JSTR
ASEAN Journal of Scientific and Technological Reports

Research article

Efficiency Enhancement with Rule-Based Method for Credit Classification

Suriyan Anuwak[1], Krich Sintanakul[2], Charun Sanrach[3*]

[1] Department of Computer Education, Faculty of Technical Education, King Mongkut's University of Technology North Bangkok, Bangkok, 10800, Thailand; sanuwak.kmutnb@gmail.com
[2] Department of Computer Education, Faculty of Technical Education, King Mongkut's University of Technology North Bangkok, Bangkok, 10800, Thailand; krich.s@fte.kmutnb.ac.th
[3] Department of Computer Education, Faculty of Technical Education, King Mongkut's University of Technology North Bangkok, Bangkok, 10800, Thailand; jsr@kmutnb.ac.th
* Correspondence: e-mail@e-mail.com; sanuwak.kmutnb@gmail.com

Abstract: The efficiency enhancement with the Rule-based Method is a data mining technique to study the relationship between credit borrowers and deciding credit approval and reduce the risk of bad debt in the future. This research aims to efficiency enhancement with the Rule-based method for credit classification, which is the credit types data, numeric, and nominal used for the category from the cooperative savings database by using the Gain Ratio as a measurement unit of the sampling (Entropy) and filter to select important variables. Therefore, the researcher uses the K-fold cross-validation method by dividing the data to perform the test into equal K-part numbers into training and testing data sets. Then Rule-based approach of data mining techniques in WEKA software version 3.9.4viz Decision Table, RIPPER (JRip), OneR, and Partial Rule (PART) to efficiency enhancement of the model for credit classification to get more accurate and reliable by measuring the efficiency of the model with Recall, Precision, and F-measure. The results of the research can be found that both the Gain Ratio and the outlier data filter can make the efficiency of the model with the Rule-based method using the Partial Rule to get the highest Recall value of 4.1%, the highest Precision value of 4.0%, and highest F-measure value by 5.4%. Besides, the Partial Rule can make the model's efficiency for credit classification get a Recall of 86.1%, Precision of 85.9%, and F-measure of 85.6%. Thus, all values were more efficient than the Decision Table, JRip, and OneR.

Keywords: Gain Ratio; Rule-Based; Efficiency; Model

Citation:
Anuwak, S.; Sintanakul, K.; Sanrach, C. Efficiency Enhancement with Rule-Based Method for Credit Classification. *ASEAN J. Sci. Tech. Report.* **2022**, *25*(2), 10-18. https://doi.org/10.55164/ajstr.v25i2.244172.

Article history:
Received: May 14, 2021
Revised: March 25, 2022
Accepted: April 19, 2022
Available online: June 26, 2022

1. Introduction

Currently, financial institutions lending to incur income from interest and aims to expand the size of the credit amount regardless of the quality of the released loans will cause bad debt problems to affect financial institutions lacking liquidity in the system. Generating the amount of more Non-performing Loan must set aside reserves money for Non-performing loan that does not incur the additional income that financial institutions are unable to expand lending as a result of the problem of making financial institutions must set aside debt reserves money a lot of money. When the bad debt or risk assets are high, that is necessary to build financial stability and trust among members, depositors, shareholders, and stakeholders [1]. To improve consideration and credit approval by the rule-based methods, it is a data mining technique to assist in

credit analysis in predicting credit borrowers from the model. It is a search for patterns or rules to build new knowledge with large amounts of hidden data to make credit analysis more efficient. Therefore, this research presents learning with rule-based methods, namely Decision Table, RIPPER (JRip), OneR, and Partial Rules (PART) and uses the principle of the K-fold cross-validation to separate data sets into a learning data set and testing data set, including the Gain Ratio is a measurement unit of the sampling (Entropy) and filter to select important variables in comparing the model efficiency with the value of Recall, Precision, and F-measure respectively. It was used to enhance the efficiency of the credit classification model obtained from the credit database of the savings cooperatives.

2. Materials and Methods

The purpose is the efficiency enhancement of the model using the Gain Ratio for selecting variables' importance and filtering out the outliers.

3. Theory and Literature Reviews

3.1 Theory

3.1.1 Data classification with Rule-based [2] is a model that shows a set of rules to have pattern characteristics 'IF-THEN' by one of the rules shown in the form below.

