International Journal of Innovative Computing, Information and Control
Volume **7**, Number **8**, August **2011**
ICIC International ©2011 ISSN 1349-4198
pp. 4885–4897

A HYBRID APPROACH OF DEA, ROUGH SET THEORY AND RANDOM FORESTS FOR CREDIT RATING

Der-Jang Chi[1], Ching-Chiang Yeh[2,*] and Ming-Cheng Lai[3]

[1]Department of Accounting
Chinese Culture University
No. 55, Hwa-Kang Road, Yang-Ming-Shan, Taipei City 11114, Taiwan
dichi@ms43.hinet.net

[2]Department of Business Administration
[3]Graduate Institute of Business Administration
National Taipei College of Business
No. 321, Sec. 1, Jinan Road, Zhongzheng District, Taipei City 100, Taiwan
*Corresponding author: ychinc@webmail.ntcb.edu.tw; laimc@mail.ntcb.edu.tw

Received April 2010; revised September 2010

Abstract. *In recent years, credit rating analysis has attracted lots of research interest in the literature. While the operating efficiency of a corporation is generally acknowledged to be a key contributor to the corporation's risk, it is usually excluded from early prediction models. To verify the operating efficiency as predictive variables, we propose a novel model to integrate rough set theory (RST) with the random forests (RF) technique, in order to increase credit rating prediction accuracy. In our proposed method, data envelopment analysis (DEA) is employed as a tool to evaluate the operating efficiency. Furthermore, the RST approach is used for variable selection due to its reliability in obtaining the significant independent variables, and utilized as a preprocessor to improve credit rating prediction capability by RF. The effectiveness of this methodology is verified by experiments comparing the RF, and compares the accuracy of the same prediction method with and without the DEA variable. The results show that operating efficiency does provide valuable information in credit rating predictions and the proposed approach provides better classification results.*

Keywords: Credit rating, Rough set theory, Random forests, Data envelopment analysis

1. Introduction.

Credit ratings have been extensively used by bond investors, debt issuers, and governmental officials as an estimate of credit condition or the ability to pay debt. They are important determinants of risk premiums and even the marketability of bonds [1]. Numerous useful techniques for corporate credit rating prediction have been the subject of research in the academic and business community. However, whilst these are well-established credit rating prediction techniques, two main problems arise.

Firstly, early studies considered only financial factors as independent (input) variables [2-4]. Although financial factors, originally found in corporation's financial statements, can reflect some characteristics of a corporation from various aspects, the operating inefficiency of a corporation is also acknowledged to be a key contributor to a corporation's operation risk [5-8], and however, it is usually excluded from early prediction models. In this study, we believe operating efficiency, which reflects the status of the management of a corporation in credit ratings prediction, is a decisive factor affecting prediction accuracy. It is difficult to evaluate the efficiency of a corporation directly from its financial statements. An approach known as data envelopment analysis (DEA) may offer useful insights into the data by incorporating multiple inputs and outputs, and DEA is able to measure the efficiency of a corporation.

Secondly, most of these statistical techniques for constructing credit rating predictions have been being used for guide some times. These techniques include multiple regression analysis [9-11], multiple discriminant analysis [12,13], ordered linear probit model [14-17] and ordered and unordered linear logit models [15], etc. However, these conventional statistical methods have some restrictive assumptions such as the linearity, normality and independence among predictor or input variables. Considering that the violation of these assumptions for independent variables frequently occurs with financial data [18], the methods can have limitations in terms of effectiveness and validity.

Recently, artificially intelligent approaches have been proved less vulnerable to these assumptions, such as random forests (RF), a tree-based classification and regression method, developed by the late Leo Breiman and Adele Cutler [19], which, compared with other current artificially intelligent approaches, is unsurpassable in accuracy. RF has been used extensively in different applications, such as modeling [20], prediction [21,22], pattern analysis in multimedia information retrieval, intrusion detection system [23,24] and machine fault diagnosis [25]. Unfortunately, to the best of our knowledge, RF has not yet been applied in the prediction problem of credit ratings. Moreover, there are several arguments suggesting that variable selection, also called feature selection, is a fundamental problem that has significant impact on the prediction accuracy of the models. Many methods have been developed to best prepare for data inputs, such as rough set theory (RST), developed by Pawlak [26]. Consistency of data [26,27], dependency of attributes [28], mutual information [29], discernibility matrix [30] and genetic algorithm are employed to find reducts of an information system [31]. In addition, these techniques are applied to text classification [32], face recognition [33], texture analysis [34], process monitoring [35], characters recognition [36], city development [37], marketing [38] and combined evolutionarily approach for optimized partitions [39]. An extensive review about RST-based feature selection is given in Thangavel and Pethalakshmi [40].

The aim of this paper is to investigate the role of non-financial factors in credit rating evaluation. In particular, we examine whether an assessment of a bank's operating efficiency conveys any useful ex-ante information in credit rating prediction. Firstly, we use DEA to obtain a measure of a corporation's operating efficiency as a predictive variable. Secondly, we propose a novel model to integrate the RST with RF techniques (RST+RF), to increase credit rating prediction accuracy. The RST approach is used for variable selection due to its reliability in obtaining the significant independent variables. Finally, the study will use acquired variables from RST as the inputs for RF models. The obtained results can then be compared to see whether the technique including efficiency variable will provide better classification accuracy or not. The effectiveness of the methodology was verified by experiments comparing the RF.

