**Improvement of ID3 Algorithm Implementation**

Ming-hao LIU, Nan LIU* and Wei SHANG

School of Computer Science and Technology, Chongqing University of Posts and Telecoms, Chongqing, 400065, China

*Corresponding author

**Keywords:** ID3 algorithm, Data mining, Spatial information.

**Abstract.** ID3 algorithm is a classical algorithm in decision tree algorithm, commonly used in data mining. Owing to that spatial information data has the characteristics of large capacity and diversity, when ID3 algorithm does data mining for spatial information, in the process of generating decision tree, some nodes will appear special sample sets which the information gain of each classification attribute is 0 and values of result attribute are not unique in. Now, conventional implementation of ID3 algorithm can't ensure the generation of decision tree. Based on this, conventional implementation of ID3 algorithm is improved, and the fault tolerance of algorithm implementation is enhanced. Improved algorithm implementation can do data mining for spatial data set.

**Introduction**

With the rapid development of Internet technology and software technology, people have been in the age of information. Data mining is rising due to the new surge in the amount of information. Classification algorithm is one of the most common data analysis methods in data mining. It can correctly get the class of test sample by training existing data set [1]. Discriminant analysis, rule induction, decision tree, neural network, naive bias classification algorithm, genetic algorithm and so on are included in current classification algorithms. Considering the decision tree algorithm is simple and easy to understand. It can get high precision and generative the model easily. People have been widely using it. Decision tree algorithm has a great influence in data mining.

ID3 algorithm is one of the most influential algorithms in decision tree, which was first proposed by Quinlan. Each time ID3 uses all training sample. On the one hand, the decision tree built by ID3 is simple and intuitive. On the other hand, it can effectively reduce the influence of several noise data as well as deal with discrete attribute values. The disadvantage of ID3 is that it biases multivalued attribute.

When using ID3 to accomplish knowledge mining or do the research on algorithm improvement, we often take the small amount of data set as the experimental data [2]. When ID3 is used to deal with the large data set, there will be some special subset of samples, which leads the conventional algorithm cannot build tree.

**ID3 Algorithm Introduction**

For ID3 algorithm, in the process of generating tree, current node will choose the best classification attribute, according to the values of this classification attribute, produce corresponding branches and each branch generates a new node. Iterate until a tree is generated [3,4]. The current node uses information gain as the criterion of selecting split attribute.

Assuming that number of samples in current node's sample set $S$ is $n$, according to the values of result attribute, the data set $S$ is divided into $c$ different sample subsets $C_i$ $(i=1,2,\ldots,c)$, number of samples in each sample subset $C_i$ is $|C_i|$, so information entropy of dividing $S$ into $c$ different classes:
\[ E(S) = \sum_{i=1}^{c} p_i \log_2 (p_i) \]  
\[ p_i = \frac{|C_i|}{n} \]

Select the classification attribute \( A \) as test attribute. Assuming that number of different values in attribute \( A \) is \( a \), and \( S_i \) is a sample set of the new branch node which attribute \( A \) takes the value \( x_i \) (\( i=1,2,\ldots,a \)) to generate, information entropy of dividing \( S_i \) into \( c \) different classes is \( E(S_i) \). Expected entropy of selecting classification attribute \( A \) as test attribute:

\[ E(S,A) = \sum_{i=1}^{a} \frac{|S_i|}{|S|} E(S_i) \]  

Information gain \( Gain(S,A) \) of attribute \( A \) relative to sample set \( S \) is as follows:

\[ Gain(S,A) = E(S) - E(S,A) \]  

The larger the \( Gain(S,A) \) is, the more information is provided by test attribute \( A \). ID3 algorithm is to make each node select the attribute whose information gain \( Gain(S,A) \) is the largest as split attribute.

**Algorithm Implementation Improvement**

**Conventional Algorithm Implementation**

Conventional ID3 algorithm [5] is as follows:

1. Create node \( N \);
2. If values of the result attribute of the sample set whose name is Samples in the node \( N \) are the same class \( C \), then return \( N \) as leaf node, and mark leaf node with the class \( C \);
3. If the remaining attribute set whose name is Attribute_list in the node \( N \) is null, then return \( N \) as leaf node, and mark leaf node with the most common class in “Samples”, according to the principle of majority voting;
4. Select the attribute whose name is Best_attribute which has the highest information gain from the remaining attribute set “Attribute_list”;
5. According to the values of “Best_attribute” to generate branch nodes, the sample set of each branch node is the sub sample set whose name is A_samples, “A_samples” is composed of those records in “Samples” whose “Best_attribute” take corresponding value, the remaining attribute set of each branch node is the sub attribute set whose name is A_attribute_list which attribute set “Attribute_list” delete attribute “Best_attribute” to get;

Spatial information data has the characteristics of large capacity and diversity. When ID3 algorithm does data mining for spatial information, conventional algorithm implementation can’t ensure the generation of decision tree. So ID3 algorithm is rarely used to process spatial information data.

**Information Gain Analysis**

ID3 algorithm has the characteristic that when the information gain of each classification attribute is 0, it can’t choose the split attribute. This characteristic determines conventional implementation has poor fault tolerance.

