Bayes Risk Post-Pruning in Decision Tree to Overcome Overfitting Problem on Customer Churn Classification

Devina Christianti[1], Sarini Abdullah[2] , Siti Nurrohmah[3]
{devina.christianti@sci.ui.ac.id[1], sarini@sci.ui.ac.id[2], snurrohmah@sci.ui.ac.id[3]}

Department of Mathematics, Universitas Indonesia, Depok, Indonesia[1,2,3]

Abstract. Classification is the process of assigning a set of data into an existing class. Decision tree is claimed to be faster and produces better accuracy compared to another classifier. However, it has some drawbacks in which the classifier is susceptible to overfitting. This problem can be avoided by post-pruning that trimming the small influence subtree in conducting the classification to improve model performance in predicting data. This paper proposes a Post-Pruning method by applying Bayes Risk, in which the risk estimation of each parent node compared with its leaf. This method is applied to two datasets of customer churn classification from the Kaggle site and IBM Datasets with three different sizes for training dataset (60%, 70%, and 80%). For the result, Bayes Risk Post-Pruning can improve decision tree performance and the larger the size of the training dataset was associated with higher accuracy, precision, and recall of the model.

Keywords: Binary Class, C4.5 Algorithm, Entropy, Gain Ratio, Zero-One-Loss.

1 Introduction

Classification is a supervised data mining operation that uses a dataset consisting of the number of vector attributes through its value and attribute classes to build the model. The model will be useful in predicting classes from data based on the values of its attributes. In building a classification model, the dataset is generally split into two parts, named training dataset to build the model and testing dataset to measure the model performance, therefore there are two kinds of error, called training error and testing error.

A good model should have a low training error and testing error to make the model fits the training dataset and accurately classifies the testing dataset. If the model has high training error and testing error, it is called underfitting which occurs because the training algorithm is still learning the data. Training error can be reduced by increasing data complexity, but it can increasing testing error which causes it cannot predict new data properly. This condition is called overfitting which can occur due to noise or the lack of a representative sample in the training dataset [2].

Decision tree is a method to classify data into an existing class. Compared with another classifier, decision tree works faster with better accuracy [1]. Decision tree consists of a node contains a collection of datasets and edges which represents the value of attributes splitting node and connect each node [2].

Decision tree is prone to overfitting [3], as the algorithm can produce many branches and leaves that make subtree contains noise or outlier. Moreover, the tree gets more complex and

ICSA 2019, August 02-03, Bogor, Indonesia
Copyright © 2020 EAI
DOI 10.4108/eai.2-8-2019.2290487

difficult to understand [9]. Overfitting can be avoided by adjusting the size of the tree or trimming the subtree which gives a small influence in the model, called pruning. There are two types of pruning methods named pre-pruning and post-pruning. Post-pruning tends to give better results than pre-pruning since pre-pruning can produce a premature tree. [6]

In this paper, the condition of overfitting that occurred on decision treemodel before post-pruning will be shown in two datasets of customer churn with one proportion of the training dataset (60%). Therefore Bayes Risk will be applied for post-pruning to overcome overfitting problem on decision tree in both datasets with three proportion of training dataset (60%, 70%, and 80%).Bayes risk method applies Bayes Theorem by considering the loss in making a decision.

2 Methods

This paper proposes an algorithm that adopts decision tree as a classifier and uses Bayes Risk as a post-pruning method. The principle of the method is shown in Figure 1.

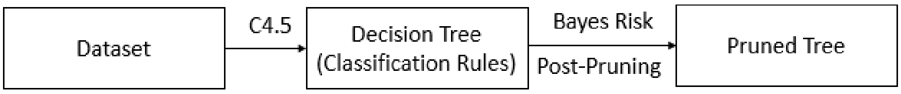

Fig. 1. Decision tree post-pruning algorithm flow based on Bayes Risk.

2.1 Decision tree

Decision tree begins from the root node which contains the entire dataset and then split by an attribute that forming nodes and edges which connecting the nodes. This process continues recursively with the rules of splitting (goodness of split criterion) until the stopping criteria are met. A node that is not splitting anymore is called a leaf node and labeled with the majority class. The algorithm to build the decision tree is C4.5 which used gain ratio as splitting criteria. It incorporates entropy, which measures the homogeneity of data on a node.

