Average Gain Ratio and Correlation-Based Feature Selection for Imprecise Classification in Algorithm C4.5

Saeful Amri
Department of Data Science, Universitas Muhammadiyah Semarang, Indonesia
saefulamri@unimus.ac.id (corresponding author)

M. Alharis
Department of Statistics, Universitas Muhammadiyah Semarang, Indonesia
alharis@unimus.ac.id

Edy Winarno
Department of Information Technology, Universitas Muhammadiyah Semarang, Indonesia
edywin@unimus.ac.id

Astrid Novita Putri
Department of Informatics Engineering, Universitas Semarang, Indonesia
astrid@usm.ac.id

Received: 8 January 2025 | Revised: 17 February 2025 and 23 April 2025 | Accepted: 27 April 2025

 DOI: https://doi.org/10.48084/etasr.10165

ABSTRACT

The attribute split in the C4.5 algorithm has proven successful in building a classification in the form of a tree to facilitate understanding and interpretation. However, the attribute split process tends to choose attributes with high values, even though they do not necessarily play a major role in the classification results, and does not consider the correlation of attributes with labels, affecting classification performance. Average Gain Ratio (AGR) has been proven to overcome weaknesses and problems in the split process. Correlation-Based Feature Selection (CBFS) is also used in the split process to calculate the correlation of attributes with labels. This study uses the AGR method to overcome the problem of attribute split criteria and the CBFS method to select attributes that correlate with labels to increase the performance of the C4.5 algorithm. The process of selecting the split attributes using AGR and the comparison with the CBFS method was shown to improve the performance of the C4.5 classifier, as indicated by the average results for accuracy (87%), sensitivity (91%), G-mean (85%), AUC (84%), and cost (3.67) on six UCI datasets.

Keywords-decision tree; C4.5 algorithm; attribute split; attribute correlation; imprecise classication

I. INTRODUCTION

Improving the efficiency and accuracy of data mining algorithms has always been a hot issue, especially in the field of machine learning [1]. In this field, classification is considered an important tool for decision support. Classification can be defined as a machine learning technique to predict group or class membership of data. Classification can be applied to support decisions in the fields of medicine, character recognition, astronomy, banking, and other fields [2]. Some algorithms that fall into the classification category are Naïve Bayes (NB), Neural Networks (NN), and Decision Tree (DT).

DTs are useful for classification due to their simple structure and easy interpretation [3], offering relatively good results and computational efficiency [4]. The DT algorithm successfully builds classifications to maximize accuracy [5]. The process of inducing a DT is called training, whereas the process of deducing is called prediction [6]. DTs have been applied in various fields, such as predicting the results of antibodies that are not suitable for kidney transplantation [7]. Soft computing techniques such as DT classifiers have been shown to be effective, offering considerable promise in the development of medical decision support systems for kidney, liver, and chronic pancreatitis diseases [8] and skin sensitivity prediction [9]. DTs have also been used in the field of natural minerals and other industrial fields [10].

ID3 is a DT algorithm introduced in 1986, with the C4.5 algorithm presented in 1993 as a further refinement [11]. Classic DT algorithms include C4.5, ID3, and CART (Classification and Regression Tree) [12]. The stages to form the C4.5 algorithm according to [13] are: (i) calculating the Entropy value, (ii) calculating the Info Gain value, and (iii) calculating the Gain Ratio value. The use of split attributes makes the C4.5 algorithm successful in building DTs and improving classification performance. However, split attribute criteria still have the disadvantage of tending to choose attributes with high values, although they do not necessarily play an important role in classification results [14]. Split attributes can also reduce costs and the risk of overfitting, but, on the other hand, increase the ability of the learning model to classify high-dimensional data in different classes [15]. According to [16], some researchers have successfully selected attributes with high values in the attribute split process, but this reduced classification performance. The split attribute process, which tends to select attributes with high values, needs to be taken into account.