IF condition THEN conclusion

Such as R1: IF age = youth AND student = yes THEN buys_computer = yes
For example, the R1 consists of two parts: the condition of 'IF' called 'rule antecedent' or 'precondition', which includes a set of attributes that comprise the various conditions. The R 1 rule consists of attributes that indicate characteristics of people of both attributes: age = youth and student (student = yes) [3]. The condition of the rule will be linked to various attributes together with an AND symbol to indicate that the data must satisfy the specified conditions. In the second part, it is called 'rule consequence', and in the example of a rule, 1R can be written as a rule.

$$R1: (age = youth) \wedge (student = yes) \Rightarrow (buys\_computer = yes)$$

If the record data x one according to the condition of the rule antecedent can tell that the R 1rule covers x record data or x record is covered by the R 1rule. Therefore, when the data set is composed of multiple records data, it will be able to find the percentage of all the records in the data set covered by the R 1rule, which it will call the coverage percentage as coverage, which can be calculated as follows.

$$coverage\ (R) = n_{covers} / |D|$$

Where n_{covers} is the number of records in the data set covered by the R rule, and $|D|$ is the number of records of all data in the dataset.
The rule's coverage value is used to evaluate the rule's data classification efficiency and the measurement of accuracy in the data classification of rules can be calculated as follows.

$$Accuracy\ (R) = n_{correct} / n_{covers}$$

Where $n_{correct}$ is the number of records was classified correctly under the rules R

3.1.2 The Rule-based classification techniques in data mining used in research are as follows.

3.1.2.1 Decision Table Technique: [4]it is managing the group in a decision table style by taking the results of learning from the input data to have similar data characteristics to be in the same group by the last attribute of the table is used for determining the group to want to divide also.

3.1.2.2 PART Technique: [5] it is managing the group by the results in the form of rules to take data of each attribute as a condition in deciding the characteristics of the similar data to be in the same group by the data value of the attribute has the continuous value to use the signs "<", "<=", "=>", ">". In considering the discontinuous data part will use the sign "=" and use the condition "AND" for linking between attributes will link till the last attribute to divide the groups.

3.1.2.3 RIPPER Technique: [6] It is a rule that results from dividing the data classification rules into groups according to the target class and then using the pruning principle to select the rule more effectively than the specified error rate to use the data prediction.

3.1.2.4 OneR Technique: [7] It is an algorithm that provides an easy and intuitive way by Rule-based principles of creation for the classification by choosing the minor error value attribute for just one attribute to be the class predictor of the data which will be got the data classification rule of a smaller number.

3.1.3 Gain Ratio

GainRatio value is a measurement unit of the sampling (Entropy) to be important in predicting and the usability of the outlier data filter. When using knowledge gain to split the dataset, will be made inclination by the considered attribute has a considerable occurrence value to make the higher knowledge gain: $\text{Gain}(A) = \text{Info}(D) - \text{Info}_A(D) = \text{Info}(D) - 0 = \text{Info}(D)$ from the inclination may arise from the knowledge gain and Info(D), known as entropy of D is the mean of the amount of data required to identify a category of records in data set which is dependent on the ratio of the number of the records corresponding to each category. [8] The researcher tries to reduce the inclination by developing new data division indicators with normalization for the knowledge gain and value calculation 'split information'.

$$\text{SplitInfo}_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2\left(\frac{|D_j|}{|D|}\right)$$

By SplitInfo_Avalue (D) shows the amount of data to consider in the data set D into v subsets according to the value in attribute A. From $\text{SplitInfo}_A(D)$ value, the calculation can find the gain ratio below.

$$\text{GainRatio}(A) = \frac{\text{Gain}(A)}{\text{SplitInfo}(A)}$$

3.2 Literature Reviews

Almost every technique in data mining can create effective rule-based prediction models. Still, they cannot make effective rule-based prediction models from the real data and many researchers were applied to select more efficient variables as follows.

Phadung and Jaree [9] propose to develop an intrusion detection model and use four classification techniques, including Decision table, Naive Bayes, RIPPER, and PART decision list in data mining. The statistics used the percentage, Precision, Recall, and F-Measure. The research findings showed that RIPPER has the highest Precision, up to 99.00%. Then, the PART decision list has a precision up to 98.20%. Follow by Decision Table has the highest Precision, which is up to 97.50%. The lowest Precision is Naive Bayes 49.40%. Therefore, the RIPPER model has the highest average significantly.

