Analysis of C4.5 Algorithm of Water Quality Dataset

Heru Mardiansyah1, Muhammad Zarlis2, Opim Salim Sitompul3
Faculty of Computer Science and Information Technology,
1Universitas Sumatera Utara, Medan, Indonesia
heru.mardiansyah@students.usu.ac.id, m.zarlis@usu.ac.id, opim@usu.ac.id

Abstract. The C4.5 algorithm still has weaknesses in predicting or classifying data if a large number of classes are used which can lead to increased decision-making time. So an approach is needed to improve the performance of the C4.5 algorithm with the selected split attributes that use the application of the average gain value to help predictions. The C4.5 algorithm is one of the Decision Tree methods in the classification process using the information entropy concept. The C4.5 algorithm uses the split criteria from ID3, the Gain Ratio is a modification of the method. The ID3 algorithm uses Information Gain (IG) for the split attribute criteria, while the C4.5 algorithm with Gain Ratio (GR), where the root value comes from high gain. The conclusion of the tests that have been carried out using the Water Quality dataset in the C4.5 method has an accuracy rate of 91.30%, with a classification error rate of 8.70%. Successful implementation using the C4.5 method in predicting the Water Quality dataset.

1. Introduction
In classifying, data requires an object that has been defined in the data. Classification algorithms are widely used in making decisions [1]. The Decision Tree algorithm performs a grouping method such as a flowchart that forms a tree-like structure, with classes divided into nodes. The classification model on the root of the leaves tests the attributes as a result of the classification in the Decision Tree [2]. The C4.5 algorithm is a classification method of data mining that is used to construct a decision tree. Decision trees are used as a reasoning procedure to get answers to the problems that are entered. The classification process can also use the entropy concept in the Decision Tree, namely the C4.5 Algorithm. In this algorithm has split criteria derived from ID3, the Gain Ratio is modified from the previous method. The ID3 Algorithm implements Information Gain in its split attribute criteria, and in the C4.5 algorithm applies a Gain Ratio where the root starts from Gain that has a high value. The classification process uses the C4.5 algorithm, namely by calculating the entropy value. The Gain Ratio value is calculated, then the attribute that has a high Gain Ratio value will be selected as the root and the low one will become the branch, then recalculate the Gain Ratio value of each attribute by not using the selected attribute as the root of the process. Previously, the next process was carried out to produce a Gain value of 0 on the remaining attributes.

In a decision tree approach, pruning is the process of cutting or removing unnecessary branches (nodes). Unnecessary nodes will cause noisy data that is less relevant [3]. This causes a large decision tree called overfitting [4]. [2] The second pruning method is to use threshold pruning. The two pruning methods were tested using the Decision Tree classification model with split criteria.
from the ID3 and C4.5 algorithms (heterogeneous cost) on six datasets and concluded that the two proposed pruning methods can be used for the decision tree classification model.

In this study, we propose a framework for eliminating irrelevant features, without eliminating information contained in the original data uses the C4.5 decision tree algorithm for classifying the data set used. The aim is to predict the classification model with the C4.5 decision tree to get relevant features and to improve C4.5 decision tree accuracy. To evaluate the proposed approach, we used the Water Quality dataset as data.

2. Research Method

In the C4.5 heuristic approach using two approaches in testing carried out with a probability rating as follows:

1. Information gain that minimizes the total entropy value of the subset bias that occurs when testing using numerical data.
2. Obtaining the Gain Ratio results obtained using Information Entropy derived from each attribute of the data used in the test.

The C4.5 algorithm is another type of decision tree method which is similar to a flowchart structure, where each internal node is represented as a test attribute. Branches that represent output by testing the nodes (leaf nodes) by determining the class. The very top node of the tree is the root node. The stages of the C4.5 algorithm are as follows[5]:

a. Calculating of the Entropy value for each attribute:

\[
\text{Entropy}(S) = \sum_{i=1}^{n} - p_i \log_2 p_i
\]  

Where:
- \( S \) = Data
- \( n \) = Number of Partitions S
- \( p_i \) = Proportion of the subsets of S on the i-particle

b. Calculating of the Information Gain value for each attribute:

\[
\text{Gain}(S,A) = \text{Entropy}(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} \times \text{Entropy}(S_i)
\]