The remainder of this paper is structured as follows: Section 2 describes the methods used in the paper: DEA, RST and RF, respectively; in Section 3, we outline the proposed approach and experiment framework used in this research; Section 4 presents the experimental results of the proposed method; finally, the conclusions and suggestions are contained in Section 5.

2. **Methods.**

2.1. **Data envelopment analysis.**

Data envelopment analysis (DEA) is an evaluation tool for decision-making units (DMUs) and solves many decision-making problems by integrating multiple inputs and outputs simultaneously. DEA is a non-parametric data analysis technique that is extensively used by various research communities [41-43]. The basic ideas behind DEA dates back to Farrel [44], and however, the recent series of discussions started with the article by Charnes et al. [45]. We will give very briefly outline the salient features of DEA. Information that is more detailed can be obtained elsewhere [46,47].

The DEA ration form, proposed by Charnes, Cooper and Rhodes (CCR) [45] is designed to measure the relative efficiency or productivity of a specific DMU. The variable constant returns to scale of the DEA model refers to CCR. The CCR model is a fractional programming, which can be transformed to linear programming (LP) as follows:

$$\begin{aligned} & Min\ \theta \\ & s.t. \ -y_k + Y\lambda \geq 0 \\ & \qquad \theta x_k - X\lambda \geq 0 \\ & \qquad \lambda \geq 0 \end{aligned} \tag{1}$$

where X is the $K \times N$ input matrix and Y is the $M \times N$ output matrix, respectively. For the k^{th} firm these are represented by the vectors x_k and y_k, respectively. λ is a $N \times 1$ vector of constraints and θ is a scalar, $0 < \theta \leq 1$ which stands for efficiency of the k^{th} firm. Solve this LP for each firm and you will then obtain the efficiency score θ for each firm. When $\theta = 1$, a technical efficiency DMU is one boundary, compared with those inefficiency peers with $\theta < 1$.

Note that Equation (1) is an input orientation DEA model under the assumption of constant returns to scale (CRS) technology. Banker, Charnes and Cooper (BCC) [46] relaxed the constraint of CRS to account for variable returns to scale (VRS) technology by adding convexity constraint to Equation (1). The BCC model is:

$$\begin{aligned} & Min\ \theta \\ & s.t. \ -y_k + Y\lambda \geq 0 \\ & \qquad \theta x_k - X\lambda \geq 0 \\ & \qquad \sum \lambda = 1 \\ & \qquad \lambda \geq 0 \end{aligned} \tag{2}$$

where the variables of X, Y, x_k, y_k, λ and θ of Equation (2) are defined as the same ones as Equation (1).

2.2. **Basic concepts of rough set theory.** Rough set theory (RST) is a machine learning method, which is introduced by Pawlak [48] in the early 1980s. It has proven to be a powerful tool for uncertainty and is usually applied to data reduction, rule extraction, data mining and granularity computation. Here, we illustrate only the basic ideas of RST that are relevant to contemporary work.

Let $I = (U, A)$ be an information system, where U is the universe, a non-empty finite set of objects. A is a non-empty finite set of attributes. For $\forall a \in A$ determines a function $f_a : U \to V_a$. If $P \subseteq A$, there is an associated equivalence relation:

$$IND(P) = \{(x, y) \in U \times U : f(x, a) = f(y, a), \forall_a \in P\} \tag{3}$$

The partition of U, generated by $IND(P)$ is denoted U/P. If $f(x, y) \in IND(P)$, then x and y are indiscernible by attributes from P. The equivalence classes of the P-indiscernibility relation are denoted $[x]_p$. The indiscernibility relation is the mathematical basis of rough set theory.

Let P, $Q \subset A$ be equivalence relations over U, then the positive, negative and boundary regions can be defined as:

$$\begin{aligned} &POS_P(Q) = \bigcup_{X\in U/Q} P_*(x) \\ &NEG_P(Q) = U - \bigcup_{X\in U/Q} P^*(x) \\ &BND_P(Q) = \bigcup_{X\in U/Q} P^*(x) - \bigcup_{X\in U/Q} P_*(x) \end{aligned} \tag{4}$$

The positive region of the partition U/Q with respect to P, $POS_P(Q)$ is the set of all objects of U that can be certainly classified to blocks of the partition U/Q by means of P. A set is rough (imprecise) if it has a non-empty boundary region.

The positive region of the partition U/Q with respect to P, $POS_P(Q)$ is the set of all objects of U that can be certainly classified to blocks of the partition U/Q by means of P. A set is rough (imprecise) if it has a non-empty boundary region.

