Classification and Regression Tree with Resampling for Classifying Imbalanced Data

Supajittree Boonamnuay, Nittaya Kerdprasop, and Kittisak Kerdprasop

***Abstract*—Data mining is the automatic process to find from data interesting and useful patterns for specific tasks such as predicting future data or classifying label or group of the new data items. Many data mining algorithms successfully applied to several real-life data are in a tree group. Among the tree-based algorithms, decision tree is the most popular and renowned one for its high accuracy on classifying data in general cases in which data in each class are quite equally distributed. But many datasets in real applications are imbalanced; amount of data in some group outnumber those in other group. Such uneven distribution among classes is a main reason why classification accuracy is not excellent even when using decision tree algorithm. Inefficiency is due to the case that in the tree growing phase, the algorithm tends to favor the majority data and ignores the minority data to be incorrectly classified. In the past many researchers try to solve this data imbalanced problem with many ways like over-sampling, under-sampling, cost-sensitive classification, or even ensemble of cost-sensitive decision tree. In this paper, we introduce a simplified method of learning classification and regression tree (CART) with resampling technique for classifying imbalanced datasets. We compare our proposed method with other methods based on several metrics including the precision on classifying the minority data as opposed to the classification on majority data, the overall accuracy regardless of minority nor majority classes, and the Matthews Correlation Coefficient (MCC). The use of MCC is suitable for imbalanced data because it takes into account all four classifying metrics: true positive, true negative, false positive, and false negative. The performance of our proposed method to combine resampling with CART is satisfied based on the MCC metric. From all five experimental imbalanced datasets, our method performs the best.**

***Index Terms*—Classification and regression tree, CART, resampling technique, imbalanced data, matthews coefficient correlation.**

I. Introduction

Data mining is a kind of data-oriented discovery science to find the patterns, important indexes or relationships from the existing databases [1]. There are many types of data mining tasks such as data classification, association rule mining, clustering, and forecasting. Among numerous potential tasks of knowledge discovery, data classification is the majority of data mining task that has been widely applied in many real applications and it is also the main mining task of our focus.

We especially aim at studying a classification problem of accurately partitioning and predicting data with imbalanced distribution among classes. This is called learning from imbalanced data [2], [3]. Typically, imbalanced data classification refers to a class of classification problems where the class distributions are not represented equally [2]. For example, suppose we have a 2-class (or binary) classification problem with 100 data instances (or rows) in total. Among these, 80 instances are labeled with class *a*, while the remaining 20 instances are labeled with class *b*. This is an example of imbalanced dataset and the ratio of data instances in class *a* to those in class *b* is 80:20, or the imbalanced ratio equals 4:1.

To deal with class imbalance, the most intuitive solution is to rebalance data with either under sampling the majority data, or over sampling data in the minority class [4], [5]. In this work, we are interested in balancing data with the bootstrapping method using resampling technique.

Our specific emphasis is on balancing data for a tree-based learning method. Tree learning is widely accepted for its easy interpretation nature. There exist several research trying to improve accuracy of tree-based learning over imbalanced data such as the work of Bartosz Krawczyk and teammates [6]. Their research introduced cost-sensitive decision tree ensembles for effective classification of imbalanced data. They tried to improve decision tree accuracy by assigning different weight to data in different imbalanced ratios.

Efficient learning from imbalanced data is still a challenging problem because imbalance is common in many applications [7]-[9]. Most classification datasets do not have exactly equal number of instances in each class, but a small difference often does not matter. There are, however, problems where a class imbalance is not just ignorance; it is the main concern of the application. For example, in commerce datasets like those that characterize fraudulent transactions are imbalanced. The vast majority of the transactions will be in the "Not-Fraud" class and a very small but important minority will be in the "Fraud" class.

We focus our concern regarding imbalanced data with the classification and regression tree (CART) method. Our special interest in CART algorithm is because this algorithm can classify all types of target data including both categorical and numeric. This algorithm has also been reported by many researchers that it yields good results. In medical domain [10], this algorithm can help efficient diagnosis based on patients' symptom. In economy [11], CART algorithm can help deciding the way to manage business plans. Also in environmental application [12], this algorithm can help to predict rainfall and groundwater level.

Manuscript received April 19, 2018; revised June 28, 2018. This work was supported by grants from Suranaree University of Technology through the funding of Data Engineering Research Unit and the Knowledge Engineering Research Unit.