Where:
- \( S \) = The amount of data
- \( A \) = Attributes to the test data
- \( n \) = the amount of data from the attribute
- \( |S_i| \) = the measure of data that an attribute has in the data
- \( |S| \) = a measure of the amount of data tested

c. Calculating of the Split Information value for each attribute:

\[
\text{Split Info}_A(D) = - \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2 \left( \frac{|D_j|}{|D|} \right)
\]

Where:
- \( D \) = Sum of all data
- \( A \) = Attributes to data
- \( v \) = Number of partition attributes
\[ | D_j | = \text{Data size on the previous attribute to the partition} \]
\[ | D | = \text{A measure of the number of attributes in the data} \]

d. Calculating the Gain Ratio value for each attribute:

\[ \text{Gain Ratio (A)} = \frac{\text{Gain (A)}}{\text{SplitInfo (A)}} \quad (4) \]

3. Result And Discussion

The preprocessing results are represented in tabular form (x1, x2, x3 ... x32) with the same sorting representation as the data being tested as showns as table 1.

**Table 1. Results of Cleaning Water Quality Status Data**

| No | TSS (mg/L) | DO (mg/L) | COD (mg/L) | BOD (mg/L) | Total Phosphat (mg/L) | ... | Quality Status |
|----|------------|-----------|------------|------------|-----------------------|----|----------------|
| 1  | 2          | 4         | 8          | 0.06       | 0.01                  | ...| Good Condition |
| 2  | 3          | 0.04      | 19.02      | 0.01       | 0.14                  | ...| Good Condition |
| 3  | 3          | 0.04      | 16         | 0.09       | 0.12                  | ...| Good Condition |
| 4  | 4          | 0.01      | 0.71       | 0.32       | 0.18                  | ...| Good Condition |
| 5  | 4          | 0.02      | 8          | 0.05       | 0.11                  | ...| Good Condition |
| 6  | 5          | 0.01      | 8          | 0.06       | 0.01                  | ...| Good Condition |
| 7  | 5          | 4         | 8          | 3          | 0.21                  | ...| Good Condition |
| 8  | 6          | 0.04      | 32         | 0.01       | 0.12                  | ...| Good Condition |
| ...| ...        | ...       | ...        | ...        | ...                   | ...| ...            |
| 117| 97         | 0.03      | 55.04      | 0.02       | 10                    | ...| Heavily Polluted|

Note:
TSS : Total Suspended Solid
DO : Dissolved Oxygen
COD : Chemical Oxygen Demand
BOD : Biochemical Oxygen Demand

The following are the stages of the C4.5 algorithm:

a. Calculation of the value of Entropy for each attribute:

\[ \text{Entropy (S)} = \sum_{i=1}^{n} - P_i \times \log_2 P_i \quad (5) \]

Where:
The total entropy calculation for each class is as follows:

\[
\begin{align*}
\text{Entropy}(S) & = \sum_{i=1}^{n} P_i \times \log_2 P_i \\
& = ((-57/117)*\log_2 (57/117)+(-60/117)*\log_2 (60/117)) \\
& = -0.487\times-1.0374+(-0.512)\times-0.9634 \\
& = 0.5052+0.4932 \\
& = 0.999
\end{align*}
\]

| Data | Class (0) | Class (1) | Entropy |
|------|-----------|-----------|---------|
| 117  | 57        | 60        | 0.69375 |

The examples of entropy calculation results for each attribute are as follows:

b. Entropy of the attribute (TSS (mg/L)):

\[
\begin{align*}
\text{Entropy}(S) & = \sum_{i=1}^{n} P_i \times \log_2 P_i \\
<20 & = ((-10/38)*\log_2 (10/38)+(-28/38)*\log_2 (28/38)) \\
& = -0.263\times-1.925+(-0.736)\times-0.4405 \\
& = 0.506+0.324 \\
& = 0.831 \\
<50 & = ((-26/39)*\log_2 (26/39)+(-13/39)*\log_2 (13/39)) \\
& = -0.666\times-0.584+(-0.333)\times-1.584 \\
& = 0.389+0.527 \\
& = 0.916
\end{align*}
\]