An important issue in data analysis is discovering dependencies between attributes. Dependency can be defined in the following way. For P, $Q \subseteq A$, P depends totally on Q, if and only if $IND(P) \subseteq IND(Q)$. That means that the partition generated by P is finer than the partition generated by Q. We say that Q depends on P in a degree k $(0 \leq k \leq 1)$, denoted $P \Rightarrow_k Q$, if

$$k = \gamma_P(Q) = |POS_P(Q)| \,/\, |U| \tag{5}$$

If $k = 1$, Q depends totally on P, if $(0 < k < 1)$, Q depends partially on P, and if $k = 0$ then Q does not depend on P. In other words, Q depends totally (partially) on P, if all (some) objects of the universe U can be certainly classified to blocks of the partition U/Q, employing P.

In a decision system the attribute set contains the condition attribute set C and decision attribute set D, i.e., $A = C \cup D$. The degree of dependency between condition and decision attributes, $\gamma_C(D)$, is called the quality of approximation of classification, induced by the set of decision attributes [49].

The goal of attribute reduction is to remove redundant attributes so that the reduced set provides the same quality of classification as the original. A reduct is defined as a subset R of the conditional attribute set C such that $\gamma_R(D) = \gamma_C(D)$. Any given decision table may have many attribute reducts. The set of all reducts is defined as:

$$\mathrm{Re}\,d = \{R \subseteq C |\, \gamma_R(D) = \gamma_C(D) \forall B \subset R, \gamma_B(D) \neq \gamma_C(D)\} \tag{6}$$

In rough set attribute reduction, a reduct with minimal cardinality is searched for. An attempt is made to locate a single element of the minimal reduct set $\mathrm{Re}\,d_{\min} \subset \mathrm{Re}d$:

$$\mathrm{Re}\,d_{\min} = \{R \in \mathrm{Re}\,d |\, \vee\, R' \in \mathrm{Re}\,d, |R| \leq |R'|\} \tag{7}$$

The intersection of all reducts is called the core, the elements of which are those attributes that cannot be eliminated. The core is defined as:

$$Core(C) = \cap\, \mathrm{Re}\,d \tag{8}$$

2.3. **Random forests.**

RF is another advanced method of machine learning. The classification is achieved by constructing an ensemble of randomized classification and regression tress (CART) [50]. For a given training dataset, $A = \{(X_1, y_1), (X_2, y_2), \ldots, (X_n, y_n)\}$, where $X_i = 1, 2, \ldots, n$, is a variable or vector and y_i is its corresponding property or class label; the basic RF algorithm is presented as follows:

2.3.1. *Bootstrap sample.*

Each training set is drawn with replacement from the original dataset A. Bootstrapping allows replacement, so that some of the samples will be repeated in the sample, while others will be "left out" of the sample. The "left out" samples constitute the "Out-of bag (OOB)" which has, for example, one-third, of samples in A which are used later to get a running unbiased estimate of the classification error as trees are added to the forest and variable importance.

2.3.2. *Growing trees.* For each bootstrap sample, a tree is grown m variables (m_{try}) are selected at random from all n variables ($m_{try} \leq n$) and the best split of all m_{try} is used at each node. Each tree is grown to the largest extent (until no further splitting is possible) and no pruning of the trees occurs.

2.3.3. *OOB error estimate.* Each tree is constructed on the bootstrap sample. The OOB samples are not used and therefore regarded as a test set to provide an unbiased estimate of the prediction accuracy. Each OOB sample is put down the constructed trees to get a classification. A test set classification is formed. At the end of the run, take k to be the class which got most of the "votes" every time sample n was OOB. The proportion of times that k is not the true class of n averaged over all samples is the OOB error estimate.

RF has several advantages over other statistical modeling methods [50]. Its variables can be both continuous and categorical. As a large number of trees are induced and averaged during the run, RF can produce the low bias and low variation results but highly accurate classification and good prediction. Since RF can make OOB error estimates which test the classifications by vote on a small number of samples this further strengthens the model. This somewhat counterintuitive strategy turns out to perform very well compared to many other classifiers, including discriminant analysis, support vector machines and neural networks, and is robust against over fitting [51]. In addition, RF has only two hyperparameters (the number of variables in the random subset at each node and the number of trees in the forest), and is usually not very sensitive to their values.

2.4. **Discussion.** Regarding the above review, some issues as the limitations of literature, are listed below.

1. Most of the prior study only adopted financial ratios as independent variables. While the operating inefficiency of a corporation is also acknowledged to be a key contributor to a corporation's operation risk [5-8], and however, it is usually excluded from early prediction models.
2. RF has been used extensively in different applications, such as modeling [20], prediction [21,22], pattern analysis in multimedia information retrieval, intrusion detection system [23,24] and machine fault diagnosis [25]. Unfortunately, to the best of our knowledge, RF has not yet been applied in the prediction problem of credit ratings.
3. Kumar and Ravi [52] provide a detailed review of artificially intelligent techniques for financial related problems and one important trend is to build a hybrid intelligent system. However, there are very few studies focusing on developing hybrid models for credit rating [53].

As a result, we examine whether an assessment of a corporation's operating efficiency conveys any useful ex-ante information in credit rating prediction and propose a novel hybrid model to integrate the RST with RF techniques (RST+RF), to increase credit rating prediction accuracy.