Nongluk et al. [10] compare the efficiency of data mining classification techniques between K-nearest Neighbor (K-NN), Rule-based, and Decision trees by using the various parameters. The variant standard data sets such as Spambase, Zoo, and Nursery from the UCI Machine Learning Repository were used by selecting different parameters and characteristics. The accuracy and efficiency of each technique were analyzed and compared by specifying various parameters. Moreover, the data set was tested in varied 12 models per test set. The most accurate approach has more effective and appropriate for the characteristics of that data set. Finally, the analysis results have shown that K-NN and Decision Tree techniques are suitable for the slight attribute data set. At the same time, the Rule-based methods are appropriate for the massive attribute data set.

Kittisak [11] proposes a relationship between health conditions and diseases. The partial Rules algorithm outperformed the Decision tree. The Precision and recall are 88.60% and 89.20%, respectively, and the F-measure is 88.80%. A group of experts was invited to examine the obtained rules. Some inappropriate

attributes and rules were eliminated. The outcome was 44 rules and these rules were embedded into the automatic health screening system that allows people to examine their health condition by themselves.

Fadi et al. (2012) [12] propose investigating the Arabic text classification problem. In particular, four different Rule-based classification approaches: Decision trees (C4.5), Rule Induction (RIPPER), Hybrid (PART), and Simple Rule (One Rule), are evaluated against the published Corpus of Contemporary Arabic text collection. Analyzing the experiment results, we determine the most suitable classification algorithms for classifying Arabic texts.

Shafiullah et al. [13] propose improving the existing approach's performance. A rule-based learning method using statistical analysis has been presented in this paper to select a unique classifier for the same application. This study has been conducted using six classifiers, namely REP Tree, J48, Decision Stump, IBK, PART, and OneR, with twenty-five datasets. WEKA tool has been used in this study to develop the prediction model.

Rajalakshmi et al. [14] propose to help improve the quality of data. Discretization is dividing a continuous attribute into a finite set of intervals to generate an attribute with a few distinct values. This paper handle continuous values of the iris data set taken from the UCI machine learning repository. Data set used in various classification algorithms, namely J48, Random Forest, REP Tree, Naive Bayes, RBF network, OneR, BF Tree, and Decision Table. The performance measures are accuracy and error rate noted before and after discretization.

Shathian [15] proposes using data mining techniques to apply knowledge from data mining as the guidelines decision support credit approval. To study this experiment with harmonized data group classification group 3 technical result rule is Decision Tree: C 4.5, Decision Tree: Partial Rules and Decision Tables trial results found that classification groups Decision Tree: Partial Rules, the result is maximum is 72.7 % by data classification.

4. Research Methodology

In this research, the researcher divided the research methods as follows.

4.1 Data Collection: All the data are collected in14 attributes, especially age, sex, status, revenue, occupation, province etc. Also, there is a data set to use in the study to be a number of 2,404 records from the savings and credit cooperative limited database. Afterwards, the data is converted into the CSV file format to use in the credit consideration, as shown in Table 1 and Figure 1.

Table 1. Attributes for the research

Attributes	Attributes Type	Meaning
Age	Numeric	Age
Sex	Nominal	Male,Female
Status	Nominal	Family, Single
Member Period	Nominal	Year
Salary	Numeric	Income
Capital stock	Numeric	Number of capital stock
Collateral	Nominal	Estate, Capital stock, Bank account
Approval	Numeric	Credit Limit Current
Instalment	Numeric	Payment
All instalments	Numeric	Term
Objectives	Nominal	Consumption,House,Car,Business
Occupation	Nominal	Governmental, Private ,Freelance
Address	Nominal	Rural ,Suburb , Town
Class	Nominal	Normal, Risk

