Sensors and Materials, Vol. 36, No. 4 (2024) 1461–1471
MYU Tokyo

S & M 3611

An Alternative Method for Upgrading the Conventional Decision Tree Algorithm

Suppakrit Kirdponpattara,[1] Veera Boonjing,[2] and Pitikhate Sooraksa[3*]

[1]School of Engineering, King Mongkut's Institute of Technology Ladkrabang, 1, Soi Chalongkrung 1, Ladkrabang, Bangkok 10520, Thailand
[2]Department of Computer Engineering, School of Engineering, King Mongkut's Institute of Technology Ladkrabang, 1, Soi Chalongkrung 1, Ladkrabang, Bangkok 10520, Thailand
[3]Department of Robotics and AI, School of Engineering, King Mongkut's Institute of Technology Ladkrabang, 1, Soi Chalongkrung 1, Ladkrabang, Bangkok 10520, Thailand

(Received February 21, 2024; accepted March 19, 2024)

Keywords: decision tree, defect prediction, feature selection, hard disk drive manufacturing

Decision tree algorithms are widely used for solving classification and regression problems. Their popularity can be attributed to their transparent nature, simplicity, easy interpretability, faster classification speed, and strong decision rules. However, decision tree induction algorithms face various inherent and external limitations, such as overfitting, high sensitivity to noise and outliers, and instability with minimal data variations. In this study, we introduce an innovative approach to enhance traditional decision tree algorithms [e.g., Iterative Dichotomiser 3 (ID3), C4.5, and Classification and Regression Trees (CART)] by incorporating feature selection techniques. The proposed approach aims to enhance the accuracy and efficiency of decision tree models. Experiments were conducted on a real-world dataset of a hard disk drive (HDD) manufacturing process using the proposed approach. In comparison with a baseline where all features were utilized, the study highlighted a significant improvement in accuracy, indicating that the approach holds immense potential for optimizing decision tree algorithms and improving the HDD manufacturing process.

1. Introduction

Decision tree algorithms have long been popular in the field of machine learning because they supply simple solutions for classification and regression tasks.(1–6) They are attractive because they imitate human decision-making processes, making them accessible to both experts and nonexperts. However, despite their popularity, traditional decision tree algorithms such as Iterative Dichotomiser 3 (ID3), C4.5, and Classification and Regression Trees (CART) have their limitations.

One of the main challenges faced by conventional decision tree algorithms is their tendency to overfit, especially when dealing with complex datasets.(7,8) They also struggle with noise and

*Corresponding author: e-mail: pitikhate.so@kmitl.ac.th
https://doi.org/10.18494/SAM5029

ISSN 0914-4935

outliers, which can affect the accuracy and reliability of the models.[9,10] Furthermore, even minor variations in the data can cause instability in model predictions.

To address these issues, alternative methods have been explored to improve traditional decision tree algorithms. One approach involves using feature selection techniques, which aim to identify the most relevant features for modeling while reducing the data's dimensionality. By focusing on informative features, these methods can enhance model performance and tackle problems like overfitting and noise.

For example, Liu and Zhou[11] proposed a feature selection approach based on decision tree rules for large-scale imbalanced data. In their approach, decision tree rules generated by the CART algorithm are used to identify relevant features on the basis of their frequency within the rules. To evaluate the effectiveness of their approach, they compared it with two other methods, chi-square and F-statistic, using a dataset from Santander Bank. The findings demonstrated that their approach yielded higher AUC, reduced computational time, and required fewer features compared with the filter-based methods.

As another example, An and Zhou[12] proposed an evaluation model for rural energy construction based on the ID3 decision tree algorithm and the Relief feature selection algorithm. The main goal of the model was to enhance the accuracy and efficiency of the decision tree algorithm. This was achieved by filtering out irrelevant attributes and assigning high importance to the significant attributes. Additionally, they conducted an analysis of the solar energy resources available in rural areas. They assessed the abundance, stability, utilization value, and suitability for grid-connected power generation of these resources. They concluded that solar energy holds immense potential for poverty alleviation and rural revitalization. Furthermore, they provided recommendations for improving the technology and application of solar energy utilization through demonstrations.