In 2012, a Heterogeneous Cost-Sensitive Learning (HCSL) algorithm was proposed to improve the attribute split solution [17]. The HCSL method applies the Average Gain method in the attribute split process and the Redu_Mc method as the normalization formula of the heterogeneity problem caused by the cost mechanism and attribute information. Redu_Mc is used as a miner in the selection of attributes with high values in the split process. As a result, HCSL can reduce heterogeneity caused by cost mechanisms and attribute information. The study in [14] focused on the problem of split attributes, with an attribute selection method that measures the correlation of attributes with labels to improve classification performance. The split attribute method in this study uses Correlation-Based Feature Selection (CBFS) calculation to find the best attribute. Feature selection is a crucial step in the machine learning workflow, focusing on selecting the most relevant and informative features from the dataset to improve model performance [18]. According to [19], the selection of attributes can improve performance more efficiently compared to the applied model. In [20], the Long-Term Spectral Pseudo-Entropy (LTSPE) method was applied to improve performance in combination with other features, increasing accuracy and performance. In [21], features were explored and identified for their influence on automatic risk prediction.

The Randomly Selected DT method was proposed in [16]. The Atts method is used to select attributes with high values in the attribute split process. Atts consist of 2 types, namely Attbest and Attproper. Attbest is the calculation of attributes with the highest Gain Ratio results, while Attproper is the calculation of attributes with the lowest Gain Ratio results. The condition formed is if the random value (0,1) is less or equal to the β value, then the attribute goes into Attbest, otherwise, if the random value (0,1) is greater than the β value, then the attribute goes into Attproper. The process of determining the value of β randomly may also reduce the performance of the algorithm. It was also highlighted that a further study for the selection of split attributes can improve the performance of the new DT algorithm.

The C4.5 algorithm is a classification algorithm used to build a DT with the main process of split attributes. The process successfully builds a classification as a tree to make it easy to understand and interpret. However, the split attribute process still tends to choose high attribute values, even though these attributes do not necessarily play a major role in the classification results and do not consider the correlation of attributes with labels. This study proposes a new method for the attribute split process using the Average Gain Ratio value with the determination of attribute correlation based on CBFS in the C4.5 algorithm. Average Gain Ratio is used to overcome the problem of attribute split criteria that tend to select high attribute values, even though these attributes do not necessarily play a major role in classification results. In contrast, the CBFS method is used to select attributes that correlate with the label to increase the performance of the C4.5 algorithm. This study uses six public datasets from the UCI machine learning repository. The evaluation results are based on the confusion matrix, accuracy, sensitivity, cost, Geometric Mean (G-Mean), and Area Under the ROC Curve (AUC).

II. DATA COLLECTION

Six real-world datasets were chosen from the UCI database repository [22], namely Breast Cancer Wisconsin, Vote, Ionosphere, Parkinson's, Horse Colic, and Statlog Heart. The datasets have data ranging from 197 to 699, attributes ranging from 9 to 34, and binominal and polynomial type labels. Table I shows the principal characteristics of each dataset. Column N shows the number of instances in the datasets, column Feat has the number of features or attribute variables, column Num has the number of numerical variables, column Nom has the number of nominal variables, and column k has the number of cases or states of the class variable (always a nominal variable).

TABLE I. DATASETS EMPLOYED IN THE EXPERIMENTAL RESEARCH

Dataset	N	Feat	Num	Nom	k
Breast Cancer Wisconsin	699	9	0	9	2
Vote	435	16	16	0	2
Ionosphere	351	34	0	34	2
Parkinson	197	23	0	23	2
Horse Colic	368	23	16	7	2
Statlog Heart	270	13	8	5	2

III. METHODOLOGY

Figure 1 shows the stages of the proposed method on imprecise classification in the C4.5 algorithm.

A. Preprocessing

Data preprocessing is crucial and involves steps before analysis or modeling, aiming to clean, modify, and prepare the data.

B. AGR-CBFS

The split attribute problem in the C4.5 algorithm has attracted the interest of many researchers. The proposed method applies the AGR and CBFS (AGR-CBFS) for split attributes in the C4.5 algorithm. The AGR method is used to overcome the problem of attribute split criteria that tend to select attributes with high values. In contrast, the CBFS method

is used to select attributes that correlate with the label. Algorithm 1 describes the AGR-CBFS method.