Suppose there are variable predictors called attributes, denoted by $\boldsymbol{x}$ and target variable that contains classes, denoted by C_i , $i = 1,2,..m$ where m is the number of classes. $p(C_i)$ is a proportion of observations that belong to class C_i. Entropy $\left(E(D)\right)$is given by equation (1).

$$E(D) = -\sum_{i=1}^{m} p(C_i).\log_2 p(C_i) \qquad (1)$$

Suppose attribute x has w different value and for each value produces subset dataset, called D_v , where$v = 1,2,..w$. $|D|$ is the number of instances in dataset D, and $|D_v|$is the number of instances in $D_v$which is the dataset partitioned by the attribute that have w different value. Information gain, which measures the homogeneity difference of data before and after a node is split by attribute x, is given by equation (2).

$$Gain(x) = E(D) - \sum_{v \in x} \frac{|D_v|}{|D|} E(D_v) \qquad (2)$$

The higher information gain value, the more homogeneous partitions generated by split node by attribute x. Information gain normalized by split info to avoid bias, called Gain Ratio, see equation (3).

$$GainRatio\ (x) = \frac{Gain\ (x)}{SplitInfo\ \ (x)} \qquad (3)$$

In building a decision tree, an attribute that has the highest gain ratio value is used to split the node.

2.2Bayes risk post-pruning

Decision rule that minimizes expectations of loss or risk by applying Bayes, called Bayes Risk, is very important because in its application costs are corresponding to misclassification error [8]. The risk corresponding to the cost of classifying the data with attributes$\boldsymbol{x}$ into class C_i , where the correct class is $C_j, j = 1,2,..m$. Conditional risk is defined when classifying $\boldsymbol{x}$into the class C_i as in equation (4).

$$l^i(\boldsymbol{x}) = \sum_{j=1}^{2} \lambda_{j,i} \Pr(C_j|\boldsymbol{x}) \qquad (4)$$

Where $\lambda_{j,i}$ =cost of classifying the data into class C_i, where the true class is C_J and $\Pr(C_j|\boldsymbol{x})$ is calculated using Bayes' Theoremgiven in equation (5).

$$\Pr(C_j|\boldsymbol{x}) = \frac{\Pr(\boldsymbol{x} \cap C_j)}{\Pr(\boldsymbol{x})} = \frac{\Pr(\boldsymbol{x}|C_j).\Pr(C_j)}{\Pr(\boldsymbol{x})} = \frac{\Pr(\boldsymbol{x}|C_j).\Pr(C_j)}{\sum_{l=1}^{m} \Pr(\boldsymbol{x}|C_l).\Pr(C_l)} \qquad (5)$$

One special case of risk matrix is zero-one-loss which has the same cost when misclassifying (classifying a subject with the i class as aj class or vice versa) as in equation (6).

$$\lambda_{j,i} = \begin{cases} 1, & ifi \neq j \\ 0, & ifi = j \end{cases} \qquad (6)$$

Bayes Risk method is applied for a post-pruning decision tree referred to Ahmed et al. (2018). A post-pruning algorithm is run from the bottom (leaf node) - up (root node) by evaluating risk each subtree based on Bayes Risk. Based on zero-one loss, risk associated with each parent node $t$isshown inequation (7).

$$R_t^i(\boldsymbol{x}) = \sum_{j=1, j \neq i}^{2} \lambda_{j,i}\ Pr\ (C_j|\boldsymbol{x}) \qquad (7)$$

Where:$R_t^i(\boldsymbol{x})$ = risk associated with node t when classifying subject with attribute $\boldsymbol{x}$into class $C_i$and $Pr(C_j|\boldsymbol{x})$= probability of a subject with attribute $\boldsymbol{x}$ predicted in class C_j .

The risk associated with the leaf node of its parent node t is shown inequation (8).

$$R_l = \sum_{l=1}^{tl} R_l^i(\boldsymbol{x}) \qquad (8)$$

Where $R_l^i(\boldsymbol{x}) =$ risk associated with leaf node l when classifying subject with attribute $\boldsymbol{x}$ into class C_i and tl = total leaf nodes in the subtree.

After the decision tree is built, an algorithm of Bayes Risk Post-Pruning evaluates each subtree by calculating the risk rates of the parent node and its leaf node. If the risk associated with parent node is smaller than the total risk associated with its leaf node, $(R_p < R_l)$,the parent node will be altered to a leaf node,if not, the parent node is retained [1].