Estherly and Prasetiyo[13] applied the C4.5 algorithm with information gain for feature selection and the bagging ensemble method to overcome class imbalance in the Original Wisconsin Breast Cancer Dataset. By incorporating information gain as a feature selection technique, they were able to achieve a remarkable average accuracy of 96.49%. This surpassed the performance of the C4.5 algorithm without information gain or bagging ensemble.

In the field of the hard disk drive (HDD) manufacturing process, decision tree algorithms play a critical role. This complex task involves assembling components from various sources, making accurate predictive models essential. The integration of feature selection techniques into conventional decision tree algorithms is being pursued to improve model accuracy and efficiency.

In their recent study, Hirunyawanakul *et al.*[14] introduced an innovative method to enhance yield prediction in the manufacturing of HDDs. By employing machine learning and statistical techniques for feature selection, they proposed a fresh approach. They compared this approach with the conventional method that relies on human expert knowledge, using seven different feature selection methods: C5, CART, SVM, stepwise regression, genetic algorithm, chi-square, and information gain. Additionally, two learning algorithms, multiple linear regression (MLR) and artificial neural networks (ANN), were utilized for yield prediction modeling. The authors assessed the performance of their approach using root mean square error (RMSE) and mean absolute error (MAE), using a real-world dataset obtained from HDD manufacturing. The outcomes demonstrated that the combination of a genetic algorithm and MLR yielded the best results, significantly reducing errors compared with the traditional method. They concluded that their approach offers improved accuracy and efficiency in yield prediction of the HDD processes..

HDDs are an integral part of many electronic devices today. However, the manufacturing process of HDDs is complex and prone to defects, leading to potential data loss and increased costs. Therefore, predicting HDD failure accurately and efficiently is of paramount importance. In this paper, an approach is proposed that enhances traditional decision tree algorithms with feature selection techniques to predict HDD failure. This approach applies filter methods, which are fast, efficient, and generalizable techniques that can handle high-dimensional data and select the most informative features for modeling. Different criteria such as information gain (IG), gain ratio (GR), and symmetrical uncertainty (SU) are used to measure the correlation between features and classes. Three widely used decision tree algorithms, namely, ID3, C4.5, and CART, which are known for their simplicity and versatility, are also used. The approach is evaluated using a real-world dataset from an HDD manufacturer and compared with a baseline that uses all features. It is shown that this approach can achieve higher accuracy, lower computational time, and fewer features than the baseline. This case study illustrates the potential of this approach for optimizing decision tree algorithms and enhancing the HDD manufacturing process. It is believed that this approach can significantly contribute to the field by providing a more accurate, efficient, and interpretable method for predicting HDD failure. Previous studies have shown the effectiveness of feature selection techniques in enhancing decision tree algorithms, but they did not compare different feature selection techniques or decision tree algorithms systematically. Moreover, some of these studies used complex or computationally intensive methods that may not be suitable for large-scale or real-time scenarios. In this study, we address these gaps by introducing a novel method that integrates feature selection techniques with traditional decision tree algorithms.

In Sects. 2 and 3, a detailed overview of the dataset used in the study, the feature selection techniques employed, and the decision tree algorithms investigated are provided. In Sect. 4, we present the results and discussion of the study. Finally, Sect. 5 is a summary of the current work and also includes suggested directions for future research.

