Determination of accuracy value in id3 algorithm with gini index and gain ratio with minimum size for split, minimum leaf size, and minimum gain

Randi Rian Putra*, Hanna Willa Dhany

Faculty of Science and Technology, Universitas Pembangunan Pancabudi Medan, Indonesia

*randirian@dosen.pancabudi.ac.id

Abstract. A process that explains and functions to distinguish data classes is called Classification. The use of Gain Ratio in ID3 algorithm is very influential on accuracy compared to the Gini Index, and if the higher the determination of the minimum size of split, the minimum leaf size and the minimum gain, the accuracy results will be greater in the gini index and gain ratio, but from both methods there is an accuracy difference of 27% for each accuracy test. In determining the minimum size of split 2, the minimum leaf size 2 and the minimum gain 0.4 with the value of accuracy at the gain ratio of 85.53% and gini index 58.67%. The minimum size of split 6, the minimum leaf size 3 and the minimum gain 0.8 with accuracy values at gain ratio 64.67% and gini index 60.67%. Then the minimum size of split 12, the minimum leaf size 6 and the minimum gain of 0.16 with the value of accuracy at the gain ratio of 86.00% and the index of 68.00%. While the minimum size of split 48, the minimum leaf size 24 and the minimum gain of 0.64 with the value of accuracy at the gain ratio of 95.33% and the index value of 72.00%. Then the Gain ratio which produces the highest accuracy value compared to the gini index.

1. Introduction

The use of the Decision Tree Model is one of the classification techniques as part of Data Mining science. Data mining does extract knowledge about data. For this classification technique, the ID3 algorithm was first developed by Ross Quinlan at the Sydney University 1975 in a book entitled: Machine Learning, vol. 1, no.1. ID3 is based on the Concept Learning System (CLS) algorithm.

The advantages of using this Decision Tree model besides being easy to understand can also be used to find rules or conditions that can be used as criteria that are useful for obtaining analysis in a decision-making process. To create a classification and identification tool for data with the Decision Tree model using ID3 algorithm and get better results with the application of techniques used to the model in the problem.

Data classification is categorizing data into different categories according to the rules.In this classification aims to change the structure of the object example. Classification algorithms are made from training sets and build models and models used to classify new objects. The decision tree evaluates the strength of the classification by analyzing the performance and results of the analysis (Patel, 2012).

Classification of data objects based on objects that have been specified in data. There are many classification algorithms but the most commonly used *Decision Tree* (Seema, 2012). *Decision Tree* algorithm is one of the most important classification measures in *Data Mining*. Classification is one type of grouping which is a flow diagram like a tree structure,

where each internal node shows a test on each attribute, each branch represents the results of the test, and each leaf node represents the class. The model for classifying a note to find the leaf root pathway to measure the attribute test and leaf attribute is the result of the classification used by Decision Tree (Qin-yun, 2016).

2. Problems Identification

From the background of the problem described earlier, the writer takes the formulation of the problem that the need for the classification process of the *Decision Tree* ID3 method by determining the gain ratio and the Gini Index in order to find out which method is more accurate in carrying out the classification.

3. Research Methods

Classification is usually related to class category forecasting and classifying data or constructing a model based on training data to define and class values in a class of attributes and using new data classes. Classification is often used in the areas of credit approval, target marketing, medical diagnosis, and analysis of the effectiveness of a decision.Classification step by describing a set of predetermined classes and using a model that serves to classify tuples of data whose class labels are unknown. These models are presented as classification rules, decision trees, or mathematical formulas. The various classifications that are often used are *Decision Tree, Bayesian Network, Adaptive Bayesian Network, Naïve Bayes,* and so on.

A. Selection of Attributes and Tree Formation

In selecting attributes to be the *root node* or internal *node* as a *test* attribute based on the size of impurity of each attribute. *Impurity* measures that are commonly used are information *gain, gain ratio* and Gini Index. The attribute that has the highest impurity value will be the test attribute.

B. Information Gain attribute A (Gain (A))

A measure of correlation in the parametric model that displays dependence on 2 random variables X and Y .

C. Gain Ratio

Gain ratio modifies from information *gain* to reduce the bias attribute that has many branches. The nature of the *gain ratio* is a large value if the data spread evenly and have a small value if all data is entered in 1 branch.

D. Decision Tree

The Decision Tree is a flow diagram that is shaped like a tree structure with each internal node testing the attributes, in its branches performing output from the test and *leaf nodes* performing class grouping or class distribution. *The* most important *node* is called the *root node* or root. A *root node* has several *edges* out but does not have an entry *edge*. The internal *node* will have one entry *edge* and several exit *edges,* while the *leaf node* will only have *one edge* without having an exit *edge*.

The Decision Tree *is* used to classify data that still does not know its class to exist classes. The path of testing data is the first step that is passed through the *root node* and the last is the *leaf node* that will predict the class for the data that has been concluded.

