JURNAL IPTEKS TERAPAN
Research of Applied Science and Education V14.i2 (99-105)

ISSN : 1979-9292

E-ISSN : 2460-5611

ACCURACY OF C4.5 ALGORITHM BASED GAIN AVERAGE VALUES IN PREDICTING STUDENT VALUES

Aminuddin Indra Permana
Sains dan Teknologi Universitas Pembangunan Pancabudi Medan
E-mail: aminuddin@dosen.pancabudi.ac.id

Abstract

C4.5 algorithm still has weaknesses in making predictions or classifying data if the classes used in large quantities can cause increased decision-making time. Then we need an approach to improve the performance of the C4.5 algorithm in the split attribute selection process is to use the average gain value that is applied to help predict students who will become the overall champion. In research conducted on Student Value is done by producing predictions from the C4.5 method by doing the highest level of accuracy that is good. From the results of the analysis that improving the performance of the C4.5 algorithm in the split attribute selection process is to use the average gain value applied. The success in predicting using the C4.5 method using Student Grades increased by 66.3%

Keyword : C4.5, Average Gain, Split Atribut, Akurasi, Siswa.

INTRODUCTION

Information Mining is portion of the KDD (Information Disclosure in Database) prepare which comprises of different stages, for illustration in information determination, pre-processing, change, information mining and assessment comes about. [1]. Classification of information objects based on objects that have been decided in a information. there are numerous classification calculations but Choice Tree is most frequently utilized. [2]. Choice Tree calculation is one of the foremost vital classification measures in information mining. The classification is one sort of gathering specifically flowcharts like a tree structure, where each inner hub appears a test on each quality, each department speaks to the comes about of the test, and each leaf hub speaks to the lesson. The demonstrate to classify a note to discover the leaf root way to degree the leaf quality and property test is the result of the classification utilized by Choice Tree. [3]. C4.5 calculation could be a choice tree that's utilized for classification with the concept of data entropy. C4.5 calculation employments the part criteria of the altered ID3 called Pick up Proportion [4]. ID3 calculation employments Data Pick up (IG) for trait part criteria, whereas C4.5 calculation employments Pick up Proportion (GR), where the quality that has the most noteworthy pick up is chosen as the root (root). The stages of the C4.5 calculation are calculating the Entropy esteem, calculating the esteem of the Pick up Proportion for each trait, the quality

Submited : 03 Jul 2020 – Accepted : 11 Jul 2020 – Published : 12 Jul 2020

ISSN : 1979-9292

E-ISSN : 2460-5611

that has the most noteworthy Pick up Proportion is chosen to be the root (root) and the quality which has the Pick up Proportion esteem lower than the root (root) is chosen to be the department (branches), calculate once more the esteem of the Pick up Proportion of each trait by not counting the trait chosen to be the root (root) within the past organize, the trait that has the most noteworthy Pick up Proportion is chosen to be a department (branches), rehashing steps 4 and 5 until the coming about Pick up esteem = for all remaining qualities.

Within the choice tree approach, pruning is the method of cutting or killing a few branches (hubs) that are not required. Pointless hubs can cause boisterous information and less significant highlights [5] This causes the measure of the choice tree to be very huge, called over-fitting. As a result there's an lopsidedness of information so that the level of precision is moo [6]. Two strategies of pruning. The primary strategy is called heterogeneous-cost touchy learning (HCSL) by adjusting the average-gain part attribute[7].Which is duplicated by the distinction in misclassification (misclassification fetched of the property some time recently it is part and after it is part). The moment pruning strategy is to utilize a edge esteem. Both of the pruning strategies were tried utilizing the Choice Tree classification demonstrate with part criteria of the ID3 and C4.5 (heterogeneous fetched) calculations on the six datasets and concluded that both of the proposed pruning strategies can be utilized for the choice tree classification model [8].

Inside the choice tree approach, pruning is the strategy of cutting or slaughtering many branches (centers) that are not required. Inconsequential centers can cause disorderly data and less critical highlights [5] This causes the degree of the choice tree to be exceptionally colossal, called over-fitting. As a result there's an lopsidedness of data so that the level of accuracy is moo [6]. Two procedures of pruning. The essential methodology is called heterogeneous-cost unstable learning (HCSL) by altering the average-gain portion attribute[7].Which is copied by the qualification in misclassification (misclassification brought of the property a few time as of late it is portion and after it is portion). The minute pruning technique is to utilize a edge regard. Both of the pruning methodologies were attempted utilizing the Choice Tree classification illustrate with portion criteria of the ID3 and C4.5 (heterogeneous brought) calculations on the six datasets and concluded that both of the proposed pruning methodologies can be use.