3. **Proposed Approach.** In this study, we would like to use operating efficiency as a predictive variable and propose the novel model, RST+RF, to increase the accuracy of credit rating prediction. To test if operating efficiency would be helpful in credit ratings predictions, we look at the operating efficiency before and after the credit rating is taken into consideration. As we would also like to see whether RST can be a good supporting tool in deciding the input variables of the RF prediction model, we also explore the performance of credit ratings predictions using the proposed RST+RF model.

Firstly, this study employs DEA are employed as a tool to evaluate the operating efficiency as a predictive variable, and the banking industry is fit for variable return to scale [8], hence, we applied the BCC model. Secondly, RST does variable selection due to its reliability in obtaining the significant independent variables. Next, we will use the obtained significant independent variables from RST as inputs for RF models. The obtained results can then be compared to see whether the one including DEA will give a better classification accuracy or not. Finally, in order to verify the applicability of this methodology, we also use the RF model as a benchmark.

4. **Research Data and Experiments.**

4.1. **Data set.**

We collected data from the Taiwan Ratings Corporation (TRC) and the Securities and Futures Institute (SFI). Information about issuer credit ratings was obtained from the TRC, which is one of the largest credit rating organizations in Taiwan, as well as being the first credit rating service established in this country. The data pertaining to financial, operational, equity structure and market information was obtained from the SFI, which plays a key role as a data center for Taiwan's securities and futures markets.

A TRC rating indicates an issuer's capacity to meet its financial commitments over a one-year period or longer. The ratings are twAAA, twAA, twA, twBBB, twBB, twB, twCCC, twCC and twD. The prefix 'tw' denotes Taiwan and the rating scale focuses on Taiwan's financial markets. The twAAA rating indicates that an organization has an extremely strong capacity to meet its commitments, whereas the twD rating denotes an organization that may be in high risk. However, the TRC rating scale does not address sovereign risk, so it is not directly comparable to standard and Poor's global scale. So far, in Taiwan, the range of ratings information for banks determined by the TRC is from twBB to twAAA. As the number of samples in extreme-rating classes is too small, we combine both twAAA and twAA as >= twAA, while twBB and lower ratings are merged as =<twBB. In this classification model, we have designed three rating categories which appear in our data set, namely: twAA and above (=<twAA); twA; and twBB and below (=<twBB).

After matching and filtering the data with missing values, we obtained a data set of 123 cases with bank credit rating, which covered 24 commercial banks from 2003 to 2007. We randomly partition the data set into two parts in a proportion of 4:1 as shown in Table 1. The first part is used for training and validation to select optimal parameters for the RF model, and the second part is used for testing.

Table 1. Description of the sample

Ratings	Numbers	Percentage	Number of samples used in	
			Training set	Testing set
twAA and the higher ratings	42	34.2%	34	8
twA	40	32.5%	32	8
twBBB and the lower rating	41	33.3%	32	9
Sum	123	100.0%	98	25

4.2. **Variable selection.**

In order to apply prediction methods to corporate credit rating, which is generally based on analyzing the financial ratios, predictor variables should be selected at first. In this paper, the predictor variables include two parts: one is the operating efficiency of corporation, which has been introduced in next section, and the other consists of several variables selected from financial ratios.

There are many financial ratios that could be derived from the financial statements. There are 22 variables (including 21 financial ratios and DEA) are listed in Table 2. These variables include the financial ratios that were available in the TRC database and are frequently used in credit rating prediction literature.

Table 2. Potential predictor variables

Definition	Frequencies of occurrence in reducts generated by RST
Investment growth ratio	6
Lending growth ratio	5
Stockholders' equity	5
Operating profit margin	4
DEA	4
Non-performing loans ratio	3
Liquidity ratio	2
Capital ratio	2
Receivables to payables	1
Earnings per share	1
Gross profit margin	1
Return on total assets	1
Interest expense to sales	1
Interest coverage ratio	1
Stock return	1
Total liabilities	1
Total assets	1
Debit ratio	1
Current ratio	1
Net income to stakeholders' equity	1
Return on equity	1
Cash flow from operating activities/current liabilities	1

For the DEA, informative input and output variables should be selected. Like previous researchers [54], input variables chosen in this study were interest expenses, fixed assets, deposits and number of employees; output variables chosen were interest income, non-interest income, investments and loans. The efficiency results of the DEA are listed in Table 3.

Table 3. Description of the efficiency

Efficiency (%)	Numbers
20=<	0
20-40	1
41-60	21
61-80	36
81-99	22
100	43

In order to find the relative importance of the independent variables, the frequencies of occurrence of the independent variables in the reducts generated are computed. The RST-based application RSES (a collection of algorithms and data structures for rough set computations, developed at the Group of Logic, Inst. of Mathematics, University of Warsaw, Poland).

Firstly, the continuous variables are discretized as the RST requires a processed dataset consisting of discrete features. In this study, the global method [55] is used to discretize the continuous variables. After constructing the discretized information table, RST was expected to find the reducts based on indiscernibility relation to generate decision rules and the genetic reduction algorithm [56] were used to obtain optimal reducts. Table 2 gives the variables' frequencies of occurrence in reducts generated.