The authors are with the School of Computer Engineering, Suranaree University of Technology (SUT), Thailand (corresponding author: Supajittree Boonamnuay; tel.: +66892865318; e-mail: eternity_faith@windowslive.com, nittaya@sut.ac.th, kerdpras@sut.ac.th).

II. Background Theories

A. Classification and Regression Tree

Classification and regression tree, or CART, is a classification method that builds a model from historical data. CART was firstly developed by Breiman, Freidman, Olshen and Stone in 1984 [13]. A CART tree is a binary decision tree in that in constructs a tree by splitting a node in half repeatedly resulting in two child nodes for each split. Tree construction begins with the root node that contains the whole learning samples. If data in the node are of mixing classes, that node has to be split. Splitting strategy is that the algorithm will search for all possible variables and all possible values in order to find the best split such that data in child nodes are of maximum homogeneity, or sometimes called purity.

The key idea of CART is recursive partitioning. The process of CART begins by taking all data for the consideration of all possible values of all variables for growing a tree. So it will select on variable or value that produces the best separation in the target attribute. If the value in focus is lower than the value at the separate point, that value will be placed on the left side of tree. For the value greater than or equal to the value at separate point, it will be sent to right side of tree like, as shown by example on Fig. 1. The tree will repeat this splitting process until it cannot find another best separate point the give the increase purity greater than the last separate point. The pseudocode of this tree growing process is illustrated in Fig. 2.

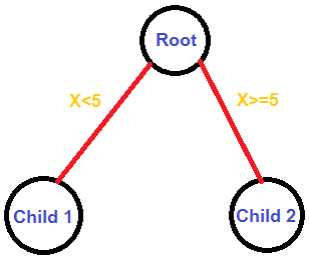


Fig. 1. Example CART model.

Classification and Regression Tree

1. Start at the root node.
2. For each ordered variable X,
   convert it to an unordered variable X' by grouping its values in the node into a small number of intervals
   if X is unordered, then set X' = X.
3. Perform a chi-squared test of independence of each X' variable versus Y on the data in the node and compute its significance probability.
4. Choose the variable X∗ associated with the X' that has the smallest significance probability.
5. Find the split set {X∗ ∈ S∗} that minimizes the sum of Gini indexes and use it to split the node into two child nodes.
6. If a stopping criterion is reached, exit.
   Otherwise, apply steps 2–5 to each child node.
7. Prune the tree with the CART method.

Fig. 2. The pseudocode of classification and regression tree.

The index that used for checking the best separate point in the pseudocode is Gini index that can be computed as in equation (1). Gini is a measure of impurity computed by counting the frequency of events that how often a randomly chosen data instance is wrongly labeled, given that that instance is to be randomly labeled based on distribution of class labels. For a binary classification with class positive and negative, p_{pos} is the probability that data instance in class positive being chosen, and $(1\text{-}p_{pos})$ is the probability that that instance is incorrectly labeled as negative. The other term can be interpreted in the same way with the class label negative, instead of positive.

$$Gini \text{ index} = p_{pos}(1\text{-}p_{pos}) + p_{neg}(1\text{-}p_{neg}) \quad (1)$$

B. Resampling

Bootstrap is a general purpose resampling technique for obtaining estimates of properties of statistical estimators without making assumptions about the distribution of the data [14]. This resampling method is often used to find

(1) standard errors of estimates,

(2) confidence intervals for unknown parameters, or

(3) *p* values for test statistics under a null hypothesis

Suppose *Y* has a cumulative distribution function, then $F(y) = P(Y<=y)$. If we have a sample of size *n* from $F(y)$, $Y1$, $Y2$, . . . , Yn, then the steps in computing resamples are as follows:

Step 1. Repeatedly simulate sample of size *n* from *F*.

Step 2. Compute statistic of interest.

Step 3. Study behavior of statistic over B repetitions.

Pretend that $Fn(y)$ is the original distribution of $F(y)$, sampling from $Fn(y)$ is thus equivalent to sampling with replacement from originally observed $Y1$, ..., Yn.

TABLE I: Confusion Matrix for Two Class Classification

		Actual Data	
		Positive	Negative
Predicted Data	Positive	*TP*	*FP*
	Negative	*FN*	*TN*

C. Classification Performance Evaluation

To evaluate each classification technique, we use the accuracy metric for assessing their overall performance. The computation of this metric is based on the values in confusion matrix [15] as shown in Table I and accuracy can compute as in equation (2).

$$Accuracy = \frac{(TP + TN)}{(TP+FN+FP+TN)} \quad (2)$$

where:

TP is the number of actual data from positive class and the model can correctly predict that data to be in a positive class,

TN is the number of actual data from negative class and the model can correctly predict that data to be in a negative class,

FP is the number of actual data from negative class but the model incorrectly predicts that data to be in a positive class,

FN is the number of actual data from positive class but the model predicts that the data incorrectly as in a negative class.