If the sample set of the current node is shown in Table 1, values of each classification attribute are unique, values of result attribute are not unique. Because values of result attribute are not unique, the current node will select split attribute. But owing to values of each classification attribute are unique,
values of Gain(S,Attribute1), Gain(S,Attribute2) and Gain(S,Attribute3) are 0. This node can’t choose split attribute. Now, conventional algorithm implementation will throw an error message.

Table 1. Special Sample Set.

<table>
<thead>
<tr>
<th>Attribute1</th>
<th>Attribute2</th>
<th>Attribute3</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>7</td>
<td>5</td>
<td>vary</td>
</tr>
<tr>
<td>3</td>
<td>7</td>
<td>5</td>
<td>unvarying</td>
</tr>
<tr>
<td>3</td>
<td>7</td>
<td>5</td>
<td>vary</td>
</tr>
<tr>
<td>3</td>
<td>7</td>
<td>5</td>
<td>unvarying</td>
</tr>
</tbody>
</table>

Sample proportion of each value for result attribute in current node is the same as sample proportion of each value for result attribute in each branch node. The description is shown as Figure 1. In Figure 1, sample proportion of the value “vary” is 1/3, sample proportion of the value “unvarying” is 2/3 in each node. Now, the calculated information gain is 0.

![Figure 1. Branches of current node.](image)

If for the sample set of the current node, select any classification attribute to generate branch nodes, sample proportion of each value for result attribute in current node is the same as sample proportion of each value for result attribute in each branch node. Now, gain information of each classification attribute is 0. This node can’t choose split attribute. Conventional algorithm implementation will throw an error message.

**Improved Algorithm Implementation**

Based on the analysis for that the information gain is 0, this paper improves conventional algorithm implementation. If the sample set of a node appears the situation described above, improved algorithm implementation will return the current node as the leaf node to ensure the generation of decision tree. That is to say, Two steps as follows are inserted between the third step and the fourth step in conventional algorithm implementation.

If values of each classification attribute in “Attribute_list” are unique, then return N as leaf node, and mark leaf node with the most common class in “Samples”, according to the principle of majority voting;

If select any classification attribute to generate branch nodes, sample proportion of each value for result attribute in current node is the same as sample proportion of each value for result attribute in each branch node, then return N as leaf node, and mark leaf node with the most common class in “Samples”, according to the principle of majority voting;

Improved algorithm implementation can ensure the generation of decision tree when ID3 algorithm is used to process spatial information data.

**Experiment**

The development language used in this paper is C#, development environment is Visual Studio 2013, and the visual selection of the tree is the TreeView control in C#.

Part of the sample set used in experiment is shown in Figure 2. There are seven classification attributes, a result attribute and 156597 pieces of data in this sample set. There are seven different integer values from 1 to 7 in attribute Dem, four different integer values from 1 to 4 in attribute
population, six different integer values from 1 to 6 in attribute Rail, six different integer values from 1 to 6 in attribute Road, five different integer values from 1 to 5 in attribute Slope, eleven different integer values from 1 to 11 in attribute Market, seven different integer values from 1 to 7 in attribute Water. There are two values in result attribute whose name is Result, one is “vary”, the other is “unvarying”.

<table>
<thead>
<tr>
<th>Dem</th>
<th>Population</th>
<th>Rail</th>
<th>Road</th>
<th>Slope</th>
<th>Market</th>
<th>Water</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>4</td>
<td>5</td>
<td>1</td>
<td>2</td>
<td>6</td>
<td>5</td>
<td>unvarying</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>6</td>
<td>5</td>
<td>unvarying</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>6</td>
<td>6</td>
<td>vary</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>6</td>
<td>6</td>
<td>vary</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>6</td>
<td>6</td>
<td>vary</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>6</td>
<td>6</td>
<td>unvarying</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>6</td>
<td>5</td>
<td>vary</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>6</td>
<td>5</td>
<td>vary</td>
</tr>
</tbody>
</table>

Figure 2. Part of data set in experiment.

Conventional algorithm implement will throw an error message when processing the data set. Figure 3 is part of decision tree which improved algorithm implement generates, and the minus sign “-” signs a branch of the tree. One “-” represents one node. The first “-” represents root node. The root node select attribute Population as split attribute. Number in parentheses is the value which parent node of current node takes to generate current node. This branch can be interpreted as following: when the value of attribute Populaion is 2, the value of attribute Rail is 1, the value of attribute Water is 1, the result is "unvary".

Figure 3. Part of generated decision tree.

Summary

Based on the analysis of that the information gain is 0, this paper improves conventional implementation of ID3 algorithm, and the fault tolerance of algorithm implementation is enhanced. Improved algorithm implementation can do data mining for spatial data set.

Acknowledgement

The authors wish to thank the helpful comments and suggestions from my colleagues in School of Computer Science, Chongqing University of Posts and Telecoms. This research was financially supported by the Science and Technology Foundation Project Affiliated to the Education Department of Chongqing Municipality (Grant NO.KJ140020), and Chongqing Key Program of application and development (cstc2014yykfb30003).

References


