Classification of Dry Docking Duration Based on Ant-Miner Algorithm

Isti Surjandari, Amar Rachman, Riara Novita, and Muthia S. Naffisah

Abstract— Classification is one of the most studied data mining tasks. Many literatures stated that predictive accuracy resulted by classification in data mining tends to be lower compared with the classification combined with optimization. The objective of this study is to estimate Dry Docking duration based on classification by using two methods; CART and Ant-Miner. It is shown that CART gives more accurate classification model than Ant-Miner with static discretization.

Keywords— Ant-Miner, CART, Data Mining, Dry Docking, Maintenance Duration

I. INTRODUCTION

A LONG with the development of technology, some techniques of data collection and storage is also rising. The use of computerized system is an example of technological developments that led to the growth of very large data. However, the amount of knowledge gained is sometimes disproportionate to its growth. This kind of situation has led to the emergence of data mining. According to Larose, there are six data mining tasks: description, estimation, clustering, association, classification, and prediction [1].

Classification is one of the most studied data mining tasks [2, 3]. It can be seen from the number of researches which developed classification model in order to improve its predictive accuracy, such as the research conducted by Izrailev et. al. (2001) which concluded that regression tree in data mining combined with Ant Colony gave a better quality tree models than the usual regression tree [4]. Boryczka et. al. (2011) found that Ant Colony Decision Trees (ACDT) is very competitive with the well-known CART algorithm [5].

The objective of this study is to estimate Dry Docking duration using CART and Ant-Miner with static discretization. Weka was used for classification and GUI Ant-Miner was used for classification induced by ant colony algorithm. Data mining is an analysis method of large amounts of data to find

Isti Surjandari is with Department of Industrial Engineering, Faculty of Engineering, University of Indonesia, Kampus UI, Depok 16424, Indonesia (corresponding author's name: +622178888805, <u>isti@ie.ui.ac.id</u>).

Amar Rachman is with Department of Industrial Engineering, Faculty of Engineering, University of Indonesia, Kampus UI, Depok 16424, Indonesia (e-mail: <u>amar@ie.ui.ac.id</u>).

Riara Novita is with Department of Industrial Engineering, Faculty of Engineering, University of Indonesia, Kampus UI, Depok 16424, Indonesia (e-mail: <u>riara.novita_ti08@yahoo.com</u>).

Muthia S. Naffisah is with Department of Industrial Engineering, Faculty of Engineering, University of Indonesia, Kampus UI, Depok 16424, Indonesia (e-mail: <u>muthiaszami@yahoo.com</u>).

a hidden pattern, in this case, the relationship between maintenance data and its duration. So that, by knowing the volume of each type of dry docking work to be done, the operator can directly estimate the dry docking duration.

II. METHODS

Classification and Regression Tree (CART) is one of the decision tree algorithms for classification by constructing a flowchart-like structure [4]. The characteristic of CART is to use a set of "if-then" conditions to perform predictions or classification of cases [4,1]. CART analysis consists of four basic steps. The first step is the tree building, in which a tree is built using recursive splitting of nodes. Parent node is an attribute with the purest node among the others. Mostly, CART uses Gini index to get node purity of each attribute, which is defined by (1).

$$\Delta i(\tau) = i(\tau) - \{p_L i(\tau_L) + p_R i(\tau_R)\}$$
(1)

Where,

$$i(\tau) = 2\left(\frac{n_{+1}}{n_{++}}\right) \left(1 - \frac{n_{+1}}{n_{++}}\right) = 2\left(\frac{n_{+1}}{n_{++}}\right) \left(\frac{n_{+2}}{n_{++}}\right)$$
(2)

$$i(\tau_{L}) = 2\left(\frac{n_{11}}{n_{1+}}\right) \left(1 - \frac{n_{11}}{n_{1+}}\right) = 2\left(\frac{n_{11}}{n_{1+}}\right) \left(\frac{n_{12}}{n_{1+}}\right)$$
(3)

$$i(\tau_R) = 2\left(\frac{n_{21}}{n_{2+}}\right) \left(1 - \frac{n_{21}}{n_{2+}}\right) = 2\left(\frac{n_{21}}{n_{2+}}\right) \left(\frac{n_{22}}{n_{2+}}\right)$$
(4)

TABLEI

CART CLASSIFICATION				
Split	Class of Y		Row Total	
	1	0	Row Total	
$x_i \leq c$	n ₁₁	<i>n</i> ₁₂	n ₁₊	
$x_i > c$	n ₂₁	n ₂₂	<i>n</i> ₂₊	
Column Total	<i>n</i> ₊₁	<i>n</i> ₊₂	<i>n</i> ₊₊	

The second step consists of stopping the tree building process. At this point, a "maximal" tree has been produced or the number of cases for each terminal node is less than minimum required, which is 5 cases or 10% of all cases in training set. The third step includes the tree "pruning", results in the creation of a sequence of simpler and simpler trees based on re-substitution error rate in every terminal node, which can be formulated as follow:

$$R(\tau_l) = r(\tau_l)p(\tau_l)$$
(5)

Where,

$$r(\tau_l) = 1 - \max(p, 1 - p) = \min(p, 1 - p)$$
(6)

$p(\tau)$ is the probability that an observation falls into node τ .