Algorithm AGR-CBFS
1: Insert Dataset.
2: Perform preprocessing. If there is a missing value, then handling is done by replacing missing values to fill in the empty data and discretizing by binning to convert numeric attribute data to nominal data (by adjusting the bin value according to the original data).
3: Calculate the Entropy value of the dataset
$Entropy\ (S) = \sum_{i=1}^{n} - pi * log_2\ pi$
S = Case
n = Sum of partitions S
pi = Proportion of Si (class) to S
4: Calculate the Entropy value of each class on the attribute
$Entropy\ (S, A) = \sum_{i=1}^{n} - pi * log_2\ pi$
S = Set Attribute
n = sum of partitions T
pi = Proportion of Si (class) to S
5: Calculate the Gain value of each attribute
$Gain(S, A) = Entropy\ (S) = - \sum_{i=1}^{n} \frac{|Si|}{|S|} * Entropy\ (Si)$
S = Set Case
A = Attribute
n = sum of partitions of attribute A
$|Si|$ = Proportion of Si to S
$|S|$ = Number of cases in S
6: Calculate the Split Info value
$$SplitInfo_A(D) = -\sum_{j=1}^{n} \left|\frac{Dj}{D}\right| * \log_2 \left|\frac{Dj}{D}\right|$$
7: Calculate the Gain Ratio Value
$$Gain\ Ratio = \frac{Gain\ (S,A)}{Split\ Info\ A(D)}$$
8: Calculate the Average Gain Ratio value of each attribute
$$AverageGainRatio = \frac{(2Averagegain(Ai,T)-1)xReduce\_Mc(Ai)}{TC(Ai)+1}$$
9: Calculate the MCs value of each attribute
$$MCs = \frac{k\bar{r}ci}{\sqrt{k+k(k-1)\bar{r}ii}} - \lambda \frac{\sum_{i=1}^{k} \text{Ctest (Ai)}}{\text{k}}$$
k = Attribute subset size
$kris$ = Average correlation of feature-classes
rii = Average intercorrelation of features
λ = Parameters for considering cost in the evaluation function (parameter λ is adjusted to vary from 0.0 to 1)
$Ctest(Ai)$ = Test Cost of feature Ai
$Ctest(Cost\ test)$ value is considered to minimize misclassification cost.
10: Calculate the Performance Average Gain Ratio value of each attribute
$AverageGainRatio = (2AG) - 1) * MC$
11: DT formation
12: Confusion matrix, Accuracy, Sensitivity, G-Mean, and AUC.

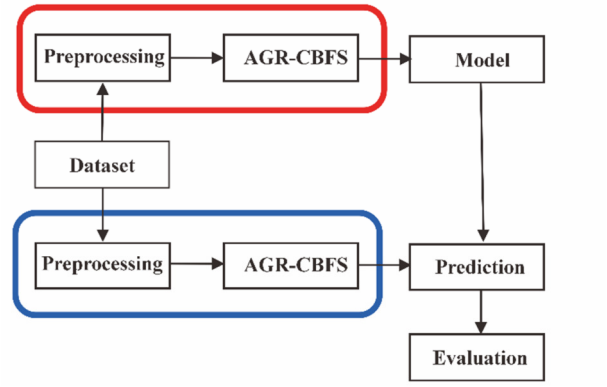

Fig. 1. Proposed method.

IV. EXPERIMENTS AND RESULTS

The SGR method is used to overcome the problem of attribute split criteria that tend to select high attribute values. In contrast, the CBFS method is used to select attributes that correlate with the label. To determine the performance of the proposed method, AGR-CBFS was compared with C4.5, HSCL, CBFS, and RSDT. All these methods were tested on the split attribute process to determine which has the best performance.

A. Accuracy Results

Table II shows a comparison of the accuracy of the compared methods, where it can be observed that AGR-CBFS has the highest average accuracy.