2. Data

The dataset used in this study is a real-world dataset obtained from a hard disk drive manufacturer, who provided us with the data of a single production timeframe. The dataset contains information about the components and machines that are used to assemble the hard disk drives, as well as the defect status of each hard disk drive after the assembly process. The defect status is labeled as either F (for defective) or P (for non-defective). The dataset has 53451 instances, each representing a hard disk drive, and 43 features, each representing a component or a machine. All the features and the class label are categorical, meaning that they have a finite number of possible values. For example, a feature related to a component vendor can have values such as V0, V1, V2, etc., indicating different vendors that supply the component. Similarly, a feature related to an assembly machine can have values such as M0, M1, M2, etc., indicating different machines that perform the assembly operation. The dataset has 26 features related to component vendors and 17 features related to assembly machines. The dataset also has a class imbalance, as there are more non-defective instances (46286) than defective instances (7165). One of the advantages of the dataset is that it is complete, meaning that there are no missing values in any of the features or the class label. This ensures the robustness and validity of the analysis, as there is no need to deal with the uncertainty or inconsistency caused by missing data.

3. Methodology

In this section, we explain the feature selection techniques and the decision tree algorithms used in the study, as well as the experimental procedure and the performance evaluation metrics. It also provides the formulas and definitions of the feature selection criteria, such as IG, GR, and SU, and the decision tree algorithms, such as ID3, C4.5, and CART. It also outlines the steps involved in the experimental procedure, such as applying feature selection techniques, ranking features, training and testing models, and comparing results.

3.1 Feature selection techniques

Feature selection techniques are used to reduce the dimensionality of the data and select the most relevant features for modeling. This can improve the accuracy, efficiency, and interpretability of the models, as well as avoid overfitting, noise, and redundancy. This study employs three feature selection techniques, namely, IG, GR, and SU, to rank the features in order of importance. These techniques are based on different criteria, such as entropy, split information, impurity, and correlation, to measure the degree of association between the features and the class.

Information Gain (IG): This technique measures the reduction in entropy or uncertainty of the class variable after splitting the data on the basis of a feature. Entropy is a measure of the randomness or disorder of a system. A high entropy means that the system is unpredictable, whereas a low entropy means that the system is predictable. The formula for IG is

$$IG(S,F) = H(S) - H(S \mid F), \tag{1}$$

where S is the data set, F is the feature, $H(S)$ is the entropy of the data set, and $H(S|F)$ is the conditional entropy of the dataset given the feature. The higher the IG, the more information the feature provides about the class.

Gain Ratio (GR): This technique is an extension of IG and considers the split information or intrinsic value of a feature. Split information is a measure of how evenly the data is distributed among the branches after splitting based on a feature. High-split information means that the data is evenly distributed, whereas low-split information means that the data is skewed. The formula for GR is

$$GR(S,F) = \frac{SI(S,F)}{IG(S,F)}, \tag{2}$$

where S is the dataset, F is the feature, $IG(S,F)$ is IG of the feature, and $SI(S,F)$ is the split information of the feature. The higher the GR, the more balanced the feature is.

Symmetrical Uncertainty (SU): This technique measures the correlation or mutual information between the feature and the class. Correlation is a measure of how closely the feature and the class are related. A high correlation means that the feature and the class are strongly associated, whereas a low correlation means that the feature and the class are weakly associated. The formula for SU is

$$SU(S,F) = 2 \times \frac{IG(S,F)}{IH(S) + H(F)}, \tag{3}$$

where S is the dataset, F is the feature, $IG(S,F)$ is IG of the feature, $H(S)$ is the entropy of the dataset, and $H(F)$ is the entropy of the feature. The higher the SU, the more correlated the feature and the class are.

3.2 Decision tree algorithms

Decision tree algorithms are widely employed in machine learning for their capacity to represent classification or regression rules in a treelike structure. Each node within the tree corresponds to a feature, whereas branches depict decision rules, and leaf nodes signify classes. These algorithms partition data recursively into smaller subsets on the basis of feature values until a predefined stopping criterion is met. Despite their ease of interpretation, mirroring human decision-making processes, decision tree algorithms may encounter challenges such as overfitting, susceptibility to noise and outliers, and sensitivity to minor data variations.

We utilize three prominent decision tree algorithms, namely, ID3, C4.5, and CART, to develop predictive models for classifying HDDs into passers or defectives, leveraging selected features.