In our experiments, we also evaluate classification by class. This measurement is called precision and its computation is in equation (3). For the case of classifying data with imbalanced distribution among classes, many researchers [16], [17] use

Matthews Correlation Coefficient (MCC) as an effective metric for fair comparison because the MCC computes performance based on all values in the confusion matrix. MCC computation is shown in equation (4).

$$\text{Precision}_{\text{Positive}} = \frac{TP}{TP + FP}$$

$$\text{Precision}_{\text{Negaive}} = \frac{TN}{TN + FN} \quad (3)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (4)$$

III. Materials and Experimentation

A. Datasets and Methods

In this research, we use standard imbalanced datasets that publicly available for download from KEEL Repository [18]. The five datasets and their details are summarized in Table II. All datasets are two classes. We show class distribution as proportion of data instances in minority class to those in the majority class. The imbalanced ratios (IR) are computed as the fraction of instances in majority to instances in minority class. IR equals to 1 means the data are well balance. The higher IR infers the more imbalanced situation among classes.

Our research methodology is that firstly classifying the selected datasets using the tree-based algorithms including decision tree induction, CART, AdaBoost, and bagging of decision trees. Then perform on the same datasets our proposed method of decision tree learning using CART with the applied resampling technique.

TABLE II: Data Sets Used in the Experiments

Dataset	Features	Objects	No. Classes	Class distribution	IR
Pima	8	768	2	268:500	1.87
Yeast	8	1484	2	429:1055	2.46
Vehicle	18	846	2	199:647	3.25
Segment	19	2308	2	329:1979	6.02
Page-blocks	10	5472	2	559:4913	8.79

B. Experimental Setup

To test the performance of the proposed method (CART + resampling), we compare it against the other four techniques (decision tree, CART, AdaBoost, and Bagging). Comparative performance metrics are precision by class, overall accuracy, and MCC. These metrics are computed from the confusion matrix that to be obtained by running a classifier model ten times using the 10-fold cross validation method, which is conceptually shown in Fig. 3. We use 10-folds cross validation because we want to fairly compare the models using every data instance as train and test data in every model comparison. To apply 10-fold cross validation, we have to separate our dataset into 10 parts and repeat the experiment 10 times. At round one, we use parts 1-9 as training set and use part 10 as test set. Then in round two, we use parts 1-8 and 10 as training set and use part 9 as test set. We repeatedly do it in such manner 10 times and average, precision, accuracy, and MCC values from that 10 rounds to compare performance of each classification technique.

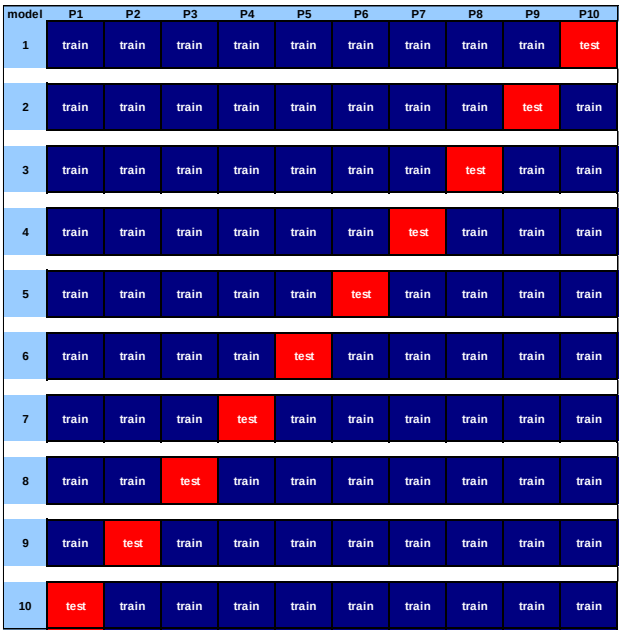
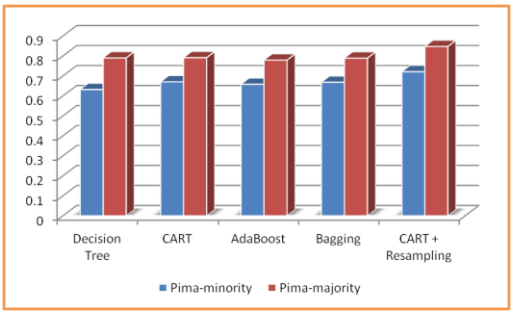


