Analysis Of c4.5 and ID3 Methods in Determining Student Graduation

Titin Qowidho¹, M Zarlis¹, Erna Budhiarti Nababan¹, Alfian Agusnady¹, Berti Sari Sembiring¹

¹ Department of Information Technology, Faculty of Computer Science and Information Technology, Universitas Sumatera Utara, Medan - Indonesia

* titinqowidho123@gmail.com

Abstract. From the performance measurement of the two algorithms that have been done based on the amount of data, it can be concluded that ID3 algorithm has better performance (precision, recall, and accuracy) than C4.5 algorithm. The performance measurement of the two algorithms based on the number of attributes indicates that the ID3 algorithm has better performance (precision, recall, and accuracy) than the C4.5 algorithm. Overall, From the experiments that have been carried out it can be concluded that ID3 has a better performance than C4.5 algorithm. Performance measurement of a data mining algorithm can be done based on several criteria such as accuracy of prediction, speed / efficiency, reliability, scalability and interpretation. In this study the authors suggest using the ID3 algorithm method.

1. Introduction

Every year the Helvetia health institute conducts a new student admission process. Every year the Helvetia health institute also graduates students. The data cannot be maximized maximally, and the data is left to accumulate.

The number of students who graduate and new students who enter each year is not comparable, for that there needs to be a system that can be used to predict student graduation rates. This student graduation prediction system requires the information available to find out whether a student can graduate on time or not. If student graduation can be known early on, the academic party can apply a policy to minimize the number of students who do not graduate on time according to their studies.

Techniques are needed to process the data to provide added value, which is one of them by utilizing data mining to overcome these problems. In this study, the authors apply C4.5 algorithm and Iterative Dichotomiser 3 (ID3). C4.5 This is supervised learning algorithm and it is advanced version of Id3 algorithm. This algorithm is developed by Ross Quinlan. It uses the concept of Id3 algorithm of information entropy [1].

2. Relates Research

Annas Gading Pertiwi, et al. Implement the C4.5 algorithm in analyzing dropout rates. The C4.5 algorithm was chosen to be the method because it proved the results showed an accuracy rate of 71.2%. Based on these results it can be concluded that the development of provincial classification based on dropout rates using the accurate C4.5 algorithm to be used [2].
Xue Bao, et al. Implement a decision tree algorithm to predict crude oil output. This algorithm has the advantage of a simpler tree structure and higher operating efficiency, making it a promising method in the production of crude oil [3].

Erie Aji Panji Kurniawan, et al. In research using iterative algorithms Dichotomizer Tree (ID3) used to process data on red onion plant pests. System accuracy testing results between detection with the results of the calculation of the Iterative Dichotomizer 3 (ID3) method has an accuracy rate of 80% [4].

3. Proposed Method

3.1. Decision Tree
Decision Tree (Pohon Keputusan) is a classification method (taxonomy) that uses a tree representation in which each node represents an attribute, the branch represents the value of the attribute, and the leaf represents the class [5]. Decision tree pruning aims to optimize decision tree and simplify production rules. It usually has two methods: prepruning and post-pruning [6].

3.2. 4.5 Algorithm
C4.5 algorithm is populated in 1993 by Ross Quinlan, who is also introduced ID3 algorithm, the previous version of C4.5 [7]. It splits the Training data with the help of information gain. Those attributes which is having high frequency is considered for splitting the data based on the information available in the dataset [1]. In general, the process flow of C4.5 algorithm for building decision trees in data mining is [5] [8]:

a. Select an attribute as the root node
b. Create a branch for each value.
c. For cases in branches
d. Repeat the process for each branch until all cases in the branch have the same class

The selection of attributes as nodes, either root (root) or internal node is based on the highest Gain value of the existing attributes. The calculation of the Gain value is used by the formula as in Equation 1.