E. C4.5 algorithm

This algorithm develops in order to improve the ID3 algorithm. This algorithm is based on *binary* decisions as seen in CL. So in addition to having characteristics such as ID3, C4.5 has some different characteristics that are also characteristic of C4.5 as improvements to ID3 it can handle numeric attributes, can handle *missing value,* do *pruning* to obtain the most efficient models, used a *gain ratio* to determine the best type of *split*.

F. Characteristics of Decision Tree

The characteristics of the *decision tree* are as follows:

- The Decision tree is a nonparametric approach to building a classification model
- A technique developed to establish a *decision tree* that is used to build models quickly from large size *training sets*.
- *Decision tree* small-sized *tree* that is relatively easy for interpretation.
- An easy description is given of learning discrete value functions in the *decision tree*.
- The number of attributes does not reduce the accuracy of the *decision tree.*
- The number of *records* is smaller because it uses a top-down approach. While the *leaf node* is too small the number of *records* that makes decisions statistically.
- A sub tree that is doubled multiple times in a *decision tree* but does not cause the *decision tree to* be more complex and more difficult to interpret (Sibaroni, 2008).

G. Basic Concept of ID3 Algorithm

Calculations to produce a piece of information are quite difficult for the algorithm. The ID3 algorithm searches using the Decision Tree search and involves adding vertices to existing trees. The function of the ID3 algorithm is as follows:

- A mathematical algorithm for building decision tree models.
- Created by Ross Quinlan J. in 1979.
- The information theory used was discovered by Shannon in 1948.
- Build trees from top to bottom by not backing down
- Profit information is used to select the attributes that are most useful for classification.

An example of the ID3 algorithm is Target Attribute, Attributes which is written as:

- Create the root node for the tree
- If all examples are positive, return the single Root tree node, with label = +
- If all are negative, return the single Root tree node, with the label = -
- If the number of single nodes is Root Tree, with the label = general value most of the target attributes in the example

4. Results and Discussion

In testing the system, several methods are needed to obtain a more efficient comparison of results in analyzing accuracy in each data set. In this study, the author analyzed the *Decision Tree ID3* Algorithm

Table 4.1 Iris dataset

	Attribute				class
	sepal length	*sepal width*	*petal length*	*petal width*	
Amount of data	150	150	150	150	Setosa
					Versicolor
					Virginica

By testing the Rapid miner software, the authors determine the attributes based on the Gini Index and gain ratio. The outputs from testing the accuracy of the Iris dataset using the Gini Index, Minimum Size for Split, Minimal Leaf Size, and Minimum Gain are as follows:

Table 2. Results of Accuracy Values

	Minimal Size for Split	Minimal Leaf Size	Minimum Gain	Accuracy
Gini index	2	2	0.4	58.67% +/- 16.81%
gain ratio				85.33% +/- 8.33%
Gini index	6	3	0.8	60.67% +/- 16.72%
gain ratio				84.67% +/- 8.46%
Gini index	12	6	0.16	68.00% +/- 15.14%
gain ratio				86.00% +/- 8.14%
Gini index	24	12	0.32	72.00% +/- 15.43%
gain ratio				95.33% +/- 4.27%
Gini index	48	24	0.64	72.00% +/- 15.43%
gain ratio				95.33% +/- 4.27%

5. Conclusion

From the test results are shown in Table 4.2 it can be concluded that the use of Gain Ratio on the ID3 algorithm greatly affect the accuracy compared with the Gini Index, and if the higher determination of the minimal size of the split, minimal leaf size, and a minimum gain is the result accuracy will be the greater the Gini Index and gain ratio, but from the two methods there is an accuracy difference of 27% for each test of accuracy. The researcher suggested using the Gain Ratio to find the accuracy value of the data in the ID3 algorithm testing.

References

[1]. AA Nababan, OS Sitompul, and sincere Attribute Weighting Based on nearest neighbor Using Heading Gain Ratio IOP Conf. Series: Journal of Physics: Conf. Series 795 (2018) 012007

[2]. Dai Qin-Yun,. Zang Chun-Ping., Wu Hao. 2016. *Research of Decision Tree Classification Algorithm in Data Mining* . Dept. of Electric and Electronic Engineering,

Shijiazhuang Vocational and Technology Institute. China
[3]. Manasi M. Phadatare, Sushma S. Nandgaonkar. 2014. *Uncertain Data Mining using Decision Tree and Bagging Technique.* Department of Computer Engineering, India.
[4]. Priyadarsini, RP, Valarmathi, ML, Sivakumari, S. 2011. Gain Ratio Based Feature Selection Method for Privacy Preservation. *ICTACT Journal on Soft Computing* 1 (4): 201-205.
[5]. Seema., Rathi Monika., Mamta. 2012. *Decision Tree: Data Mining Techniques* . Department of Computer Science Engineering. India.
[6]. Turban Efraim., E Jay., Aronson., Liang Ting-Peng. 2005. *Decision Support System and Intelligent System*. Andi Offset.