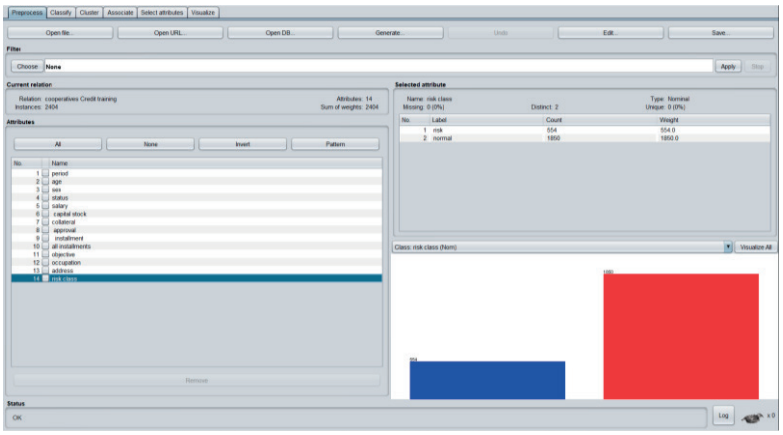

Figure 1. The number of credit data set before preprocessing

4.2 Data preprocessing: [16] Data transformations of the attributes analysis or the selection of variables are important for forecasting with Gain Ratio. Outlier data filters with the Gain Ratio will be used for creating trees. If any data is not classified with C4.5, the outlier data will be removed from the above preparation dataset, decreasing the data by only 1.33,% as shown in Table 2.

Table 2. The number of data before Rule-based methods

Class	Raw data	GainRatio	Outlier filter
Normal	1,850	1,850	1,823
Risk	554	554	549
Total	2,404	2,404	2,372

As for filtering out the outliers, it is an outlier by detecting an error value in the data set that is higher than or lower than normal. It cannot be classified in any group and falls outside the group. For example, no outlier is the value at 2,372 and yes outlier is the value at 32. Thereby, it should be discarded by an outlier of 32 values as in Figure 2 and Figure 3 . Besides, one raw data set was not filtered out of the attributes. In contrast, the same raw data set was filtered out the attributes by Gain Ratio. Afterwards, both groups were predicted by Rule-based methods. Therefore, it was found that the results of the testing are different from Table3.