Iterative Dichotomiser 3 (ID3) was proposed by Quinlan in 1986.(15) It employs IG as the feature selection criterion. The algorithm begins by considering the entire dataset as the root node. Subsequently, for each feature, it calculates IG concerning the class variable and selects the feature with the highest IG as the splitting feature. This iterative process continues until all instances within a branch belong to the same class or there are no more features to be split.

C4.5, an extension of ID3 proposed by Quinlan in 1993,(16) addresses several limitations of its predecessor. It utilizes GR instead of IG as the feature selection criterion and handles missing values, continuous features, and pruning. Similar to ID3, C4.5 begins with the entire dataset as the root node. For each feature, it calculates GR with respect to the class variable and selects the feature with the highest GR as the splitting feature. Additional steps are undertaken for handling continuous features and missing values, followed by pruning to enhance model accuracy.

CART, or classification and regression trees, proposed by Breiman *et al.*(17) in 1984, is a versatile algorithm capable of addressing both classification and regression tasks. It employs Gini impurity as the feature selection criterion. CART commences by considering the entire dataset as the root node. For each feature, it calculates the Gini impurity with respect to the class variable and selects the feature with the lowest Gini impurity as the splitting feature. Similar to previous algorithms, CART iteratively partitions the data until a stopping criterion is met, followed by pruning to refine model accuracy.

Despite their effectiveness, all three algorithms may be prone to overfitting and exhibit biases towards multilevel features. However, they remain widely utilized in machine learning owing to their adaptability and versatility in handling diverse datasets and conditions.

3.3 Experimental procedure

The experimental procedure aims to evaluate the performance of different decision tree algorithms and feature selection techniques in predicting HDD failure. It also compares the results with a baseline that uses all features. The procedure involves the following steps and is illustrated in Fig. 1. The program used to implement the procedure is R-3.6.3 software.

1. Start with the HDD data and prepare it for analysis.
2. Apply the three feature selection techniques (IG, GR, and SU) to compute the feature importance for all 43 features.
3. Rank the features by their importance for each feature selection technique.
4. Use the three decision tree algorithms (ID3, C4.5, and CART) to train and test models with all features as a baseline.
5. Train and test models using the ranking of each feature selection. Add one feature at a time from the top until the highest accuracy is achieved. Repeat this step for the top 20 features for each feature selection technique.
6. Use threefold cross-validation to evaluate the performances of the models in terms of accuracy, efficiency, and interpretability.
7. Compare the results of the models with the baseline using statistical tests and report the p-values and the 95% confidence intervals.
8. Conclude the procedure and summarize the findings.

We employ threefold cross-validation, where one fold is designated as the test data, while the remaining two serve as the training data. Performance evaluation of the models encompasses accuracy, efficiency, and interpretability. Accuracy is gauged using the confusion matrix, delineating correct and incorrect classifications for each class. Efficiency is quantified by the total time required to construct the models, encompassing feature selection, training, and testing durations. Interpretability is assessed through the top N features used in the generated models, indicative of their simplicity and clarity.

Additionally, we conduct statistical tests to juxtapose mean accuracy values and total time between each model and the baseline, with a predetermined significance level of 0.05. The results provide p-values and 95% confidence intervals that are used to elucidate the likelihood and range, respectively, of observing the results under the null hypothesis.