METHOD

C4.5 is a supervised learning classification algorithm to form a decision tree of data developed by J. Ross Quinlan as a development of the ID3 algorithm. If ID3 (Iterative Dichotomiser 3) uses entropy for split criteria, whereas C4.5 decision tree uses modified split criteria called Gain Ratio (Mitchael, 1997) in the process of selecting split attributes. Split attribute is the main process in forming a decision tree. C4.5 algorithm can work on continuous variables and missing values

Submited : 03 Jul 2020 – Accepted : 11 Jul 2020 – Published : 12 Jul 2020

JURNAL IPTEKS TERAPAN
Research of Applied Science and Education V14.i2 (99-105)

ISSN : 1979-9292

E-ISSN : 2460-5611

[8] Attributes that have the highest Gain Ratio chosen [9].

C4.5 uses two heuristic approaches to test the probability ranking, namely: (1) Information gain, minimizing the total entropy of the {Si} subset where bias occurs when tested with numerical data. (2) Gain ratio, division of Information gain by entropy information of each attribute. The C4.5 algorithm is one of the decision tree variants that is similar to a flowchart structure, each of which is an internal node expressed as a test attribute. Each branch represents the output of the test and each node (leafnode) determines the class label. The topmost node of a tree is the root node. C4.5 algorithm uses information gain as a determinant of root, internal and leaf nodes. The stages of the C4.5 algorithm are as follows[10].

(1) calculates the Entropy value on each attribute

$$Entropy(S) = \sum_{i=1}^{n} - pi * \log_2 pi$$

Where:

- *S = Case Set*
- *n = Number of Partitions S*
- *pi = Proportion of Subset of Spada i-partition*

(2) calculate the Information Gain value for each attribute:

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

Where:

- S = Overall Dataset
- A = Subset Attribute
- n = Number of Attribute Partitions A
- | Si | = Size of Subset of the dataset owned by attribute A on the i-th partition
- $|S|$ = Size of Number of Cases in a Dataset

(3) calculates the Split Information value for each attribute:

Where:

- D = Overall Dataset
- A = Subset Attribute
- v = Number of Attribute Partitions A
- | Dj | = Size of Subset of Dataset that belongs to the jth partition A attribute
- | D | = Size of Number of Cases in Dataset

(3) calculate the Gain Ratio value for each attribute:

$$Gain\ Ratio\ (A) = \frac{Gain\ (A)}{SplitInfo\ (A)}$$

(3) the attribute that has the highest Gain Ratio is selected to be the root (splitting-attribute) and the attribute which has a Gain Ratio value lower than the root (root) is chosen to be branches (branches),
(4) calculate again the Gain Ratio value of each attribute by not including the selected attribute as the root (root) in the previous stage,
(5) attributes that have the highest Ratiot Gain are selected as branches.
(6) repeats steps 4 and 5 until the resulting Gain value = 0 for all remaining attributes

Drainage testing using the Confusion Matrix is tabulated into a table called a confusion matrix (Witten & Frank, 2005). Confusion Matrix is a good or bad parameter for classifying test data in different classes, namely positive and negative classes (two-class prediction). The following table 1 explains the Confusion Matrix (two-class prediction):

Submited : 03 Jul 2020 – Accepted : 11 Jul 2020 – Published : 12 Jul 2020

JURNAL IPTEKS TERAPAN
Research of Applied Science and Education V14.i2 (99-105)

ISSN : 1979-9292

E-ISSN : 2460-5611

Tabel 1 *Confusion Matrix (two-class prediction)*

Two-Class Prediction		**Predicted Class**	
		Yes	No
Actual Class	Yes	True Positive	False Negative
	No	False Positive	True Negative

Table 1 explains the parameters of the 2 class classification model namely yes and no. True Positives (TP) and True Negatives (TN) are the number of classifications that are true. The False Positive (FP) parameter will be concluded when the resulting prediction is incorrect or has a yes (positive) value when the expected prediction is no (negative). Conversely, the False Negative Parameter (FN) will be concluded when the resulting prediction is incorrect or has no (negative) value when the expected prediction is yes (positive). The result of the Confusion Matrix parameter is accuracy.

The Confusion Matrix equation for calculating the accuracy value is

$$\frac{TP + TN}{TP + TN + FP + FN}$$

TP (True Positive) is the amount of data in the class yes whose prediction class results are indeed classified into the actual class that has a value of yes

TN (True Negative) is the amount of data in class no whose prediction class results are indeed classified into the actual class whose value is no

FP (False Positive) is the amount of data that is actually included in the actual class which has no value but the results of the prediction class are classified into classes that have a value of yes.

FN (False Negative) is the actual amount of data included in the actual class that has a value of yes but the results of the prediction class are classified into classes of value null.