TABLE II. COMPARISON OF ACCURACY

Dataset	DT model				
	C4.5	HCSL	CBFS	RSDT	AGR-CBFS
Breast Cancer Wisconsin	94.56%	91.56%	91.42%	90.99%	94.56%
Vote	95.86%	93.10%	94.71%	93.56%	96.15%
Ionosphere	85.19%	79.77%	87.18%	84.90%	87.18%
Parkinson	85.13%	75.90%	84.10%	82.56%	88.22%
Horse Colic	64.95%	61.41%	63.59%	63.86%	77.98%
Statlog Heart	75.19%	74.07%	75.19%	74.07%	77.89%

B. Sensitivity Results

Table III shows the sensitivity comparison of the compared methods, where it can be observed that AGR-CBFS has the highest average sensitivity.

TABLE III. COMPARISON OF SENSITIVITY

Dataset	DT model				
	C4.5	HCSL	CBFS	RSDT	AGR-CBFS
Breast Cancer Wisconsin	96.94%	96.29%	96.29%	96.07%	97.38%
Vote	94.12%	90.40%	94.38%	96.14%	97.37%
Ionosphere	97.78%	79.56%	87.56%	84.89%	91.89%
Parkinson	91.67%	84.72%	93.06%	90.97%	93.55%
Horse Colic	93.03%	36.29%	29.84%	28.23%	85.74%
Statlog Heart	66.67%	66.67%	67.48%	66.67%	75.00%

C. G-Mean Result

Table IV shows a comparison of the G-Mean of the methods, where AGR-CBFS has the highest average G-Mean.

TABLE IV. COMPARISON OF G-MEAN

Dataset	DT model				
	C4.5	HCSL	CBFS	RSDT	AGR-CBFS
Breast Cancer Wisconsin	0.934	0.892	0.889	0.884	0.922
Vote	0.955	0.927	0.948	0.929	0.961
Ionosphere	0.783	0.799	0.870	0.849	0.863
Parkinson	0.782	0.657	0.740	0.732	0.826
Horse Colic	0.300	0.519	0.491	0.481	0.748
Statlog Heart	0.739	0.730	0.742	0.730	0.764

D. AUC Results

Table V shows a comparison of the AUC of the compared methods, where it can be observed that AGR-CBFS has the highest average AUC.

TABLE V. COMPARISON OF AUC

Dataset	Decision Tree Model				
	C4.5	HCSL	CBFS	RSDT	AGR-CBFS
Breast Cancer Wisconsin	0.935	0.894	0.892	0.887	0.923
Vote	0.955	0.927	0.948	0.930	0.961
Ionosphere	0.802	0.799	0.870	0.849	0.845
Parkinson	0.792	0.679	0.759	0.749	0.840
Horse Colic	0.510	0.552	0.553	0.550	0.750
Statlog Heart	0.705	0.733	0.746	0.696	0.749

E. Cost Result

Table VI shows the cost comparison of the methods, where AGR-CBFS has the highest average cost.

TABLE VI. COMPARISON OF COST

Dataset	Decision Tree Model				
	C4.5	HCSL	CBFS	RSDT	AGR-CBFS
Breast Cancer Wisconsin	4.00	5.00	5.00	5.00	4.00
Vote	6.00	7.00	6.00	7.00	5.00
Ionosphere	4.00	4.00	3.00	4.00	3.00
Parkinson	1.00	1.00	1.00	1.00	1.00
Horse Colic	3.00	3.00	3.00	3.00	2.00
Statlog Heart	9.00	8.00	7.00	9.00	7.00

V. DISCUSSION

C4.5 is a classification algorithm to build DTs with the main process in the form of split attributes. The process is successful in building classifications in the form of trees to be easy to understand and interpret. However, the split attribute process still tends to choose attributes with high values, although they do not necessarily play a major role in the classification results and do not consider the correlation of attributes with labels, thus affecting the classification performance. AGR is used to overcome the problem of split attribute criteria that tend to choose high attribute values. The CBFS method is used to select attributes that correlate with the label, increasing the performance of the C4.5 algorithm. Table VI illustrates a performance comparison with previous studies. The proposed method achieved the best accuracy (0.87%), sensitivity (0.91%), G-Mean (0.85), AUC (0.84%), and Cost score (3.67%)