3 Model performance evaluation

Performance measurement is important to show how well the model classifies data. In general, the confusion matrix table as in Table 1 is usedwhen the true values are known to evaluate themodel performanceon a set of test data [4]. Model performance is measured by calculating the value of accuracy, recall, and precision.

Table 1. Confusion matrix.

		Predicted	
		Negative	**Positive**
Actual	**Positive**	False Negative (FN)	True Positive (TP)
	Negative	True Negative (TN)	False Positive (FP)

Where

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \; ; \; Precision = \frac{TP}{TP+FP} \; ; \; Recall = \frac{TP}{TP+FN}$$

Accuracy is obtained by calculated the true predicted observation divided by total observations. Precision also known as positive predicted values is obtained by calculated thetrue predicted positive observations divided by total predicted positive observations. Recall also called sensitivity is the fraction oftrue predicted positive observations of the totalactual positive observation.

4 Data analysis

4.1 Customer churn classification

Customer churn is a condition when a customer stops using products/services from the company for a certain time. Companies usually use predictive analysis applications for modeling customer churn to find out customers who have the potential to unsubscribe from the company's products and know the causal factors that must be avoided in the future. One of the company's strategies to retract customers who have the potential to unsubscribe is to provide attractive promos for these customers. There are two datasets about customer churn in the telephone company with the target variable or class for both datasets are 0 = No-Churn and 1 = Churn.

In the first dataset (Dataset I) from Kaggle site (Telecom Customer Churn), there are 3333 customers with 16 attributes (x) that describe behavior of customer such as account length, voicemail message, day mins, eve mins, night mins, intl mins, customer service calls, int Plans, day calls, day charges, eve calls, eve charges, night calls, night charges, intl calls, and intl charge.

In the second dataset (Dataset II) from IBM Dataset (Telco Customer Churn), there are 7043 customers with 19 attributes (x) that describe behavior of customer such as gender, senior citizen, partner, dependents, tenure, phone service, multiple lines, internet service, online security, online backup, device protection, tech support, streaming TV, streaming movies, contract, paperless billing, payment method, monthly charges and total charges.

4.2Pre-processing data

Pre-processing data aims to check whether the data contains rows or columns that are duplicated, empty (null), check whether variable manipulation is needed, and do the treatment on variables with continuous/categorical data types before fitting the data. In both datasets of customer churn, there are no missing data and duplicate rows or columns. Variables with categorical data types were changed to numeric so that they could be operated in a matrix.

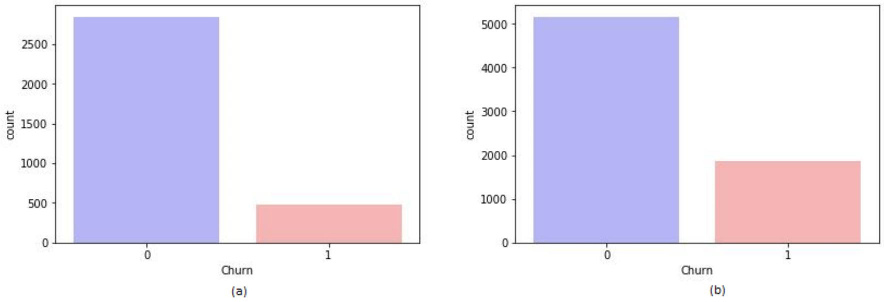

Fig. 2. Proportion of classes 0= No-Churn Class and 1= Churn Class in (a) first dataset, (b) second dataset.

Based on Figure 2, it shows that both datasets are between-class imbalance because the number of customers belongs to churn and no-churn classes are far different. The problem with the imbalance class is a classification error because the minority class cannot be classified properly compared to the majority class [7]. Therefore, a random oversampling method in the training dataset was needed by adding observations to the minority class randomly so that the proportion of minority classes (churn class) increased until the proportion of minority and majority class was balanced.

