{2}\frac{6}{11} - \frac{5}{11}\log_{2}\frac{5}{11}\right) = 0.50918 \]

At this point the information gain is:

\[ Gain(dressing index) = I - E(dressing index) = 0.48359 \]

Choose the temperature as the test attribute, then get the information entropy of it as followings:

\[ E(temperature) = 9\left(-\frac{4}{9}\log_{2}\frac{4}{9} - \frac{5}{9}\log_{2}\frac{5}{9}\right) + 11\left(-\frac{7}{11}\log_{2}\frac{7}{11} - \frac{4}{11}\log_{2}\frac{4}{11}\right) = 0.48191 \]

At this point the information gain is:

\[ Gain(temperature) = I - E(temperature) = 0.52820 \]

Choose the humidity as the test attribute, then get the information entropy of it as followings:

\[ E(humidity) = 12\left(-\frac{8}{12}\log_{2}\frac{8}{12} - \frac{4}{12}\log_{2}\frac{4}{12}\right) + 8\left(-\frac{3}{8}\log_{2}\frac{3}{8} - \frac{5}{8}\log_{2}\frac{5}{8}\right) = 0.46457 \]

At this point the information gain is:

\[ Gain(humidity) = I - E(humidity) = 0.52820 \]

Choose the wind power as the test attribute, then get the information entropy of it as followings:

\[ E(wind power) = 7\left(-\frac{4}{7}\log_{2}\frac{4}{7} - \frac{3}{7}\log_{2}\frac{3}{7}\right) + 11\left(-\frac{3}{11}\log_{2}\frac{3}{11} - \frac{2}{11}\log_{2}\frac{2}{11}\right) = 0.98757 \]

At this point the information gain is:

\[ Gain(wind power) = I - E(wind power) = 0.00520 \]

Comparison shows that:

\[ Gain(humidity) > Gain(temperature) > Gain(dressing index) > Gain(wind power) \]

so first select the humidity to classify, repeat this process to get the final decision tree shown in figure 2.

![Diagram of decision tree by improved algorithm](image)

Figure 2. decision tree by improved algorithm
Get the classification rules as follows according to figure 2:

6) IF(humidity = great and dressing index = more or many) THEN (class = counter example)

7) IF(humidity = normal and dressing index = more or normal) THEN (class = counter example)

8) IF(humidity = normal and dressing index = many and temperature = moderate) THEN (class = counter example)

9) IF(humidity = normal and dressing index = many and temperature = high and wind power = no) THEN (class = positive example)

10) IF(humidity = normal and dressing index = many and temperature = high and wind power = great) THEN (class = counter example)

Among them, the rules (1), (2), (3), (5) are more suitable for realistic conditions, and can be used to predict the actual weather comfort. Comparing rule extraction of the two decision tree algorithms, the resulting decision trees and classification rules by ID3 algorithm and improved algorithm are very different. But the improved algorithm has reduced the importance of dress index which can effect on the weather comfort. Because it is far away from the root node, so increases the importance of humidity, temperature and wind power in the classification. Obviously, this is more consistent with the actual situation, so as to achieve the desired objective of the improved algorithm.

The experimental results show that: Improved ID3 algorithm compared with the traditional ID3 algorithm has better classification accuracy. This is because: the improved algorithm through introducing attribute-importance emphasizes the attributes with fewer values and higher importance, dilute the attributes with more values and lower importance, and solve the classification defect of inclining to choose attributes with more values.

V. Conclusion

The improved decision tree is by introducing attribute-importance in attribute information entropy to calculate information gain of each attribute and make the attribute with the maximum information gain as the splitting attribute. Its size is determined by experience, knowledge, or the actual situation in this field of the decision-makers. The improved algorithm through introducing attribute-importance emphasizes the attributes with fewer values and higher importance, dilute the attributes with more values and lower importance, and solve the classification defect of inclining to choose attributes with more values. The analysis of the experimental data show that improved ID3 algorithm compared with ID3 algorithm has better classification accuracy and can get more reasonable, more effective and more objectively actual classification rules.

REFERENCES