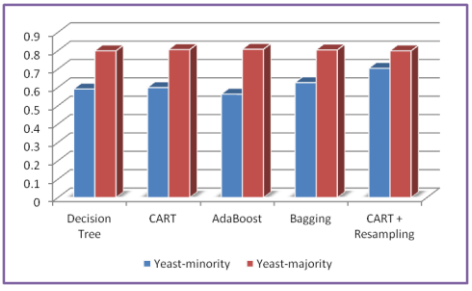
Fig. 3. Building and testing a model with10-folds cross validation.

IV. Experimental Results and Discussions

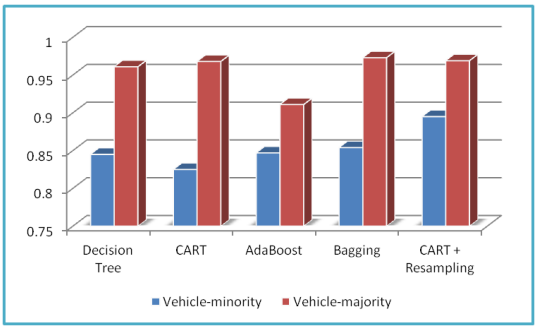
To observe classification performance of the proposed CART + resampling method on imbalanced data, we firstly analyze its precision on classifying minority cases as compared to the classification on the majority cases. The results are shown in Fig. 4. It can be noticed from the results that resampling can help improving classifying data in the minority class as well as yielding good precision on the majority class. It is true in almost all datasets, except the segment dataset that even though resampling can improve the precision of classifying majority cases, the precision on classifying minority cases is still low, comparative to the bagging technique. From observing precision on predicting minority and majority classes, we can conclude that CART + resampling performs well on four out of five datasets.



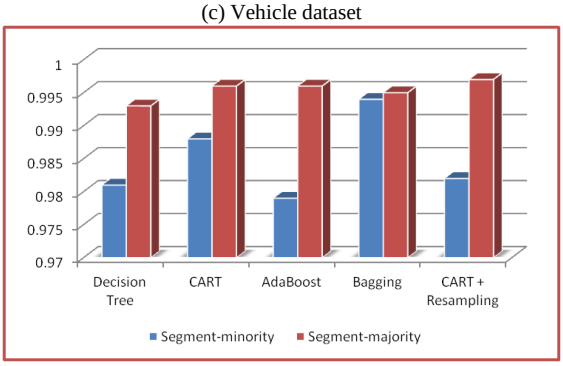
(a) Pima dataset



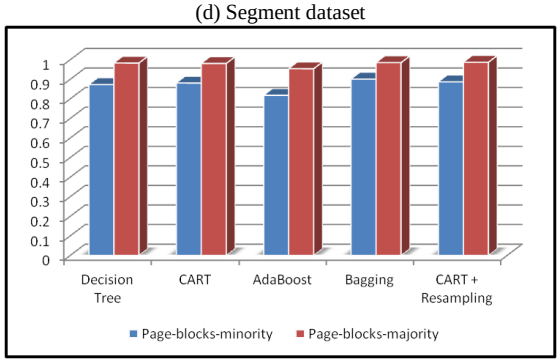
(b) Yeast dataset



(c) Vehicle dataset



(d) Segment dataset



(e) Page-blocks dataset

Fig. 4. Precision by class and by dataset of the CART + Resampling method comparative to other classification methods.