4.3 Experiment

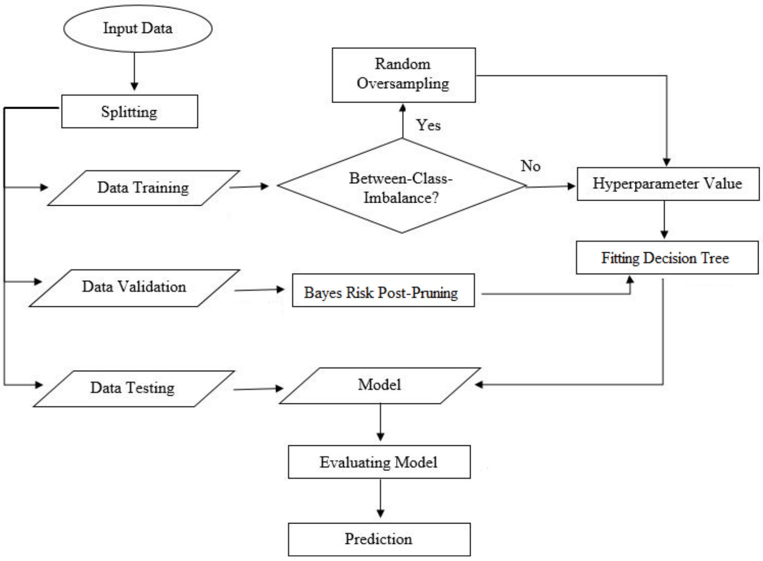

Fig. 3. The experiment process of modeling customer churn using decision tree with Bayes Risk Post-Pruning method.

Each dataset was divided into three parts, namely training dataset for building a decision tree, validation dataset for learning Bayes Risk Post-Pruning method, and testing dataset for evaluating the model as in Figure 3. There were three types of composition for training, validation, and testing datasets which were 60%, 20%, and 20% ; 70%, 15%, and 15% ; 80%, 10%, and 10% respectively.

The hyperparameters used in building a decision tree model were splitting criteria: Gain Ratio and stopping criteria: internal node was altered to leaf node if it satisfied one of the following: all observation in internal node was in the same class, no attribute had positive gain ratio, and the minimum number of observation on leaf node was 2 [5].

4.4Result

Figure 4 shows the comparison of model performance in training and testing dataset for both customer churn datasets without post-pruning and after Bayes Risk Post-Pruning was applied.The model performance to show the existence of overfitting conditions wasmeasured by the value of accuracy.

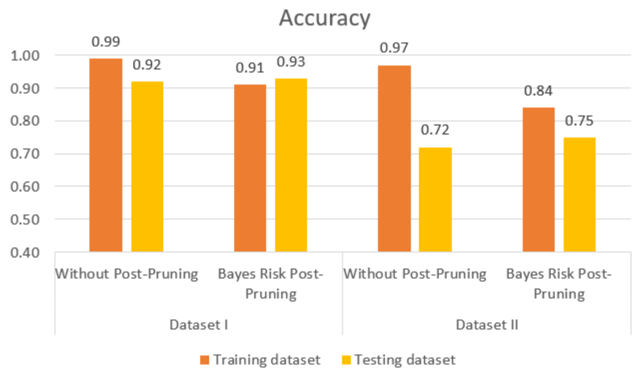

Fig. 4. Comparison of accuracy between training and testing datasets.

Based on Figure 4, the model performance for both datasetswithout post-pruning in the training dataset was better than the performance in the testing dataset. This showed the condition of overfitting that the model couldnot predict new data properly. Therefore Bayes Risk Post-Pruning is needed to overcome overfitting problems. After Bayes Risk Post-Pruning was applied to the model, the model performance in the training dataset decreased and the model performance in the testing dataset increased. This showed that the model performance in predicting new data was more accurate than before.

Figures5, 6, and 7 show the comparison of decision tree performance without post-pruning and after Bayes Risk Post-Prunning was applied in two dataset customer churn with three compositions of the training dataset (60%, 70%, and 80%). Model performance was measured by the value of accuracy, precision, and recall and shown in Figures5, 6, and 7 respectively.

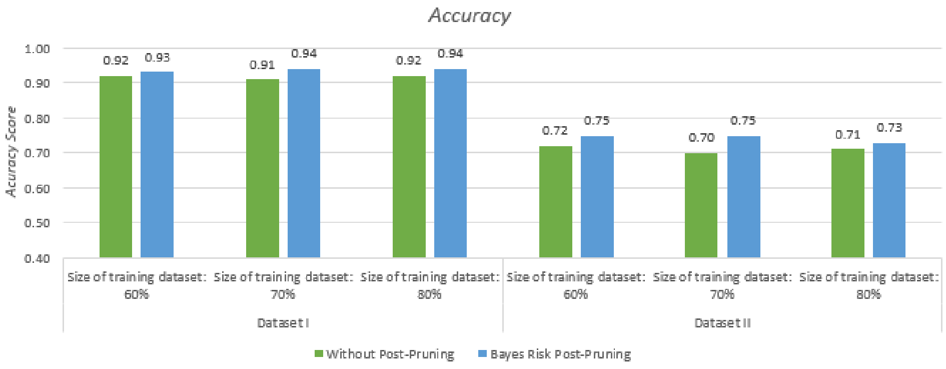

Fig. 5. Comparison of accuracy for three compositions of training dataset for dataset I and dataset II