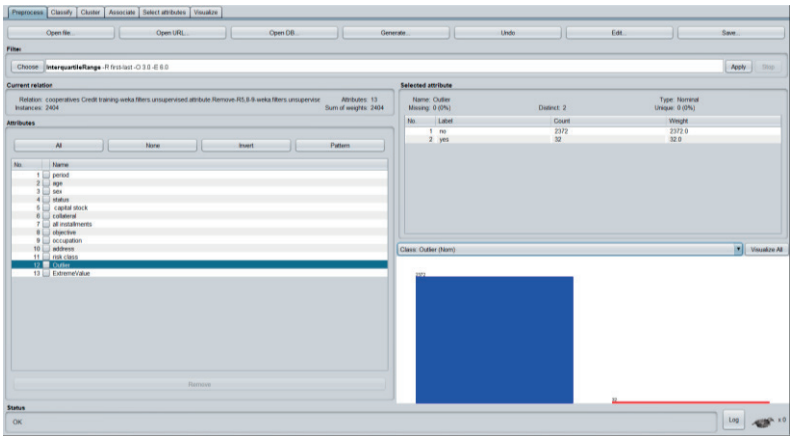

Figure 2. Before filtering out the outliers of the credit data set

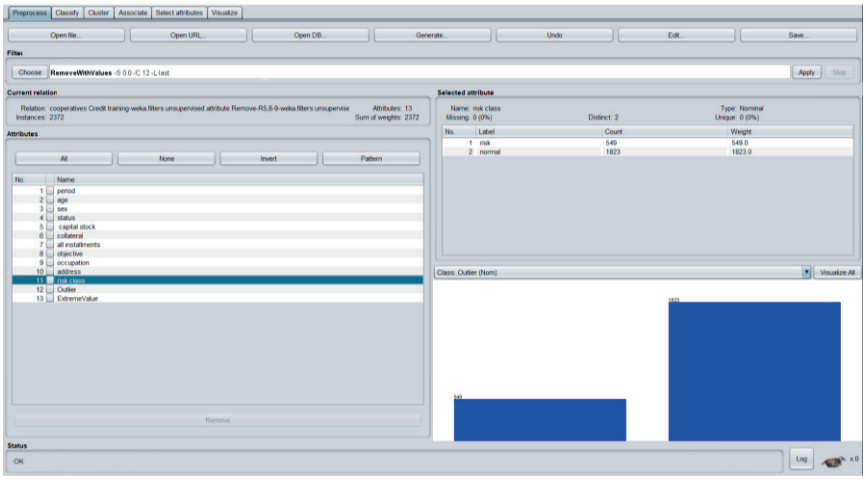

Figure 3. The number of credit data set after filtering outlier

4.3 Modeling: A prediction modelling by Rule-based can choose four methods, namely Decision Table, PART, RIPPER, and OneR. Therefore, the efficiency measurement of the model experiments to divide 2data sets into a training data set and a testing data set using the -10fold cross-validation principle by dividing the data into several parts)K-fold cross-validation) for example, 10-fold cross-validation. It is to divide the data into 10 parts. Each of the parts has a number of the data equal parts. After that, [17] the data of one part is used to test the model's efficiency to cycle until the complete number of splitting and use Recall, Precision, and F-measure value demonstrates the model's efficiency below.

4.3.1 Recall is a measurement of correct of the model by considering the separation of each class: Recall = TP / TP+FN

4.3.2 Precision is a measurement of the model accuracy by considering the separation of each class: Precision = TP / TP+FP

4.3.3 F-measure is a simultaneous measurement of Precision and recall of the model by considering the separation of each class: F-measure = 2*(Precision*Recall) / Precision+Recall

Where accuracy is a measure of the model's accuracy by considering including all classes.

5. Results

Classification by Decision Table, JRip, OneR, and PART with each method and the efficiency measurement of the model by Recall, Precision, and F-measure. The test results of Table 3, Figure 3, Figure 4, and Figure 5 are as follows.

Table 3. Recall, Precision, and F-measure of Rule-based methods

Rule-based methods (10-fold)	Recall			Precision			F-measure		
	Raw data	GainRatio	Outlier filter	Raw data	GainRatio	Outlier filter	Raw data	GainRatio	Outlier filter
DecisionTable	0.639	0.700	0.683	0.426	0.514	0.490	0.511	0.593	0.571
JRip	0.689	0.750	0.743	0.789	0.814	0.812	0.590	0.668	0.664
OneR	0.721	0.730	0.725	0.738	0.694	0.700	0.672	0.666	0.665
PART	0.762	0.820	**0.861**	0.766	0.819	**0.859**	0.740	0.802	**0.856**

Recall values were derived after selecting the variable which the method of Decision Table and JRip get higher values at 6.10%. As for the technique of OneR, and PART is more value at 0.9% and 5.8%. Afterwards,

the variable selection and filtering of the outlier decreased efficiency by 1.7%, 0.7%, and 0.5%, except that PART increased by 4.1%. Therefore, it was found that the results of the method were shown the recall of the model viz PART was the highest value at 86.10%, OneR was the value at 73.0%, JRip was the value of 75.0%, , and Decision Table was the value at 70% as Figure 4.

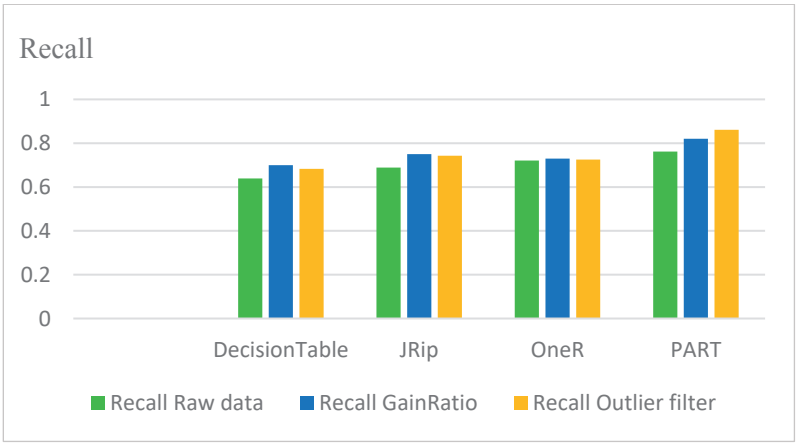

Figure 4. Recall values of the prediction rule

Precision values were derived after selecting the variable which the method of Decision Table, JRip and PART get higher values at 8.8%, 2.5%, and 5.3%. The method of OneR decreased by 4.4%. Afterwards, the variable selection, and filtering of the outlier, the efficiency is reduced by 2.4% and 0.2% except for OneR, and PART is increased by 0.6% and 4.0%. Therefore, it was found that the results of the method were shown Precision of the model viz PART was the highest value at 85.9%, OneR was the value at 73.8%, the JRip was the value at 81.4%, and Decision Table was the value at 51.4% as Figure 5.