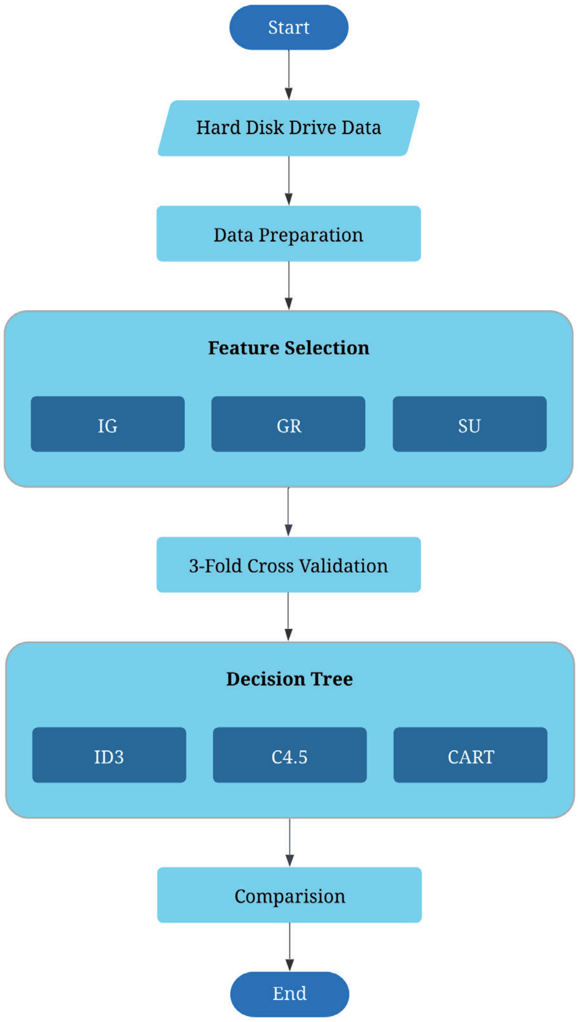

Fig. 1. (Color online) Methodology flow chart.

4. Results and Discussion

Table 1 presents the accuracy of each decision tree algorithm when using all features and each feature selection technique. The table also includes the minimum and maximum values of the 95% confidence intervals for each accuracy value, which represent the range of potential accuracy values if the experiment were to be conducted with different data samples.

From the data in Table 1, it is evident that the C4.5 algorithm using the IG feature selection technique achieved the highest accuracy, with a mean value of 0.88172. This is closely followed by the ID3 algorithm also using the IG technique, with a mean value of 0.88083. The CART algorithm using all features yielded the lowest accuracy, with a mean value of 0.87138. The non-overlapping confidence intervals for most cases indicate that these differences in accuracy are statistically significant.

Table 1
Accuracy of each decision tree algorithm at 95% confidence interval between all features, IG, GR, and SU.

Decision Tree Algorithm	Feature Selection Technique	Accuracy	Minimum	Maximum
ID3	All Features	0.87476	0.87307	0.87646
	IG	0.88083	0.87984	0.88181
	GR	0.88058	0.88010	0.88106
	SU	0.88056	0.87999	0.88114
C4.5	All Features	0.87759	0.87654	0.87864
	IG	0.88172	0.88144	0.88201
	GR	0.88073	0.88034	0.88113
	SU	0.88170	0.88142	0.88199
CART	All Features	0.87138	0.86952	0.87324
	IG	0.88084	0.87952	0.88217
	GR	0.88053	0.87997	0.88108
	SU	0.88056	0.87999	0.88114

These results suggest that feature selection techniques can enhance the accuracy of decision tree algorithms, since models using these techniques outperformed those using all features. Among the feature selection techniques, IG was the most effective, followed by SU and GR. This indicates that these techniques can efficiently identify the most relevant and informative features for predicting HDD failure. In terms of decision tree algorithms, C4.5 and ID3 outperformed CART, suggesting that these algorithms can generate more accurate and robust models for the given dataset.

Table 2 shows the top N features for building models using different decision tree algorithms and feature selection techniques. The table indicates that feature selection techniques can significantly reduce the number of features required for modeling compared with using all features. This can improve the efficiency and interpretability of the models, as well as avoid overfitting, noise, and redundancy. Among the feature selection techniques, IG was the most effective, followed by GR and SU. This suggests that these techniques can efficiently identify the most relevant and informative features for predicting hard disk drive failure. In terms of decision tree algorithms, C4.5 and ID3 outperformed CART, as they required fewer features to achieve comparable or higher accuracy. This implies that these algorithms can generate more accurate and robust models for the given dataset.