As described in Table 2, independent variables with the highest frequency of occurrence in the reducts generated is investment growth ratio, followed by lending growth ratio, stockholders' equity, operating profit margin, DEA, non-performing loans ratio, liquidity ratio, and capital ratio. The remaining independent variables below 2 reducts and are eliminated in the subsequent experiment. Therefore, the reduced 8 variables (including 7 financial ratios and DEA) are selected as potential predictor variables. The selected variables are taken as input of the classifier of RF.

4.3. **Results and analysis.** After the significant independent variables are picked out, RF classifiers are implemented. We employed the RF available in the R package random forest [57]. This implementation is based on the original Fortran code authored by Leo Breiman, the inventor of RF. Following the suggestions of [19,57] and http://www.stat.berkeley.edu/breiman/RandomForests/, we consider different parameter configurations for the values of $m_{try} = 2 \times \sqrt{number\ of\ variables}$ (number of variables randomly selected for each tree), $n_{tree} = \{250, 500, 1000, 1500, 2000\}$ (number of trees to build).

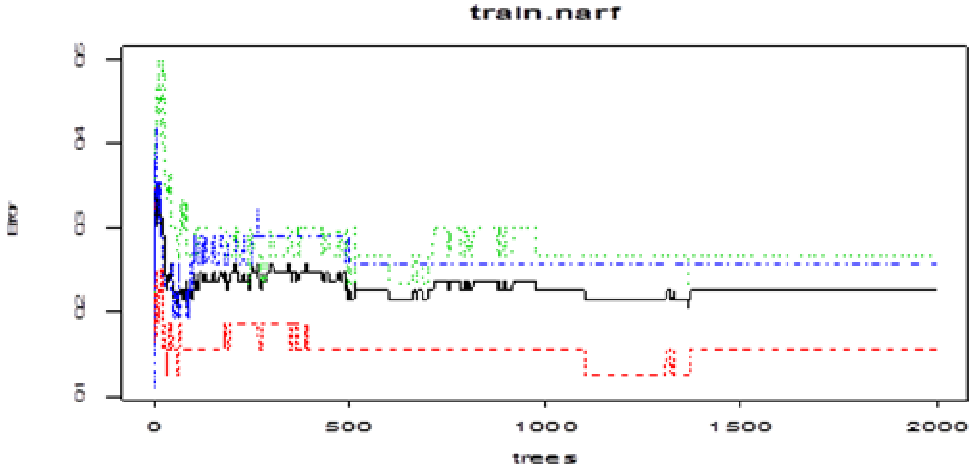

Figure 1. OOB error for the training data of random forest

Figure 1 shows the graph of the OOB accuracy estimation (in terms of relative error rates) with increase in number of classification trees. OOB estimated error rates are computed continuously classification tree are built. Estimates based on the model are hypothesized to be unbiased estimates of error, and however, estimates based on a small number of trees are likely unreliable. As Figure 1 shows, black lines represent overall relative error rates, green lines represent "twA" class relative error rates, blue lines represent "twBBB>=" class relative error rates, and red lines represent "twAA>=" class relative error rates. It illustrates the convergence of the RF algorithm: this depicts the running relative rate of error for the RF model. We can see that when the number of trees about is 1500, the relative error rate is almost 0.021333 after that it starts decreasing and ends with an error rate of 0.02777, which indicates that the obtained RF has reached its optimal performance.

For testing whether operating efficiency will be helpful in credit ratings prediction, we tested these two possible hybrid models. The RST+RF model including both financial ratios and DEA is model 1, and the RST+RF model using only financial ratios as independent variables is model 2.

To evaluate the prediction performance, we followed the 5-fold cross-validation procedure, which has shown good performance in model selection [58]. Hence, every subset will be trained and tested 5 times, and the average prediction performance can be obtained consequently. The results of the confusion matrix for the testing sample using these two models can be summarized in Tables 4 and 5 respectively. Table 4 shows the confusion matrix of a full-variable model, RST+RF, while Table 5 shows the confusion matrix consisting only the financial ratios variables model. We can observe that the average correct classification rate is 88.00% for the model 1, and 84.00% for the model 2. From the improved correct classification rate of the model using both financial ratios and DEA, we can conclude that operating performance is helpful in improving the classification accuracy of the prediction model.

Table 4. Model 1 classification results

Actual rating	>= twAA	twA	=<twBBB
>=twAA	7	1	0
twA	0	7	1
=<twBBB	0	1	8
Average correct classification rate: 88.00%			

Table 5. Model 2 classification results

Actual rating	>= twAA	twA	=<twBBB
>=twAA	7	0	1
twA	1	7	0
=<twBBB	1	1	7
Average correct classification rate: 84.00%			

In this study, we also tested these two possible hybrid models. The RF model including both financial ratios and DEA is model 3, and the RF model using only financial ratios as independent variables is model 4. The prediction results of the confusion matrix using the two prediction models are summarized in Tables 6 and 7, respectively.