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

Where :
- \( S \) : Case set
- \( A \) : Attribute
- \( n \) : Number of partition attributes \( A \)
- \( |S_i| \) : Number of case on the – i partition
- \( |S| \) : Number of cases in \( S \)

To calculate the Entropy value can be seen in Equation 2.

\[
Entropy(S) = \sum_{i=1}^{n} - p_i \cdot \log_2 p_i
\] (2)

Where :
- \( S \) : Case set
- \( n \) : Number of partitions \( S \)
- \( p_i \) : Proportion of \( S_i \) to \( S \)
3.3. ID3 Algorithm

ID3 algorithm is the most basic algorithm in the decision tree and is a method for building the decision tree of the Computer Science Faculty, Universitas Brawijaya is looking for a solution. The ID3 algorithm performs a thorough search on all possible trees [4]. The following is the ID3 algorithm calculation by finding the value of entropy and information gain [4]:

$$E_{\text{entropy}}(S) = -P \cdot \log_2 P - P' \cdot \log_2 P'$$

Information:

- $P+$ is the probability of a sample $S$ that has a positive class. $P+$ calculated by dividing the number of positive samples ($S+$) by the total number of samples ($S$) so that $P+ = \frac{S+}{S}$
- $P-$ is the probability of the sample $S$ having a negative class. $P-$ calculated by dividing the number of negative samples ($S-$) by the total number of samples ($S$) so that $P- = \frac{S-}{S}$

### 4. Results and Discussion

In this study the authors conducted a trial using 4 attributes (gender, distance, ipk and length of study) data mining process was carried out with the help of data mining software, that is using the RapidMiner application. The following is a comparison analysis of C4.5 and ID3 methods.

#### Table 1. C4.5 and ID3 Method Performance

| Test Data | Accuracy (%) | Classification Error (%) | Precision (%) | Recall (%) | Execution Time (Second) |
|-----------|--------------|--------------------------|---------------|------------|-------------------------|
| 25        | 28.57        | 71.43                    | 50.00         | 35.00      | 0.002                   |
| 50        | 33.33        | 66.67                    | 38.63         | 42.50      | 0.003                   |
| 100       | 43.33        | 56.67                    | 43.33         | 42.50      | 0.005                   |
| 150       | 62.22        | 37.78                    | 62.20         | 79.54      | 0.010                   |
| 200       | 43.33        | 56.67                    | 44.78         | 79.76      | 0.017                   |
| 230       | 34.78        | 65.22                    | 34.96         | 80.00      | 0.021                   |

From table 1 can be seen the comparison method C4.5 and ID3 using 25 data up to 230 data ID3 method has better accuracy in the form of processing time. Next is a graph of several C4.5 and ID3 method comparison processes.
From figure 1 shows that the ID3 algorithm reaches the highest accuracy value in the amount of data 100, whereas C4.5 algorithm achieves the highest accuracy value on 150 data amounts. Then it can be concluded that the ID3 algorithm accuracy value is better than C4.5 algorithm.

In figure 2 the highest precision value is achieved by the C4.5 algorithm on 150 data numbers from the number of attributes 62. The ID3 algorithm is the highest precision value achieved at 150 data numbers from the number of attributes 58.

In figure 3 the highest ID3 algorithm recall value is on the amount of data 100 with attribute 70, while the highest C4.5 recall value is on the amount of 100 data with attribute 80.
From figure 4 shows that the ID3 algorithm achieves a better execution time value with a value of 0.16, from the c4.5 algorithm with a value of 0.21. The next step is to from the decision tree of the ID3 and C4.5 algorithms.