Table 3 presents the total time required for building models using different decision tree algorithms and feature selection techniques. It is observed that the ID3 algorithm with the IG feature selection technique had the shortest total time of 9.4 s, indicating that this combination is the most efficient. On the other hand, the CART algorithm with the SU feature selection technique had the longest total time of 37.2 s, suggesting that this combination is the least efficient.

In summary, the results in Tables 1–3 suggest that feature selection techniques, particularly Information Gain (IG), can significantly improve the accuracy and efficiency of decision tree algorithms. IG is an efficient and effective feature selection technique when integrated with decision tree algorithms because it can identify the most relevant and informative features for predicting the class variable. By using IG, decision tree algorithms can reduce data

Table 2
Top N features for building models.

Decision Tree Algorithm	Feature Selection Technique	Top N Features
ID3	All Features	43
	IG	7
	GR	9
	SU	9
C4.5	All Features	43
	IG	6
	GR	10
	SU	10
CART	All Features	43
	IG	6
	GR	9
	SU	9

Table 3
Total time required for building models (second).

Decision Tree Algorithm	Feature Selection Technique	Total time
ID3	All Features	48.0
	IG	9.4
	GR	9.8
	SU	14.4
C4.5	All Features	42.5
	IG	13.7
	GR	17.8
	SU	20.6
CART	All Features	50.2
	IG	25.3
	GR	30.2
	SU	37.2

dimensionality, avoid overfitting, noise, and redundancy, and improve the accuracy, efficiency, and interpretability of the models.

The integration of IG with three decision tree algorithms (ID3, C4.5, and CART) varies in effectiveness depending on the characteristics of the dataset and the application domain. For instance, ID3 and C4.5 are more suitable for categorical features, whereas CART can handle both categorical and continuous features. ID3 and C4.5 use entropy as the measure of impurity, whereas CART can use either entropy or Gini impurity. C4.5 uses normalized IG or gain ratio as the splitting criterion, whereas ID3 uses the raw IG value. ID3 and C4.5 are more prone to overfitting, whereas CART uses pruning to avoid overfitting. ID3 and C4.5 are more popular in machine learning and natural language processing, whereas CART is more versatile and can be used for both classification and regression tasks.

However, the specific choices of the decision tree algorithm and feature selection technique may depend on the acceptable trade-off between accuracy and computational efficiency for a given application. The IG feature selection technique was found to be the most effective and efficient across all decision tree algorithms, affecting the total time required for building models.

5. Conclusions

We presented a novel approach to enhance traditional decision tree algorithms by integrating feature selection techniques. The objective of this method is to enhance the accuracy and efficiency of decision tree models in predicting HDD failure. To assess the effectiveness of this approach, experiments were conducted on a real-world dataset from an HDD manufacturing process. Three feature selection techniques (IG, GR, and SU) and three decision tree algorithms (ID3, C4.5, and CART) were utilized. The outcomes revealed that feature selection techniques have a significant impact on improving the accuracy of decision tree algorithms, but they also influence the time required to build models. Among the feature selection techniques, IG was found to be the most effective and efficient across all decision tree algorithms. Moreover, IG can reduce the number of features needed for modeling, which can improve the interpretability and robustness of the models. This proposed method has enormous potential for optimizing decision tree algorithms and enhancing the HDD manufacturing process. Future research should explore other feature selection techniques, decision tree algorithms, and datasets to validate and generalize the findings of this study.