RESULT

1. Test Result

In the process of forming the decision tree C4.5 classification model, the results of preprocessing data from the Student dataset obtained by 167 observational data were then divided into 90% of the data as training data and 10% of the data as test data.

In this study, only two preprocessing processes were carried out. The first is handling missing values. Missing value on factors that have numerical value is replaced by the mean value of the factors in the same column. Whereas the missing value of the nominal value factor is replaced by the highest possible value of the factor in the same column. Next is the cleaning process carried out by removing data duplication so that the amount of observation data that was originally 170 records into 167 records. The next process is data normalization done by standardizing the data so that the interval or range of the data becomes more proportional by using the Z-Score method as follows.:

$$z = \frac{\mathrm{x} - \mu}{\sigma}$$

z: standard score, x: observation data, μ: mean per variable and σ: standard deviation per variable. The results of the

Submited : 03 Jul 2020 – Accepted : 11 Jul 2020 – Published : 12 Jul 2020

JURNAL IPTEKS TERAPAN
Research of Applied Science and Education V14.i2 (99-105)

ISSN : 1979-9292

E-ISSN : 2460-5611

Z-score are data with mean = 0 and standard deviation = 1.

Put simply, the Z-score process is: each observation data on a variable is reduced by the mean of the variable and divided by its standard deviation (in other words, each row per column is reduced by the column's mean, divided by the same standard column deviation).

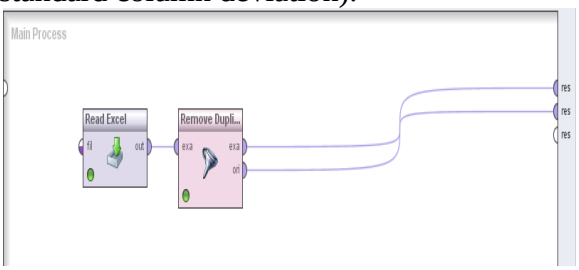

Figure 1: Remove Data Duplicate (Rapidminer) Process

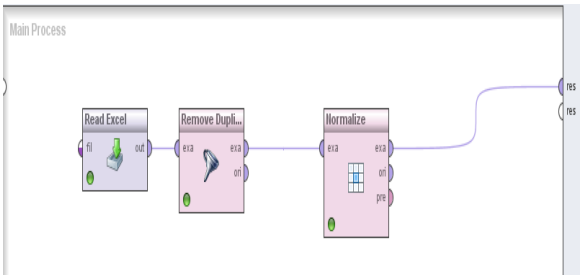

Figure 2: Process Normalize Data (Rapidminer)

This study seeks to apply the decision tree classification technique C4.5 to classify the dominant factors that influence student grades, in this case the Student Value dataset is used as trial data. The decision tree classification technique C4.5 uses two heuristic approaches to test the probability ranking, namely: (1) Information gain, minimizing the total entropy of the {Si} subset where bias occurs when tested with numerical data. (2) Gain ratio, division of Information gain by entropy information of each attribute.

Tabel 2 Hasil Data *Preprocessing*

ID Siswa	X1	X2	X3	X4	X5	X6	X7	X8	Target
1	89	96	96	93	88	84	96	127	1
2	88	95	95	90	90	85	92	115	1
3	83	93	94	87	89	82	90	109	1
4	89	90	92	83	89	85	86	118	1
5	83	90	86	89	76	82	91	118	1
6	80	88	90	85	78	82	91	118	1
7	82	84	84	89	91	84	90	100	1
8	82	87	80	90	79	85	87	106	1
9	79	88	90	81	84	81	86	118	1
10	84	78	85	89	83	84	90	112	1
11	84	85	87	80	83	83	85	82	1
12	78	92	85	78	81	83	91	100	1
13	82	84	84	82	82	82	90	106	1
14	81	74	89	84	75	81	87	94	1
15	81	88	81	78	75	81	89	100	1
16	75	80	91	80	77	83	84	109	1
17	82	78	86	78	75	81	89	91	1
18	75	84	80	81	75	81	86	93	1
19	81	74	84	80	84	82	83	112	1
…	…	…	…	…	…	…	…	…	…
167	70	76	70	76	75	78	78	93	2

2. Testing accuracy using Confusion Matrix

After obtaining the C4.5 classification model in the form of rules, then the classification model is then tested using Student Value test data.