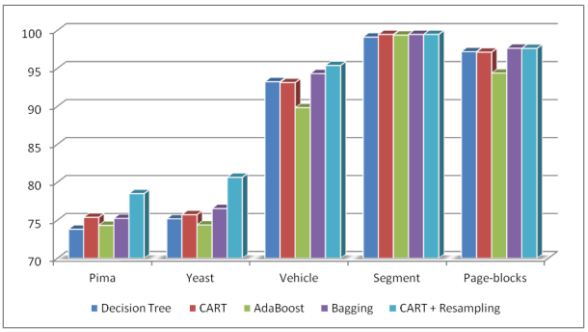


Fig. 5. Overall classification accuracy of the studied CART + Resampling method compared against other four methods

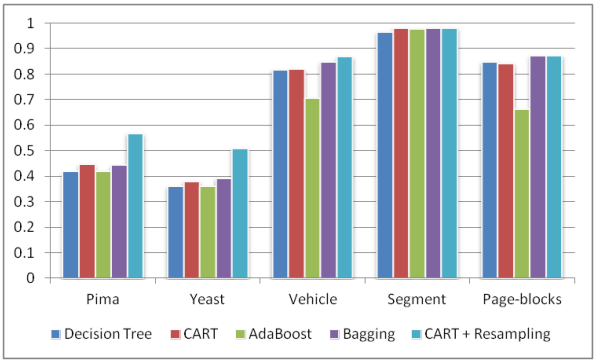


Fig. 6. Class imbalanced classification based on the Matthew correlation coefficient metric of each classification method on different datasets.

We then check the overall accuracy performance of the CART + resampling method. The results are summarized and illustrated in Table III. The graphical comparison is also shown in Fig. 5. By ignoring importance of minority versus majority and just evaluate the overall predictive accuracy, we find that our proposed method performs almost the best in every dataset, except in the page-blocks dataset that bagging method is a little bit better with insignificant difference.

To take into account both precision on predicting minority and majority classes as well as penalty on misclassification, we compare the models' performance with the MCC metric. The results are illustrated in Table IV and graphically shown in Fig. 6. When consider both correct and incorrect classification cases, we can now clearly see the power of CART + resampling method as it performs the best in every dataset.

TABLE III: OVERALL ACCURACY FOR EACH CLASSIFICATION TECHNIQUE ON DIFFERENT DATASETS

Dataset	Decision Tree	CART	AdaBoost	Bagging	CART + Resampling
Pima	73.8281	75.3906	74.349	75.2604	**78.5156**
Yeast	75.2022	75.7412	74.3935	76.5499	**80.6604**
Vehicle	93.2624	93.1442	89.8345	94.3262	**95.3901**
Segment	99.1334	**99.4801**	99.3934	**99.4801**	**99.4801**
Page-blocks	97.2222	97.1491	94.3896	**97.6608**	97.6425

TABLE IV: MATTHEWS CORRELATION COEFFICIENTS FOR EACH CLASSIFIER ON EACH DATASET

Dataset	Decision Tree	CART	AdaBoost	Bagging	CART + Resampling
Pima	0.417	0.444	0.417	0.441	**0.566**
Yeast	0.360	0.379	0.360	0.391	**0.507**
Vehicle	0.815	0.817	0.704	0.847	**0.866**
Segment	0.964	**0.979**	0.975	**0.979**	**0.979**
Page-blocks	0.847	0.841	0.660	0.870	**0.871**

V. CONCLUSION

To improve the performance of imbalanced data classification using tree learning algorithms, we can apply the method or technique to improve accuracy by resample the datasets. For classifying categorical and numerical data, classification and regression tree (CART) algorithm is renowned for better classifying imbalanced data than normal decision tree. We propose in this work that we can further improve the performance of CART by handling the imbalanced data through the resampling technique.

The experimental results show that our proposed technique can help improving classification performance when several measurements including precision by class, overall accuracy, and Matthews Correlation Coefficient (MCC). The MCC metric is the most discriminative measurement confirming the power of resampling when applied to the CART algorithm. This method has been proven work well on datasets with numerous imbalanced ratios (IR); in our experiments the IR ranges from 1.87 up to 8.79. We notice that the CART + resampling is extremely powerful when IR is lower than 4.

For future work, we plan to further our investigation that how much data is enough to effectively represent the whole dataset. Such knowledge can facilitate our application of bootstrapping for under-sampling and over-sampling as well.

REFERENCES

[1] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufmann, 2006.

[2] N. Chawla. "Data mining for imbalanced datasets: An overview," in *Data Mining and Knowledge Discovery Handbook*, O. Maimon and L. Rokash, Eds., pp. 875-886, Springer, 2010.

[3] B. Krawczyk, "Learning from imbalanced data: Open challenges and future directions," *Progress in Artificial Intelligence*, vol. 5, no. 4, pp. 221-232, 2016.

[4] N. Chawla, K. Bowyer, L. Hall, and W. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif Int Res*, vol. 16, no. 1, pp. 321-357, 2002.