Table 6. Model 3 classification results

Actual rating	>= twAA	twA	=<twBBB
>=twAA	6	1	1
twA	0	7	1
=<twBBB	1	1	7
Average correct classification rate: 80.00%			

Table 7. Model 4 classification results

Actual rating	>= twAA	twA	=<twBBB
>=twAA	6	1	1
twA	1	7	0
=<twBBB	1	2	6
Average correct classification rate: 76.00%			

From the results revealed in Tables 6 and 7, we can observe that the average correct classification rate is 80.00% for model 3, and 76.00% for model 4. Again, from Table 8, the improved correct classification rate of the model incorporating both financial ratios and DEA demonstrates that operating performance does provide valuable information in credit rating prediction model.

Table 8. Predictive accuracies of the constructed model

Model	Average accuracy	
	Training set	Testing set
RST+RF model classification results with both financial ratios and DEA (Model 1)	91.84%	88.00%
RST+RF model classification results with only financial ratios (Model 2)	85.71%	84.00%
RF model classification results with both financial ratios and DEA (Model 3)	81.63%	80.00%
RF model classification results with only financial ratios (Model 4)	79.60%	76.00%

By examining the input variables and accuracies of the models, we can provide useful information about the credit rating process. Firstly, the models including both financial ratios and operating performance provide better classification results than the models only using financial ratios. Furthermore, the importance rankings in RST can understand as being relative to a particular variable. Overall, it shows that the operating performance, are more important than the financial ratios in effecting the credit rating.

These findings have obvious managerial implications since the operating performance can influence credit rating for commercial banks in Taiwan. For some financial ratios (e.g., receivables to payables and earnings per share), these results show that it may have little effect on credit rating and should therefore not be a priority item for managerial action. Conversely, since the credit rating is more strongly influenced by operating performance. The above phenomenon implies that operating performance does provide valuable information in predicting credit rating.

Secondly, we proposed RST+RF model provides better classification results than the RF model, even when only considering financial ratios or the model including both financial ratios and operating performance. Hence, we believe that the proposed RST+RF model is a better alternative since it exhibits the capability of identifying important independent variables, which may provide valuable information for further credit rating purposes.

5. **Conclusions and Suggestions.** In this study, we use operating efficiency as the predictive variable and propose a novel model, RST+RF, in order to increase the accuracy of credit rating prediction. Thus, we use DEA as the variable for efficiency of a corporation

in the credit rating prediction models. To verify the applicability of this methodology, we also designed a RF model as the benchmark, and applied it to the commercial banks in Taiwan.

We have then applied this proposed method to the credit rating prediction of an empirical examination of credit rating and showed that the proposed model can be used to improve the predictions to the in credit rating problem as well as understand the relative importance and order of the input variables. The results of this study can be summarized as follows.

Firstly, the non-financial efficiency indicators are useful ex-ante determinants of credit rating. The new input variables, i.e., DEA as the variable for efficiency of a corporation, are intended to enhance the classification effectiveness of credit ratings. We have done this by emphasizing the links between the corporation's performance and financial decisions. In particular, we have shown that managerial inefficiencies are an important ex-ante indicator of the corporation's credit rating. Secondly, the proposed RST+RF model provides better classification results than RF, both when only financial ratios are considered or the model includes both financial ratios and DEA. Hence, the RST+RF model appears to be an efficient alternative. The above-mentioned research findings justify the presumption that the RST+RF model should be a better choice in conducting credit ratings prediction tasks. Thus, the forecasting technique (RST+RF) can provide an accurate investment guide for investors and governments.

However, several problems are worthy of further research. Firstly, we used operating efficiency as the predictive variable in credit ratings prediction. Future research may use potential non-financial factors as part of the bank's overall credit rating. Secondly, before RF is implemented, the (m_{try}, n_{tree}) parameters have to be optimized in order to construct a first-class classifier. Consequently, the extraction of optimal parameters is crucial when implementing RF, and must be taken into account in the further. Finally, we should continue to compare our proposed feature selection approach with state-of-the-art approaches in the future.

REFERENCES

[1] Z. Huang, H. Chen, C. J. Hsu, W. H. Chen and S. Wu, Credit rating analysis with support vector machine and neural networks: A market comparative study, *Decision Support Systems*, vol.37, pp.543-558, 2004.

[2] E. I. Altman, Financial ratios, discriminant analysis, and the prediction of corporate bankruptcy, *Journal of Finance*, vol.23, pp.589-609, 1968.

[3] W. P. H. Poon, Are unsolicited credit ratings biased downward? *Journal of Banking and Finance*, vol.27, pp.593-614, 2003.

[4] J. Pettit, C. Fitt, S. Orlov and A. Kalsekar, *The New World of Credit Ratings*, http://ssrn.com/abstract=593522, 2004.

[5] H. Secrist, *National Bank Failures and Non-Failures: An Autopsy and Diagnosis*, Principia Press, Bloomington, 1938.

[6] L. D. Seballos and J. B. Thomson, Understanding causes of commercial bank failures in the 1980s, *Economic Commentary*, 1990.