The fourth step provides the optimal tree selection, where the tree that fits the information in the learning dataset, but does not overfit the information, is selected from among the sequence of pruned trees by using testing data, as in (7) and cross validation, as in (8).

$$R^{ts}(T_*) = \min_k R^{ts}(T_k)$$
(7)

$$R^{\frac{CV}{V}}(T_*) = \min_k R^{\frac{CV}{V}}(T_k)$$
(8)

Where,

$$R^{\frac{CV}{V}}(T_k) = R^{\frac{CV}{V}}(T_{(\alpha_k)})$$
(9)

$$R^{\frac{CV}{V}}(T_{(\alpha)}) = N^{-1} \sum_{i=1}^{J} \sum_{j=1}^{J} n_{ij}(\alpha)$$
(10)

$$n_{ij}(\alpha) = \sum_{v=1}^{v} n_{ij}^{v}(\alpha), i, i = 1, 2, ..., J$$
(11)

Steps in building the classification model using Ant-Miner algorithm are as follows:

- 1) Determine the parameters, such as number of iterations and number of ants.
- 2) Tree building. Like CART, Ant-Miner also select the attributes used in classification model by using the following equations:

$$p_{i} = \frac{\tau_{E,L,x_{i}} \cdot \eta_{i}}{\sum_{i \in F} \tau_{E,L,x_{i}} \cdot \eta_{i}}$$
(12)

Where,

$$\eta_{x_i} = \text{GainRatio}(S, x_i) \tag{13}$$

Gain Ratio (S, A) = $\frac{\text{Gain}(S,A)}{\text{SlitInformation}(S,A)}$ (14)

$$Gain(S, A) = Entropy(S) - \sum_{v=1}^{T} \frac{|S_v|}{|S|} \cdot Entropy(S_v)$$
(15)

SplitInformation (S, A) =
$$\sum_{v=1}^{T} - \frac{|s_v|}{|s|} \cdot \log_2 \frac{|s_v|}{|s|}$$
 (16)

Entropy
$$(S) = \sum_{c=1}^{m} -p_c \cdot \log_2 p_c$$
 (17)

- 3) Stopping tree building. The tree-building process will stop if the number of cases in each node is smaller than the minimum number required or there is no more attributes that can be used as a node.
- Pruning. Using error-based pruning method, terminal nodes will be cut off from the tree if their errors tend to be larger than their parent node's error.
- 5) Updating pheromone on each path that has been passed by the best ant candidate. This process itself can either be pheromone addition or subtraction.

$$\tau_{(E,L,x_{i})} = \begin{cases} \rho \cdot \tau_{E,L,x_{i}} & \text{if } (E,L,x_{i} \notin \text{tree}_{ib} \\ \rho \cdot \tau_{E,L,x_{i}} & \text{if } (E,L,x_{i} \in \text{tree}_{ib} \end{cases}$$
(18)

Where,

$$Q = \frac{N - Error}{N}$$
(19)

- 6) Repeating the 2nd to the 5th step until the number of classification rule is larger or equal to the number of ants or the current ant have built the same classification model.
- 7) After generating the number of classification model as well as determining the number of ants and iterations, the next step is selecting the best classification model from the set of classification models that have been generated before.

The attributes and its specification for each of the dry docking works are summarized in Table 2. Attributes reduction is done based on interview with the relevant operator who is an expert in ship maintenance. The attributes identified as irrelevant are those whose influence on the overall work duration in dock is assessed as not significant.

Redundant attributes are those which contain information that is already contained in some other attribute(s). On the other hand, attributes whose values are the same for all or almost instances in sample, are considered as useless because it was obvious that such attributes do not have influence or have very little influence on the learning model [6]. There are 28 significant attributes out of 35 attributes.

TABLE II DRY DOCKING ATTRIBUTES SPECIFICATION				
Attributes	Data Specification			
ukuran, tank, bak	Volume (m ³)			
grt, pelat	Mass (ton)			
scraping, sandblasting, cuci, cat	Wide (m ²)			
zinc, sealprop, ringprop, packingprop, chrome, ringkem, packingkem, sealkem, chest, valve, scrupper, manhole, plug	Amount			
ut, las	Amount (point)			
grease	Mass (kg)			
propeller, porosprop, kemudi, poroskem	A = no maintenance, B = recondition, C = balancing, D = change, E = recondition and balancing, F = change and balancing, G = change and recondition, H = change, recondition, and balancing			
bantalanprop, bantalankem	A = no maintenance, B = recondition, C = change, D = balancing			
shaftseal, jangkar	Options (YES, NO)			
panjanglas	Length (m)			

m = meter and kg = kilogram.