The results obtained from the calculation of each attribute will be shown in the table 3:

| TSS (mg/L) | Data | Class (0) | Class (1) | Entropy | Total Entropy | Information Gain | Split Info |
|------------|------|-----------|-----------|---------|---------------|-----------------|-----------|
|            | TOTAL| 117       | 57        | 60      | 0.999         |                 |           |
|            | <20  | 38        | 10        | 28      | 0.831         |                 |           |
|            | <50  | 39        | 26        | 13      | 0.918         |                 |           |
|            | <90  | 21        | 15        | 6       | 0.863         |                 |           |
|            | <200 | 19        | 6         | 13      | 0.899         |                 |           |
|            | TOTAL|           |           |         | 0.887         | 0.122           | 1.925     |

| TSS (mg/L) | 0.88 | 0.119 | 1.426 |
Classifying the decision tree from the dataset being tested, so that the calculation is carried out in the test and then described in a table called confusion matrix [6]. All about the good and bad parameters in classifying with test data in different classes, namely the process of the confusion matrix as shown in table 4.

|                | Predicted Class |
|----------------|-----------------|
|                | Actual Class    |
|                | 1               |
| DO (mg/L)      |                 |
| <4             | 33              |
| <6             | 64              |
| <8             | 20              |
| COD (mg/L)     |                 |
| <15            | 53              |
| <40            | 45              |
| >100           | 19              |
| BOD (mg/L)     |                 |
| <3             | 47              |
| <7             | 35              |
| >10            | 35              |
| Total phospat (mg/L) |       |
| <0.1           | 38              |
| <0.16          | 43              |
| >0.17          | 36              |
| Fecal Coliform (mg/L) |     |
| >750           | 39              |
| >7500          | 43              |
| <9000          | 35              |
| Total Coliform (mg/L) |     |
| <35000         | 63              |
| >42000         | 54              |
| Pij            |                 |
| 117            |                 |

Based on the table above, it is followed by calculating the value of Accuracy, Particularity and the level of classification error from the C4.5 classification model.
a. **Accuracy** = \( \frac{TP+TN}{TP+TN+FP+FN} = \frac{8+13}{8+13+1+1} = \frac{21}{23} = 0.9130 \times 100\% = 91.30\% \)

The level of closeness between the class prediction and the actual class or the number of correct class predictions from the C4.5 classification model is 91.30%.

b. **Classification Error** = \( \frac{FP+FN}{TP+TN+FP+FN} = \frac{1+1}{8+13+1+1} = \frac{2}{23} = 0.0869 \times 100\% = 8.70\% \)

### 4. Conclusion

The Water Quality Status dataset is obtained from the results of research [7] where the data is the result of data collection by the Ministry of Environment on Water Quality Provisions which are classified into four categories. The Water Quality Status dataset is 117 records with 8 attributes and consists of 2 classes. The conclusion of the tests that have been carried out using the Water Quality dataset in the C4.5 method has an accuracy rate of 91.30%, with a classification error rate of 8.70%. Successful implementation using the C4.5 method in predicting the Water Quality dataset.

### References

[1] Seema R M, Mantu. 2012. *Decision Tree: Data Mining Techniques*. Department of Computer Science Engineering.

[2] Qing-Yun Dai, Chun-ping Zhang and Hao Wu 2016 Research of Decision Tree Classification Algorithm in Data Mining, Int. J of Database Theory and App. 9 5 1-8.

[3] Zhang S. 2012. Decision tree classifiers sensitive to heterogeneous costs. *J of Systems and Software*, 85 4, 771–779. doi:10.1016/j.jss.2011.10.007

[4] Hou S, Hou R, Shi X, Wang J, and Yuan, C 2014 Research on C5.0 Algorithm Improvement and the Test in Lightning Disaster Statistics. *Int. J. of Control and Automation* 7 1 181-190.

[5] Mitchell T M 1997 Machine learning. (Boston: McGraw-Hill).

[6] Witten I 2011 Data Mining: Practical Machine Learning Tools and Techniques Third Edition. (Burlington : Elsevier)

[7] Danades A, Pratama D, Anggraini D, Anggriani D. 2016 Comparison of Accuracy Level K-Nearest Neighbor Algorithm and Support Vector Machine