[7] T. V. Gestel, B. Baesens, J. Suykens, D. V. Poel, D. E. Baestaens and M. Willekens, Bayesian kernel based classification for financial distress detection, *European Journal of Operational Research*, vol.172, no.3, pp.979-1003, 2006.

[8] Y. H. Chiu, C. M. Ma and M. Y. Sun, Efficiency and credit rating in Taiwan banking: DEA estimate, *Applied Economics (SSCI)*, 2009.

[9] J. O. Horrigan, The determination of long-term credit standing with financial ratios, *Journal of Accounting Research*, vol.4, pp.44-62, 1966.

[10] T. F. Pogue and R. M. Soldofsky, What's in a bond rating? *Journal of Financial and Quantitative Analysis*, vol.4, pp.201-228, 1969.

[11] R. R. West, An alternative approach to predicting corporate bond ratings, *Journal of Accounting Research*, vol.8, pp.118-125, 1970.

[12] G. E. Pinches and K. A. Mingo, A note on the role of subordination in determining industrial bond ratings, *Journal of Finance*, vol.30, pp.201-206, 1975.

[13] E. I. Altman and S. Katz, Statistical bond rating classification using financial and accounting data, *Proc. of the Conf. on Topical Research in Accounting*, New York, pp.205-239, 1976.

[14] R. S. Kaplan and G. Urwitz, Statistical models of bond ratings: A methodological inquiry, *Journal of Business*, vol.52, pp.231-261, 1979.

[15] L. H. Ederington, Classification models and bond ratings, *The Financial Review*, vol.20, pp.237-262, 1985.

[16] J. A. Gentry, D. T. Whitford and P. Newbold, Predicting industrial bond ratings with a probit model and fund flow components, *The Financial Review*, vol.23, pp.269-286, 1988.

[17] R. C. Hwang, K. F. Cheng and C. F. Lee, On multiple-class prediction of issuer credit ratings, *Applied Stochastic Models in Business and Industry*, vol.25, no.5, pp.535-550, 2008.

[18] E. B. Deakin, A discriminant analysis of predictors of business failure, *Journal of Accounting Research*, vol.10, pp.167-179, 1972.

[19] L. Breiman, *Manual on Setting up, Using, and Understanding Random Forests, v4.0*, ftp://ftp.stat.berkeley.edu/pub/users/breiman/, 2003.

[20] P. Xu and F. Jelinek, Random forests and the data sparseness problem in language modeling, *Journal of Computer Speech and Language*, vol.21, no.1, pp.105-152, 2007.

[21] L. Guo, Y. Ma, B. Cukic and H. Singh, Robust prediction of fault-proneness by random forests, *Proc. of the 15th Int. Symposium on Software Reliability Engineering*, Brittany, France, pp.417-428, 2004.

[22] B. Lariviere and D. van den Poel, Predicting customer retention and profitability by using random forests and regression forests techniques, *Expert Systems with Applications*, vol.29, no.2, pp.472-482, 2005.

[23] S. K. Dong, M. L. Sang and S. P. Jong, Building lightweight intrusion detection system based on random forest, *Lecture Notes in Computer Science*, vol.3973, pp.224-230, 2006.

[24] J. Zhang and M. Zulkernine, A hybrid network intrusion detection technique using random forests, *Proc. of IEEE the 1st Int. Conf. on Availability, Reliability and Security*, Kingston, Canda, pp.1-8, 2006.

[25] B. S. Yang, X. Di and T. Han, Random forests classifier for machine fault diagnosis, *Journal of Mechanical Science and Technology*, vol.22, pp.1716-1725, 2008.

[26] Z. Pawlak, *Rough Sets: Theoretical Aspects of Reasoning about Data*, Springer-Verlag, New York, 1991.

[27] J. S. Mi, W. Z. Wu and W. X. Zhang, Approaches to knowledge reduction based on variable precision rough set model, *Information Sciences*, vol.159, no.3-4, pp.255-272, 2004.

[28] G. Wang, H. Hu and D. Yang, Decision table reduction based on conditional information entropy, *Chinese Journal of Computers*, vol.25, no.7, pp.1-8, 2002.

[29] A. Skowron and C. Rauszer, The discernibility matrices and functions in information systems, *Intelligent Decision Support: Handbook of Applications and Advances of Rough Set Theory*, pp.331-362, 1992.

[30] J. Wang and D. Q. Miao, Analysis on attribute reduction strategies of rough set, *Journal of Computer Science and Technology*, vol.13, no.2, pp.189-193, 1998.

[31] H. Moradi, J. W. Grzymala-Busse and J. A. Roberts, Entropy of English text: Experiments with humans and a machine learning system based on rough sets, *Information Sciences*, vol.104, no.1-2, pp.31-47, 1998.

[32] R. W. Swiniarski and L. Hargis, Rough sets as a front end of neural networks texture classifier, *Neurocomputing*, vol.36, no.1-4, pp.85-102, 2001.

[33] H. Liu and R. Setiono, Some issues on scalable feature selection, *Expert Systems with Applications*, vol.15, no.3-4, pp.333-339, 1998.