[5] R. Dubey, J. Zhou, Y. Wang, P. Thompson, and J. Ye, "Analysis of sampling techniques for imbalanced data: An N=648 ADNI study," *Neuroimage*, vol. 87, pp. 220-241, 2014.

[6] B. Krawczyk, M. Wozniak, and G. Schaefer, "Cost-sensitive decision tree ensembles for effective imbalanced classification," *Applied Soft Computing*, vol. 14, pp. 554-562, 2014.

[7] P. Gutierrez, M. Lastra, J. Benitez, and F. Herrera, "SMOTE-GPU: Big Data preprocessing on commodity hardware for imbalanced classification," *Progress in Artificial Intelligence*, 2017.

[8] S. Pouyanfar and S. Chen, "Automatic video event detection for imbalanced data using enhanced ensemble deep learning," *Int J of Semantic Computing*, vol. 11, no. 1, 2017.

[9] F. Li, S. Li, C. Zhu, X. Lan, and H. Chang, "Cost-effective class-imbalance aware CNN for vehicle localization and categorization in high resolution aerial images," *Remote Sensing*, vol. 9, no. 6, 2017.

[10] S. C. Lemon, J. Roy, and M. A. Clark, "Classification and regression tree analysis in public health: Methodological review and comparison with logistic regression," *Annals of Behavioral Medicine*, vol. 26, no. 3, pp. 172-181, 2003.

[11] A. I. Irimia-Dieguez, A. Blanco-Oliver, and M. J. Vazques-Cueto, "A comparison of classification/regression trees and logistic regression in failure models," *Procedia Economies and Finance*, vol. 23, pp. 9-14, 2015.

[12] Y. Zhao, Y. Li, L. Zhang, and Q. Wang, "Groundwater level prediction of landslide based on classification and regression tree," *Geodesy and Geodynamics*, vol. 7, no. 5, pp. 348-355, 2016.

[13] L. Breiman, J. H. Friedman, R. Olshen, and C. J. Stone, *Classification and Regression Tree*, Wadsworth & Brooks/Cole Advanced Books & Software, Pacific California, 1984.

[14] A. Benner. (October 2016). *Resampling and the Bootstrap*. [Online]. Available: http://www.bioconductor.org/workshop/203/NGFN03/resamplig.pdf

[15] D. M. Powers, "Evaluation: from Precision, recall and F-measure to ROC, informedness, markedness and correlation," *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37-63, 2011.

[16] B. W. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," *Biochimica et Biophysica Acta (BBA) – Protein Structure*, vol. 405, no. 2, pp. 442-451, 1975.

[17] S. Boughorbel, F. Jarray, and M. El-Anbari, "Optimal classifier for imbalanced data using Matthews correlation coefficient metric," *PLoS ONE*, vol. 12, no. 6, article e0177678, June 2017..

[18] J. Alcala-Fdez, A. Ferndedez, J. Luengo, J. Derrac, S. Garcia, L.Sanchez, and F. Herrera, "Keel data-mining software tool: data set repository integration of algorithms and experimental analysis framework," *Journal of Multiple-Valued Logic and Soft Computing*, vol. 17, no. 2-3, pp. 255-287, 2011.

S. Boonamnuay is currently a master student with the School of Computer Engineering, Suranaree University of Technology, Thailand. She received her bachelor degree in computer engineering from Suranaree University of Technology, Thailand, in 2015. Her current research of interest includes data mining, classification and regression tree, and imbalanced data classification.

Kittisak Kerdprasop is an associate professor at the School of Computer Engineering and Chair of the School. He is also the head of Knowledge Engineering Research Unit, Suranaree University of Technology, Thailand. He received his bachelor degree in Mathematics from Srinakarinwirot University, Thailand, in 1986, master degree in computer science from the Prince of Songkla University, Thailand, in 1991 and doctoral degree in computer science from Nova Southeastern University, U.S.A., in 1999. His current research includes machine learning, artificial intelligence and probabilistic knowledge bases.

Nittaya Kerdprasop is an associate professor and the head of Data Engineering Research Unit, School of Computer Engineering, Suranaree Uiversity of Technology. She received her B.S. in radiation techniques from Mahidol University, Thailand, in 1985, M.S. in computer science from the Prince of Songkla University, Thailand, in 1991 and Ph.D. in computer science from Nova Southeastern University, U.S.A., in 1999. Her research of interest includes knowledge discovery in databases, data mining, artificial intelligence, logic and constraint programming, deductive and active databases.