TABLE VII. ACCURACY, SENSITIVITY, G-MEAN, AUC, AND COST SCORE COMPARISON

No.	M	Results				
		Acc	Sens	G-Mean	AUC	Avg. cost
1	C4.5	0.83	0.90	0.75	0.78	4.50
2	HCSL	0.79	0.76	0.75	0.76	4.67
3	CBFS	0.83	0.78	0.78	0.79	4.20
4	RSDT	0.82	0.77	0.77	0.78	4.83
5	AGR-CBFS	0.87	0.90	0.85	0.84	3.67

VI. CONCLUSION

C4.5 is a classification algorithm to build a DT in the form of split attributes. This process aims to build a classification in the form of a tree to be easy to understand and interpret. However, the split attribute process still tends to choose attributes with high values, even though they do not necessarily play an important role in the classification results, and does not take into account the correlation of attributes with labels. The AGR method is used to overcome the problem of attribute split criteria that tend to choose high attribute values. The CBFS method is used to select attributes that correlate with the label to increase the performance of the C4.5 algorithm. AGR-CBFS was tested on six datasets from the UCI Machine Learning Repository. The evaluation results show that the proposed method can solve the split attribute problem and improve the performance of the C4.5, HCSL, CBFS, and RSDT algorithms. This is demonstrated by the results, with average values of accuracy at 87%, sensitivity at 91%, cost at 3,67, G-Mean at 85%, and AUC at 84%. This research contributes to improving the classification performance of the C4.5 algorithm by applying the AGR and CBFS methods. Future research can further improve the classification performance of the C4.5 algorithm for general use.

REFERENCES

[1] H. B. Wang and Y. J. Gao, "Research on C4.5 algorithm improvement strategy based on MapReduce," *Procedia Computer Science*, vol. 183, pp. 160–165, 2021, https://doi.org/10.1016/j.procs.2021.02.045.

[2] K. N. Singh and J. K. Mantri, "A clinical decision support system using rough set theory and machine learning for disease prediction," *Intelligent Medicine*, vol. 4, no. 3, pp. 200–208, Aug. 2024, https://doi.org/10.1016/j.imed.2023.08.002.

[3] A. Ashfaq, N. Cronin, and P. Müller, "Recent advances in machine learning for maximal oxygen uptake (VO_2 max) prediction: A review," *Informatics in Medicine Unlocked*, vol. 28, 2022, Art. no. 100863, https://doi.org/10.1016/j.imu.2022.100863.

[4] O. Sagi and L. Rokach, "Explainable decision forest: Transforming a decision forest into an interpretable tree," *Information Fusion*, vol. 61, pp. 124–138, Sep. 2020, https://doi.org/10.1016/j.inffus.2020.03.013.

[5] I. D. Mienye, Y. Sun, and Z. Wang, "Prediction performance of improved decision tree-based algorithms: a review," *Procedia Manufacturing*, vol. 35, pp. 698–703, 2019, https://doi.org/10.1016/j.promfg.2019.06.011.

[6] L. Wang, Z. Zhang, X. Zhang, X. Zhou, P. Wang, and Y. Zheng, "A Deep-forest based approach for detecting fraudulent online transaction," in *Advances in Computers*, vol. 120, Elsevier, 2021, pp. 1–38.

[7] T. Shaikhina, D. Lowe, S. Daga, D. Briggs, R. Higgins, and N. Khovanova, "Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation," *Biomedical Signal Processing and Control*, vol. 52, pp. 456–462, Jul. 2019, https://doi.org/10.1016/j.bspc.2017.01.012.