References

1 Z. He, Z. Wu, G. Xu, Y. Liu, and Q. Zou: Trans. Knowl. Data Eng. **35** (IEEE, 2023) 1. https://doi.org/10.1109/TKDE.2021.3075023
2 A. Law and A. Ghosh: IEEE Trans. Emerging Top. Comput. Intell. **6** (IEEE, 2022) 3. https://doi.org/10.1109/TETCI.2021.3075717
3 P. Zang: Proc. Integrated Intelligence and Communication (IEEE, Kalaburagi, 2023) 1–6. https://doi.org/10.1109/ICIICS59993.2023.10421137
4 M. Abdurohman, A. G. Putrada, and M. M. Deris: IEEE Access **10** (2022) 56937. https://doi.org/10.1109/ACCESS.2022.3177225
5 S. Dong and D. Fei: Proc. Computer Engineering and Artificial Intelligence (IEEE, Shanghai, 2021) 325–330. https://doi.org/10.1109/ICCEAI52939.2021.00065
6 Z. Zhang: Proc. Machine Learning, Big Data and Business Intelligence (IEEE, Taiyuan, 2021) 280–283. https://doi.org/10.1109/MLBDBI54094.2021.00059
7 A. Amro, M. Al-Akhras, K. E. Hindi, M. Habib, and B. A. Shawar: J. Intell. Syst. **30** (2021) 438. https://doi.org/10.1515/jisys-2020-0061
8 R. García Leiva, A. Fernández Anta, V. Mancuso, and P. Casari: IEEE Access **7** (2019) 99978. https://doi.org/10.1109/ACCESS.2019.2930235
9 Priyanka and D. Kumar: Inf. Decis. Sci. **12** (2020) 246. https://doi.org/10.1504/IJIDS.2020.108141
10 I. D. Mienye, Y. Sun, and Z. Wang: Procedia Manuf. **35** (2019) 698. https://doi.org/10.1016/j.promfg.2019.06.011
11 H. Liu and M. Zhou: Proc. Wireless and Optical Communication Conf. (IEEE, Newark, 2017) 1–6. https://doi.org/10.1109/WOCC.2017.7928973
12 Y. An and H. Zhou: Energy Rep. **8** (2022) 1004. https://doi.org/10.1016/j.egyr.2022.01.239
13 B. Estherly and B. Prasetiyo: Proc. AIP Conf. 2614 (Bali, 2023) 040012-1-040012-8. https://doi.org/10.1063/5.012594
14 A. Hirunyawanakul, N. Kaoungku, N. Kerdprasop, and K. Kerdprasop: Electr. Electron. Eng. Telecommun. **9** (2020) 420. https://api.semanticscholar.org/CorpusID:231668647
15 J. R. Quinlan: Mach. Learn. **1** (1986) 81. https://doi.org/10.1023/A:1022643204877
16 J. R. Quinlan: C4.5: Programs for Machine Learning. (Morgan Kaufmann Publishers Inc., San Francisco, 1993).
17 L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone: Classification and Regression Trees 1 (Routledge, 1984). https://doi.org/10.1201/9781315139470

About the Authors

Suppakrit Kirdponpattara is currently working towards a Ph.D. degree in robotics and artificial intelligence (RAI) at King Mongkut's Institute of Technology Ladkrabang. Prior to this, he completed his B.Eng. degree in computer engineering at Prince of Songkla University, Thailand, and followed it with an M.S. degree in computer science from King Mongkut's Institute of Technology Ladkrabang, Thailand. (63601230@kmitl.ac.th)

Veera Boonjing is an associate professor of information technology at the Department of Computer Engineering, School of Engineering at King Mongkut's Institute of Technology Ladkrabang, Thailand. He received a B.Sc. degree in mathematics from Ramkhamhaeng University, Thailand, M.S. degree in computer science from Chulalongkorn University, Thailand, and Ph.D. degree in decision sciences and engineering systems from Rensselaer Polytechnic Institute, USA. (veera.bo@kmitl.ac.th)

Pitikhate Sooraksa is a professor at the School of Engineering, King Mongkut's Institute of Technology Ladkrabang, Thailand. He received his B.Ed. (Hons) and M.Sc. degrees in physics from Srinakharinwirot University, Thailand, his M.S. degree from George Washington University, USA, in 1992, and his Ph.D. degree from the University of Houston, USA, in 1996, both in electrical engineering. His research interests include cyber-physical applications and rapid prototypes in robotics and AI.
(pitikhate.so@kmitl.ac.th)