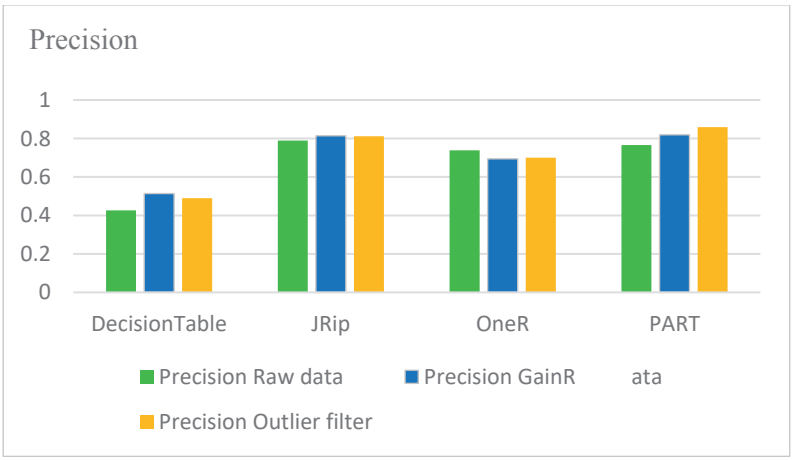

Figure 5. Precision values of the prediction rule

F-measure values were derived after selecting the variable in which the Decision Table, JRip, and PART get higher values at 8.2%, 7.8%, and 6.2%. As the method of OneR is decreased by 0.6%. Afterwards, the variable selection, and filtering of the outlier, the efficiency is reduced by 2.2%, 0.4%, and 0.1%, except for PART is increased by 5.4%. Therefore, it was found that the results of the method were shown the F-measure of the model viz PART was the highest value at 85.6%, OneR was the value of 67.2%, JRip was the value of 66.8%, and Decision Table was the value at 59.3% as Figure 6.

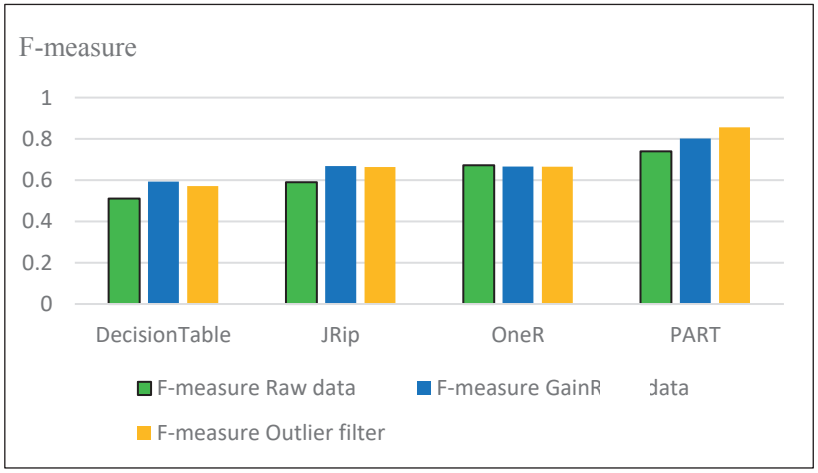

Figure 6. F-measure values of the prediction rule

6. Conclusions

The efficiency of the predictive rule model (Rule-based) uses the Gain Ratio to select the attributes and choose the data with the outlier to remove them. The model was used to compare four methods: Decision Table, JRip, OneR, and PART. Consequently, the results were found that both the Gain Ratio and the outlier filter to make the model's efficiency get PART method by the higher recall at 4.1%, higher Precision at 4.0%, and higher F- measure at 5.4%. Therefore, the PART method created the model with a Recall of 86.1%, Precision of 85.9%, and F-measure of 85.6%. Hence, all values were higher than the method of Decision Table, JRip, and OneR, respectively. Furthermore, the researcher studied the data structure, data cleaning, and data reduction to choose an appropriate PART model to apply as a guideline for future credit approval of savings cooperatives.