[8] M. Alexiuk, H. Elgubtan, and N. Tangri, "Clinical Decision Support Tools in the Electronic Medical Record," *Kidney International Reports*, vol. 9, no. 1, pp. 29–38, Jan. 2024, https://doi.org/10.1016/j.ekir.2023.10.019.

[9] D. S. Macmillan and M. L. Chilton, "A defined approach for predicting skin sensitisation hazard and potency based on the guided integration of in silico, in chemico and in vitro data using exclusion criteria," *Regulatory Toxicology and Pharmacology*, vol. 101, pp. 35–47, Feb. 2019, https://doi.org/10.1016/j.yrtph.2018.11.001.

[10] V. A. Dev and M. R. Eden, "Formation lithology classification using scalable gradient boosted decision trees," *Computers & Chemical Engineering*, vol. 128, pp. 392–404, Sep. 2019, https://doi.org/10.1016/j.compchemeng.2019.06.001.

[11] S. Abolhosseini, M. Khorashadizadeh, M. Chahkandi, and M. Golalizadeh, "A modified ID3 decision tree algorithm based on cumulative residual entropy," *Expert Systems with Applications*, vol. 255, Dec. 2024, Art. no. 124821, https://doi.org/10.1016/j.eswa.2024.124821.

[12] C. C. Aggarwal, "Data Classification," in *Data Mining*, Springer International Publishing, 2015, pp. 285–344.

[13] M. Qiu, "Path Planning Algorithm and ID3 Decision Tree Model Application of Scenic Intelligent Navigation System," *Procedia Computer Science*, vol. 247, pp. 1187–1196, 2024, https://doi.org/10.1016/j.procs.2024.10.143.

[14] V. Bolón-Canedo, I. Porto-Díaz, N. Sánchez-Maroño, and A. Alonso-Betanzos, "A framework for cost-based feature selection," *Pattern Recognition*, vol. 47, no. 7, pp. 2481–2489, Jul. 2014, https://doi.org/10.1016/j.patcog.2014.01.008.

[15] S. Bakhshandeh, R. Azmi, and M. Teshnehlab, "Graph Based Feature Selection Using Symmetrical Uncertainty in Microarray Dataset," *Journal of Information Systems and Telecommunication*, vol. 7, no. 1, pp. 35–40, 2019.

[16] C. Qiu, L. Jiang, and C. Li, "Randomly selected decision tree for test-cost sensitive learning," *Applied Soft Computing*, vol. 53, pp. 27–33, Apr. 2017, https://doi.org/10.1016/j.asoc.2016.12.047.

[17] S. Zhang, "Decision tree classifiers sensitive to heterogeneous costs," *Journal of Systems and Software*, vol. 85, no. 4, pp. 771–779, Apr. 2012, https://doi.org/10.1016/j.jss.2011.10.007.

[18] Y. E. Touati, J. B. Slimane, and T. Saidani, "Adaptive Method for Feature Selection in the Machine Learning Context," *Engineering, Technology & Applied Science Research*, vol. 14, no. 3, pp. 14295–14300, Jun. 2024, https://doi.org/10.48084/etasr.7401.

[19] F. Mozaffari, I. R. Vanani, P. Mahmoudian, and B. Sohrabi, "Application of Machine Learning in the Telecommunications Industry: Partial Churn Prediction by using a Hybrid Feature Selection Approach," *Journal of Information Systems and Telecommunication (JIST)*, vol. 4, no. 44, Dec. 2023, Art. no. 331, https://doi.org/10.61186/jist.38419.11.44.331.

[20] M. R. Kahrizi, "Long-Term Spectral Pseudo-Entropy (LTSPE) Feature," *Journal of Information Systems and Telecommunication*, vol. 6, no. 4, pp. 204–208, 2018, https://doi.org/10.21227/H2G05K.

[21] M. Irfan, S. Basuki, and Y. Azhar, "Giving more insight for automatic risk prediction during pregnancy with interpretable machine learning," *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 3, Jun. 2021, https://doi.org/10.11591/eei.v10i3.2344.

[22] C. L. Blake and C. J. Merz, "UCI repository of machine learning databases," 1998.