[34] R. W. Swiniarski and A. Skowron, Rough set methods in feature selection and recognition, *Pattern Recognition Letters*, vol.24, no.6, pp.833-849, 2003.

[35] D. Dubois and H. Prade, Putting fuzzy sets and rough sets together, *Intelligent Decision Support: Handbook of Applications and Advances of Rough Set Theory*, pp.203-232, 1992.

[36] W. Kasemsiri and Y. Shi, Thai characters recognition based on tolerant rough sets with fuzzy C-mean, *ICIC Express Letters*, vol.3, no.4(B), pp.1399-1404, 2009.

[37] C.-Y. Huang, G.-H. Tzeng, C.-C. Chan and H.-C. Wu, Semiconductor market fluctuation indicators and rule derivations using the rough set theory, *International Journal of Innovative Computing, Information and Control*, vol.5, no.6, pp.1485-1504, 2009.

[38] L.-C. Lin, J. Zhu, W. Junzo, K. Tomoko and I. Hiroaki, A rough set approach to classification and its application for the creative city development, *International Journal of Innovative Computing, Information and Control*, vol.5, no.12(B), pp.4859-4866, 2009.

[39] C. Bodie, M. Tshilidzi and L. Monica, Evolutionarily optimized rough sets partitions, *ICIC Express Letters*, vol.3, no.3(A), pp.241-246, 2009.

[40] K. Thangavel and A. Pethalakshmi, Dimensionality reduction based on rough set theory: A review, *Applied Soft Computing*, vol.9, no.1, pp.1-12, 2009.

[41] H. Hong, S. Ha, C. Shin, S. Park and S. Kim, Evaluating the efficiency of system integration projects using data envelopment analysis (DEA) and machine learning, *Expert Systems with Applications*, vol.16, no.3, pp.283-296, 1999.

[42] H. Seol, J. Choi, G. Park and Y. Park, A framework for benchmarking service process using data envelopment analysis and decision tree, *Expert Systems with Applications*, vol.32, no.2, pp.432-440, 2007.

[43] S. Sohn and T. Moon, Decision tree based on data envelopment analysis for effective technology commercialization, *Expert Systems with Applications*, vol.26, no.2, pp.279-284, 2004.

[44] M. J. Farrell, The measurement of productive efficiency, *Journal of the Royal Statistical Society: Series A (General)*, vol.120, no.3, pp.253-290, 1957.

[45] A. Charnes, W. W. Cooper and E. Rhodes, Measuring the efficiency of decision making units, *European Journal of Operations Research*, vol.2, no.6, pp.429-444, 1978.

[46] R. D. Banker, A. Charnes and W. W. Cooper, Some models for estimating technical and scale inefficiencies in data envelopment analysis, *Management Science*, vol.9, pp.1078-1092, 1984.

[47] A. Charnes, W. W. Cooper, A. Y. Lewin and L. M. Seiford, *Data Envelopment Analysis: Theory, Methodology and Applications*, Springer-Verlag, New York, 1995.

[48] Z. Pawlak, Rough sets and intelligent data analysis, *Information Sciences*, vol.147, no.1-4, pp.1-12, 2002.

[49] R. E. Schapire, Y. Freund, P. Bartlett and W. S. Lee, Boosting the margin: A new explanation for the effectiveness of voting methods, *The Annals of Statistics*, vol.26, pp.1651-1686, 1998.

[50] L. Breiman, Random forests, *Machine Learning*, vol.45, pp.5-32, 2001.

[51] L. Breiman, Bagging predictors, *Machine Learning*, vol.24, no.2, pp.123-140, 1996.

[52] P. R. Kumar and V. Ravi, Bankruptcy prediction in banks and firms via statistical and intelligent techniques – A review, *European Journal of Operational Research*, vol.180, pp.1-28, 2007.

[53] C. F. Tsai and M. L. Chen, Credit rating by hybrid machine learning techniques, *Applied Soft Computing*, vol.10, pp.374-380, 2010.

[54] K. H. Lu, M. L. Yang, F. K. Hsiao and H. Y. Lin, Measuring the operating efficiency of domestic banks with DEA, *International Journal of Business Performance Management*, vol.9, no.1, pp.22-42, 2007.

[55] J. Bazan, H. S. Nguyen, S. H. Nguyen, P. Synak and J. Wróblewski, Rough set algorithms in classification problem, in *Rough Set Methods and Applications*, L. Polkowski, S. Tsumoto and T. Lin (eds.), New York, Physica-Verlag, 2000.

[56] K. Komorowski, A. Øhrn and A. Skowron, The ROSETT a rough set software system, in *Handbook of Data Mining and Knowledge Discovery*, W. Klösgen and J. Zytkow (eds.), Oxford University Press, 2000.

[57] A. Liaw and M. Wiener, Classification and regression by random forest, *R News*, vol.2/3, pp.18-22, 2002.

[58] S. M. Weiss and C. A. Kulikowski, *Computer Systems That Learn: Classification and Prediction Methods from Statistics, Neural Networks, Machine Learning and Expert Systems*, Morgan Kaufmann, San Mateo, 1991.