Author Contributions: The first author built models, the efficiency enhancement results. The second author contributed to editing and revising the manuscript. The last author proposed the research idea for the data mining technique and was responsible for examining the model results. All authors read and approved the final manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] Thongkumpra, M.; Rattanasiriwongwut, M. Application of techniques to Classify data to predict customer owes doesn't cause Income: A Case Study of a financial institution, *The Twelfth National Conference on Computing and Information Technology NCCIT* 2016; pp. 605-610.

[2] Rick L. Lawrence; Andrea Wright. Title of the chapter. "Rule-Based Classification Systems using Classification and Regression Tree (CART) Analysis. *Photogrammetric Engineering & Remote Sensing*, 2011; 67(10), pp. 1137-1142.

[3] Daoruang, B.; Mingkhwan, A.; Sanrach, C. The Comparison of Data Classification Efficiency for Decision in the Selection of Information Technology Students Academic Subjects, Faculty of Technology and Industrial Management, King Mongkut's University of Technology North Bangkok, Prachinburi Campus. June 16, 2018; pp. 1-9.

[4] R, Kohavi. The Power of Decision Tables. presented at the in *8 th European Conference on Machine Learning*, 1995; pp. 2-17.

[5] E, Frank; I. H. Witten: Generating Accurate Rule Sets Without Global Optimization. presented at the In *Fifteenth International Conference on Machine Learning*,1998; pp. 1-15.

[6] Cohen, WW. Fast Effective Rule Induction. In *Proceedings of the Twelfth International Conference on Machine Learning*; July 9-12; Tahoe. California: Morgan Kaufmann, 1995; pp. 115-123.

[7] R. C. Holte. "Very Simple Classification Rules Perform well on Most Co mmonly used datasets." *Machine Learning*, 1993; 11, pp. 91-63.

[8] J. R, Quinlan. *Book Review*, C :4.5 Programs for Machine Learning: Morgan Kaufmann Publishers,1994; pp. 235-240.

[9] Nanaumphai, P.; Thongkam, J. Intrusion Detection Using Classification Techniques in Data Mining. Faculty of Informatics at Mahasarakham University *Journal of Information Technology and Innovation Management*,2019; Year 6 No. 2 July - Dec. pp. 111-118.

[10] Promthong, N.; Oenchamrush, S.; Tangprasert, S. Comparative Efficiency of Classification with Data Mining Technique between K-Nearest Neighbor Rule-Based and Decision Tree. *The Eighth National Conference on Computing and Information Technology NCCIT* 2012; pp. 540-546.

[11] Kittisak, Summary. Basic Health Screening by Using Data Mining Techniques. *Master of Science in Web Engineering, Faculty of Information Technology, Dhurakij Pundit University*, 2012.

[12] Fadi, Thabtah; Omar, Gharaibeh; Rashid, Al-Zubaidy. Arabic Text Mining Using Rule-Based Classi cation. *Journal of Information & Knowledge Management*, 2012; *11*(1), pp. 1-10.

[13] GM, Shafiullah; A B M S, Ali; A, Thompson. Rule-Based Classification Approach for Railway Wagon Health Monitoring. *Proceedings of the Neural Networks 2010 International Joint Conference, IEEE, Piscataway, N. J*. February 2010; pp 1-7.

[14] Rajalakshmi, A.; Vinodhini, R.; Fathima, Bibi K. Data Discretization Technique Using WEKA Tool A. Rajalakshmi et al *IJCSET*, August 2016; *6*(8); pp. 293-298.

[15] Wichareung, S. Applying Data Mining Technique in Loan Approval Process. Master of Science in Computer and Communication Technology, Faculty of Information Technology, Dhurakij Pundit University, 2010.

[16] Ian H. Witten; Eibe Frank; Mark A. Hall; Christopher J. Pal. 2nd ed.; *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers, Inc., Sanfrancisco, CA, USA ,2016; pp. 61-76 and pp. 97-119.

[17] Pacharawongsak, E. 2nd ed.; *An Introduction to Data Mining Techniques*. Bangkok: Asia Digital Printing Company, 2014; pp. 50-83.