**Figure 5. Decision Tree 150 Data with C4.5**

**Description :**

\[
\text{IPK} = \leq 2.5
\]

| DISTANCE = Far |
|----------------|
| GENDER = Man: 6 Years \( \{5 \text{ Years}=1, 6 \text{ Years}=3, 4 \text{ Years}=0\} \) |
| GENDER = Woman: 5 Years \( \{5 \text{ Years}=1, 6 \text{ Years}=0, 4 \text{ Years}=0\} \) |
| DISTANCE = Normal: 6 Years \( \{5 \text{ Years}=0, 6 \text{ Years}=1, 4 \text{ Years}=0\} \) |

\[
\text{IPK} = \leq 3.0: 6 \text{ Years} \ (5 \text{ Years}=28, 6 \text{ Years}=39, 4 \text{ Years}=4) \\
\text{IPK} > 3.0: 5 \text{ Years} \ (5 \text{ Years}=36, 6 \text{ Years}=9, 4 \text{ Years}=28)
\]

**Figure 6. Decision Tree 150 Data with ID3**

**Description :**

\[
\text{IPK} = \leq 2.5
\]

| DISTANCE = Far |
|----------------|
| GENDER = Man: 6 Years \( \{5 \text{ Years}=1, 6 \text{ Years}=3, 4 \text{ Years}=0\} \) |
| GENDER = Woman: 5 Years \( \{5 \text{ Years}=1, 6 \text{ Years}=0, 4 \text{ Years}=0\} \) |
| DISTANCE = Normal: 6 Years \( \{5 \text{ Years}=0, 6 \text{ Years}=1, 4 \text{ Years}=0\} \) |

\[
\text{IPK} = \leq 3.0: 6 \text{ Years} \ (5 \text{ Years}=28, 6 \text{ Years}=39, 4 \text{ Years}=4) \\
\text{IPK} = > 3.0: 5 \text{ Years} \ (5 \text{ Years}=36, 6 \text{ Years}=9, 4 \text{ Years}=28)
\]
5. Conclusions

From the performance measurement of the two algorithms that have been done based on the amount of data, it can be concluded that ID3 algorithm has better performance (precision, recall, and accuracy) that c4.5 algorithm, see figure 1,2,3,4. The performance measurement of the two algorithms based on the number of attributes indicates that the ID3 algorithm has better performance (precision, recall, and accuracy) that the C4.5 algorithm. Overall, from the experiments that have been carried out it can be concluded that ID3 has a better performance than C4.5 algorithm. Performance measurement of a data mining algorithm can be done based on several inter-criteria such as predictive accuracy, speed / efficiency, reliability, scalability and interpretation. In this study the authors suggest using the ID3 algorithm method.

References

[1] Lokesh S. Katore, et al, Novel Profesional Career Prediction and Recommendation Method For Individual Through Analytics on Personal Traits Using C4.5 Algorithm, Proceeding of 205 Global Conference on Communication Technologies (GCCT 2015).
[2] Annas gading Pertiwi, et al, Classification of Province Based on Dropout Rate Using C4.5 Algorithm, International Conference on Sustainable Information Engineering and Technology, 2017.
[3] Xue Bao, et al, A Method of Predicting Crude Oil Output Based on RS-C4.5 Algorithm, International Conference on Information Science and Control Engineering, 2016.
[4] Erie Aji Panji Kurniawan, et al, Implementasi Metode Iterative Dichotomizer Tree (ID3) Untuk Diagnosis Penyakit Pada tanaman Bawang Merah, Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer.
[5] Yudhi Andrian, et al, Analisis Kinerja Data Mining Algoritma C4.5 Dalam Menentukan Tingkat Minat Siswa Yang Mendaftar di Kampus ABC, KETIK, 2015.
[6] Yusheng Lin, et al, A Method of Satellite Network Fault Synthetic Diagnosis Based on C4.5 Algorithm and Expert Knowledge Database, IEEE, 2015
[7] H. Siahaan, H. Mawengkang, S. Efendi, A. Wanto, and A. P. Windarto, “Application of Classification Method C4 . 5 on Selection of Exemplary Teachers,” in IOP Conference Series, 2018, pp. 1–6.
[8] S. Sudirman, A. P. Windarto, and A. Wanto, “Data Mining Tools | RapidMiner: K-Means Method on Clustering of Rice Crops by Province as Efforts to Stabilize Food Crops In Indonesia,” IOP Conf. Ser. Mater. Sci. Eng., vol. 420, no. 12089, pp. 1–8, 2018