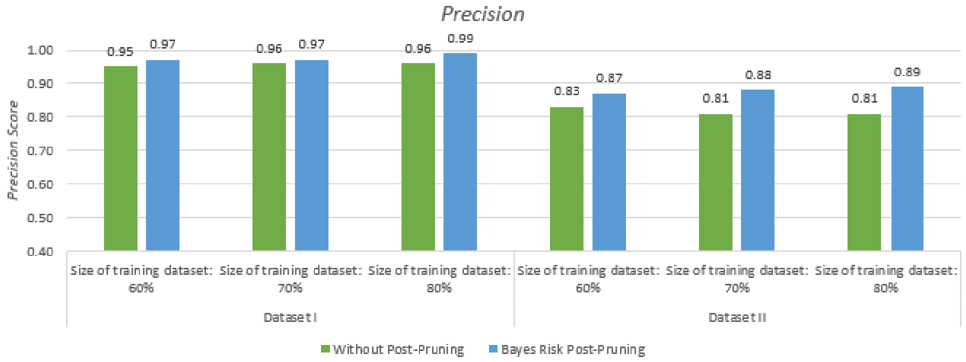

Fig. 6. Comparison of precision for three compositions of training dataset for dataset I and dataset II

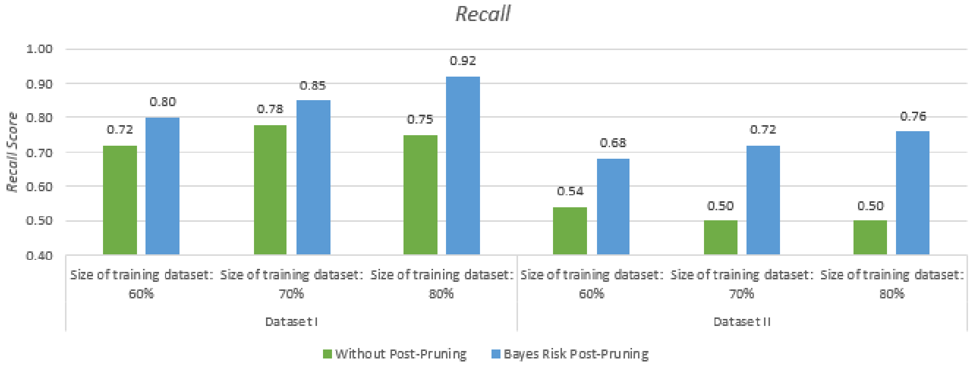

Fig. 7. Comparison of recall for three compositions of training dataset for dataset I and dataset II

From the three figures, the value of accuracy, precision, and recall for decision tree after Bayes Risk Post-Pruning applied were better than without post-pruning. Recall value was increasing higher than accuracy and precision by 17% in the first dataset and 26% in the second dataset if Bayes Post-Pruning was applied and the proportion of training dataset was 80%. It also showed that the larger the size of the training dataset was associated with higher performance of Bayes Risk Post-Pruning. The best model for both customer churn datasetswas obtained if Bayes Risk Post-Pruning was applied and the proportion of training dataset was 80%.Model performance in the first dataset (Telecom Customer Churn) was better than model performance in the second dataset (Telco Customer Churn) because the second dataset was larger than the first dataset.

The precision value for both datasets was higher than the accuracy and recall value. It showed that the model formed could save company costs because promo fees given to customers who had the potential to unsubscribe were right on target, but it not used for all customers who had the potential to unsubscribe because the model was less able to detect customers who had the potential to unsubscribe.