Based on the results, we get 1 record of True Positive data, because only those records have the same number of actual classes and class predictions. The amount of data that has a False Positive value is obtained by 1 record. The amount of data that has False Negative value is obtained by 2 records. The amount of data that has a True Negative value is 21 records. So we can get the evaluation results of the

Submited : 03 Jul 2020 – Accepted : 11 Jul 2020 – Published : 12 Jul 2020

JURNAL IPTEKS TERAPAN
Research of Applied Science and Education V14.i2 (99-105)

ISSN : 1979-9292

E-ISSN : 2460-5611

C4.5 classification model from the Student Value dataset using the confusion matrix as in the following table:

Table 3 *Confusion Matrix C4.5*

Classification Performance		**Predicted Class**
Actual Class	Predicted. Class 2	Predicted. Class 1
Actual. Class 2	26 (True Positive)	2 (False Negative)
Actual. Class 1	1 (False Positive)	21 (True Negative)

Based on the above table, then proceed with calculating the value of Accuracy, Particularity and classification error rate (Classification_error) of the C4.5 classification model of the Student Value dataset, the following results of the calculation:

a. Accuracy $= \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} = \frac{26+21}{26+21+1+2} = \frac{47}{50} = 0.94*100\% = 94\ \%$

The level of closeness between the class prediction with the actual class or the number of correct class predictions from the C4.5 classification model is 94%

b. Classification Error $= \frac{FP+FN}{TP+TN+FP+FN} = \frac{1+2}{26+21+1+2} = \frac{3}{50} = 0.06 * 100\% = \mathbf{6}\ \%$

CONCLUSION

In research conducted on Student Value is done by producing predictions from the C4.5 method by doing the highest level of accuracy that is good. From the results of the analysis that improving the performance of the C4.5 algorithm in the split attribute selection process is to use the average gain value applied. Success in predicting using the C4.5 method using Student Grades.

In further research the author hopes to develop a program system in predicting even greater data, because there are still shortcomings in this study so that it must be refined in future research to obtain better results than before and systematically again. Therefore the authors expect this research to continue by using other algorithms and obtain the final results in accordance with the wishes. Hopefully get greater accuracy and produce a better prediction concept.

REFERENCE

[1] Han, J., Kamber, M. & Pei, J. 2012.*Data Mining: Concepts and Techniques*. 3rd Edition. Morgan Kaufmann Publishers: San Francisco.

[2] Hussain, H., Quazilbash. N.Z., Bai. S. &Khoja, S. 2015. Reduction of Variables for Predicting Breast Cancer Survivability Using Principal Component Analysis.*International Conference on Computer-Based Medical Systems*, pp. 131-134.

[3] Kavitha, K. V., Tiwari, S., Purandare, V. B., Khedkar, S., Bhosale, S. S., Unnikrishnan, A. G. (2014). Choice of wound care in diabetic foot ulcer: A practical approach. World J

Submited : 03 Jul 2020 – Accepted : 11 Jul 2020 – Published : 12 Jul 2020

ISSN : 1979-9292

E-ISSN : 2460-5611

Diabetes.5(4):546–56.doi: 10.4239/wjd.v5.i4.546.

[4] Kavitha, R. & Kannan, E. 2016. An Efficient Framework for Heart Disease Classification using Feature Extraction and Feature Selection Technique in Data Mining. *International Conference on Emerging Trends in Engineering, Technology and Science(ICETETS)*, **pp.** 1-5.

[5] Kotu, V. & Deshpande, B. 2015. *Predictive Analytics and Data Mining*. Morgan Kaufmann Publisher: San Francisco.

[6] Larose, D.T. 2005.*Discovering Knowledge in Data: An Introduction to Data Mining, John Willey & Sons*. Inc. pp. 129-240

[7] Maimon, O. dan Last, M. 2000. Knowledge Discovery and Data Mining, The Fuzzy network (IFN) Methodology. Dordrecht: Kluwer Academik.

[8] Quinlan, J.R. 1992. C4.5 Programs for Maching Learning. San Mateo, CA: Morgan Kaufmann.

[9] Raviya. Kaushik H & Gajjar, Biren. 2013. *Performance Evaluation of Different Data Mining Classification Algoritma Using WEKA. Indian Journal of Research*. Volume. 2. Issue.1. ISSN: 2250-1991.

[10] Sahu, Mridu., Nagwani. N.K., Verma Shrish., Shirke. Saransh. 2015. *Performance Evaluation of Different Classifier for Eye State Prediction Using EEG Signal. International Journal of Knowledge Engineering*, Volume.1, No.2.

[11] Seema., Rathi Monika., Mamta. 2012. *Decision Tree: Data Mining Techniques*. Department of Computer Science Engineering. India.

[12] Sivapriya, T. R.&Nadira, B. K. 2013.Hybrid Feature Selection for Enhanced Classification of High Dimensional Medical Data.*International Conference on Computational Intelligence and Computing Research*, pp. 1-4.

[13] Steinbach, M., Karypis, G. & Kumar, V. 2000. A comparison of document clustering techniques. *KDD Workshop on Text Mining*, **pp.** 525.

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

Submited : 03 Jul 2020 – Accepted : 11 Jul 2020 – Published : 12 Jul 2020