The classification tree for the dry docking duration can be seen in Figure 1 where dry docking duration is divided into 4 classes [7]. Then, to get the same behavior from both CART and GUI Ant-Miner, we discretized dry docking duration into 4 bins by using Weka Discretization.



Fig. 1 Classification Tree of Dry Docking Duration

III. RESULT AND ANALYSIS

In this study, we compare the predictive accuracy resulted by Ant-Miner with static discretization and CART. From Table 3, it can be seen that the predictive accuracy of CART is similar to Ant-Miner with static discretization, and even is likely to be higher than Ant-Miner's. Based on calculations performed by using software, CART has higher accuracy than static discretization Ant-Miner by 0.2667%. While based on manual calculations, CART predictive accuracy is 3.03% higher than Ant-Miner with static discretization. It can be concluded that the predictive accuracy resulted by CART is higher than Ant-Miner with static discretization.

TABLE II			
PREDICTIVE ACCURACY			

Methods	Accuracy (Software) [%]	Accuracy (Manual) [%]
CART	65.4867%	63.64%
Ant-miner	65.22%	60.61%

IV. CONCLUSIONS

Maintenance is one of the most important activities in the shipyard industry. This study uses CART and Ant-Miner with static discretization to estimate the duration of dry docking. The result of this study shows that CART's predictive accuracy tends to be higher than the accuracy of Ant-Miner with static discretization. This indicates that Ant-Miner is not always better than CART.

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to University of Indonesia for financing this research, and also to Dr. Srdoc for her help in analyzing the attributes for dry docking works.

REFERENCES

- D.T. Larose, Discovering Knowledge in Data: An Introduction to Data Mining. New Jersey: John Wiley & Sons, Inc, 2005.
- [2] M. Dorigo, G.D. Caro and L.M. Gambardella, Ant Algorithms for Discrete Optimization, edited by M.A. Bedau, Artificial Life, MIT Press, 1999, no..3, volume 5
- [3] F.E.B. Otero, A.A. Freitas and C.G. Johnson, *Inducing Decision Trees with An Ant Colony Optimization Algorithm*, Applied Soft Computing Issues, 2012, vol. 12, p. 3615-3626 http://dx.doi.org/10.1016/j.asoc.2012.05.028
- [4] D.J. Hand, H. Mannila and P. Smyth, "Principles of Data Mining" Cambridge MA: MIT Press, 2001.
- [5] U. Boryczka, J. Kozak, "New Insights of Cooperation among Ants in Ant Colony Decision Trees" Proceeding of Third Worlds Congress on Nature and Biologically Inspired Computing, Spain, October 19-21 2011.
- [6] S. Izrailev and D. Agrafiotis, A Novel Method for Building Regression Tree Models for QSAR based on Artificial Ant Colony System, J. Chem. inf. Comput. Sci., 2001, vol. 41 pp. 176-180 http://dx.doi.org/10.1021/ci000336s
- [7] I. Surjandari and R. Novita, "Estimation of Dry Docking Duration Using Data Mining", Proceeding of International Conference on Operations Research and Statistics, Zurich, Switzerland, July 30-31 2013.

Isti Surjandari is Professor and Head of Statistics and Quality Engineering Laboratory in the Department of Industrial Engineering, Faculty of Engineering, University of Indonesia. She holds a bachelor degree in Industrial Engineering from University of Indonesia and a Ph.D. degree from the Ohio State University. Her areas of interest are industrial management, quality engineering and applied statistics. She has a vast experience in industrial and manufacturing systems and has published papers in national and international journals. She is a senior member of American Society for Quality (ASQ), and Indonesian Association of Industrial Engineering and Management (ISTMI). She is also Editorial Board for International Journal of Technology.

Amar Rachman is a senior lecturer in the Department of Industrial Engineering, Faculty of Engineering, University of Indonesia. He has obtained his bachelor degree in Mechanical Engineering from University of Indonesia and Master of Engineering in Industrial Management from KU Leuven Belgium.

Riara Novita is a research assistant in the Department of Industrial Engineering, Faculty of Engineering, University of Indonesia. She holds a bachelor and master degrees in Industrial Engineering from University of Indonesia.

Muthia S. Naffisah is now pursuing her bachelor degree in Industrial Engineering, Faculty of Engineering, University of Indonesia. She has become a research assistant in Statistics and Quality Engineering Laboratory University of Indonesia since 2012.