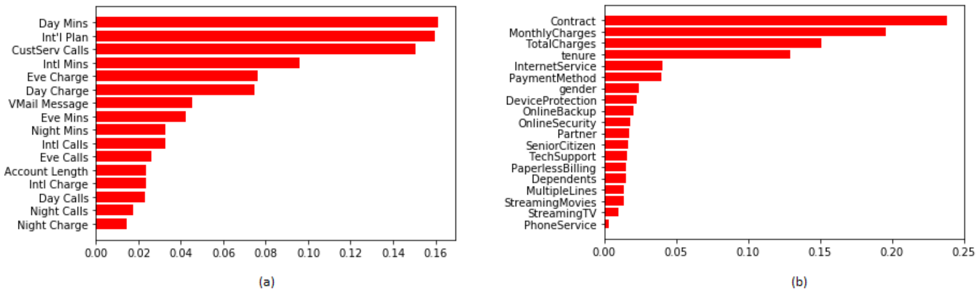

Fig. 8. Important attributes in (a) Telecom Customer Churn and (b) Telco Customer Churn dataset

FromFigures8 (a) and (b), all attributes contributed to the model decision tree. In Telecom Customer Churn dataset, the duration of daily telephone, whether or not there was a plan to subscribe to international telephone, and the intensity of customer contacting customer service were influential in determining whether the subject will churn or not. In Telco Customer Churn dataset, the terms of the customer contract, monthly charges amount, and total charges amount were influential in determining whether the subject will churn or not.

4 Conclusion and recommendation

Overfitting is a condition when model performance in the training dataset is better than the performance in the testing dataset that makes the model cannot predict new data properly. In this case, Bayes Risk Post-Pruning could overcome the overfitting problem. In the process of Bayes Risk Post-Pruning, each subtree in the decision tree is evaluated by comparing the risk associated with the parent node and the risk associated with its leaf node. The subtree is pruned if the risk of the parent node is smaller than the risk of its leaf nodesso that the parent node is altered to a leaf node. Bayes Risk Post-Pruning improved the ability of the decision tree model to predict new data better than the decision tree without post-pruning in two customer churn datasets. This was shown by the higher accuracy, precision and recall values when Bayes Risk Post-Pruning was applied. Bayes Risk Post-Pruning method was applied to three different proportions for training dataset (60%, 70%, and 80%) and the results showed that the larger the size of the training dataset was associated with higher accuracy, precision, and recall of the model.

There are several other post-pruning methods that were proposed in the literature, such as Reduced Error Pruning, Cost-Complexity Pruning, Pessimistic Error Pruning, Minimum Error Pruning. Yet, we have not examined the performance of these methods in this study. Therefore, in the next study, we plan to elaborate on these methods as well,hence it could be identified as the conditions where each method could outperform the other methods.

Acknowledgements

This research supported by the University of Indonesia with PITTA B 2019 research grant scheme, with ID number NKB-0665/UN2.R3.1/HKP.05.00/2019. We thank to all reviewers for the improvement of this article.

References

[1] Ahmed, A.M, Rizanes,A., Ulusoy, A. H. : A Novel Decision Tree Classification based on Post-Pruning with Bayes Minimum Risk. Plos One, 13(4): e0194168. Accessed on February 9, 2019 from https://doi.org/10.1371/journal. pone.0194168 (2018)

[2] Al Behadili, H. N. K., Ku-Mahamud, K. R., Sagban, R. : Rule Pruning Techniques in the Ant-Miner Classification Algorithm and Its Variants: A Review. Paper presented at conference 2018 IEEE Symposium on Computer Applications and Industrial Electronics (ISCAIE), Penang, Malaysia(2018)

[3] Berk, R. A. : Statistical Learning from a Regression Perspective. New York: Springer(2008)

[4] David, M. W. P. : Evaluation- From Precision, Recall and F-Factor to ROC, Informedness, Markedness and Correlation. Australia: School of Informatics and Engineering Flinders University of South Australia(2007)

[5] Mantas, Carlos J. and Abellan, Joaquin : Credal-C4.5: Decision Tree Based on Imprecise Probabilities to Classify Noisy Data. Expert System with Applications. Vol. 41, pp. 4625-4637(2014)

[6] Patel, N. and Upadhyay, S. : Study of Various Decision Tree Pruning Method with their Empirical Comparison in WEKA. International Journal of Computer Applications. Vol. 60, pp. 20-25.(2012)

[7] Santoso, B. et al. : Synthetic Over Sampling Methods for Handling Class Imbalanced Problems : A Review. IOP Conference Series: Earth and Environmental Science.Vol.58(2017)

[8] Webb, A. R. : Statistical Pattern Recognition (2nd ed.). John Willey and Sons,Inc, New Jersey (2002)

[9] Zhang, W. and Li, Y. : A Post-Prunning Decision Tree Algorithm based on Bayesian. International Conference on Computational and Information Sciences, pp. 988-991 (2013)