ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org

http://dx.doi.org/10.13088/jiis.2018.24.4.155

Predicting Corporate Bankruptcy using Simulated Annealing-based Random Forests*

Hoyeon Park
Department of MIS, Graduate School,
Dongguk University_Seoul
(*hoyeonni@naver.com*)

Kyoung-jae Kim
Department of MIS, Business School,
Dongguk University_Seoul
(*kjkim@dongguk.edu*)

Predicting a company's financial bankruptcy is traditionally one of the most crucial forecasting problems in business analytics. In previous studies, prediction models have been proposed by applying or combining statistical and machine learning-based techniques. In this paper, we propose a novel intelligent prediction model based on the simulated annealing which is one of the well-known optimization techniques. The simulated annealing is known to have comparable optimization performance to the genetic algorithms. Nevertheless, since there has been little research on the prediction and classification of business decision-making problems using the simulated annealing, it is meaningful to confirm the usefulness of the proposed model in business analytics. In this study, we use the combined model of simulated annealing and machine learning to select the input features of the bankruptcy prediction model. Typical types of combining optimization and machine learning techniques are feature selection, feature weighting, and instance selection. This study proposes a combining model for feature selection, which has been studied the most. In order to confirm the superiority of the proposed model in this study, we apply the real-world financial data of the Korean companies and analyze the results. The results show that the predictive accuracy of the proposed model is better than that of the naïve model. Notably, the performance is significantly improved as compared with the traditional decision tree, random forests, artificial neural network, SVM, and logistic regression analysis.

Key Words : Simulated Annealing, Random Forests, Bankruptcy Prediction, Feature Selection, Business Analytics

Received : November 30, 2018 Revised : December 11, 2018 Accepted : December 12, 2018
Publication Type : Regular Paper Corresponding Author : Kyoung-jae Kim

1. Introduction

A firm's financial bankruptcy prediction model is essential for decision making by entrepreneurs, creditors, and investors. Financial bankruptcy resulting from corporate credit risk can place socially significant economic costs, and in extreme cases can lead to economic downturns as well as corporate bankruptcies. Some examples include Korea's Benefit from IMF bailout caused by Asian financial collapse in 1997, dot-com bubble in the early 2000s due to the failure of the internet-based

* This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education (NRF-2017R1D1A1B03032770).

companies, and 2008 global financial crisis caused by the subprime mortgage crisis.

As a result, corporate bankruptcy prediction has long been considered an important research topic in the field of finance and accounting research. Most of the previous studies have used various statistical techniques or artificial intelligence techniques, among which LR (logistic regression), DA (discriminant analysis), DT (decision trees), CBR (case-based reasoning), and ANN (artificial neural network) are typical ones. Among them, ANN is the most widely used technique due to high prediction accuracy. However, ANN is a method that has critical limitations due to lack of explanatory power and overfitting problem, so it is difficult to use it actively in practice.

Based on this background, SVM (support vector machine), which is known to have more predictable performance than ANN, is widely used. Unlike ANN, SVMs use the principle of structural risk minimization, so it searches a decision boundary to minimize an upper bound of the generalization error. For that reason, overfitting may be unlikely to rise with SVMs. Therefore, several recent studies on corporate bankruptcy prediction have used standalone SVMs or hybrid techniques with SVMs as a novel classifier.

Despite the recent use of SVMs, they have the disadvantage that SVMs are sensitive to the setting of several design factors. For example, setting the kernel function and its parameters, and selecting the features to be used in the classifier are typical examples. Also, proper sampling is one of the crucial factors that determine the performance of SVM. Thus, in recent years, random forest (RF), which shows very high classification performance, is used for predicting corporate bankruptcy even though there is little need for such design factors.

In this paper, we propose a predictive model that combines machine learning and optimization techniques to improve the accuracy of the firm's financial bankruptcy. Specifically, it is a hybrid algorithm that combines simulated annealing and random forest. In order to confirm the usefulness of the proposed model, we compare the existing machine learning algorithms such as decision trees, artificial neural networks, SVM, logistic regression, and random forest-based bankruptcy prediction models.

The composition of this paper is as follows. Chapter 2 provides a brief description of the previous research and simulated annealing. In Chapter 3, we present a combined prediction model of the machine learning method and optimization technique. In this chapter, we also present the results of the experiment. Chapter 4 suggests future research including the strengths and weaknesses of the paper.

2. Prior Research

2.1 Prior Studies on Corporate Bankruptcy Prediction

Traditionally, many researchers have developed predictive models based on exploratory studies to derive important variables and use them. The first model of corporate bankruptcy prediction model was developed by Altman (1968) using the Z-score based on a traditional statistical technique. Starting with that research, many researchers have proposed

〈Table 1〉 Prior research on the prediction of corporate bankruptcies using business analytics

Reference	Suggested model	Comparative models
Tam and Kiang, 1992	BPN	DA, LR, k-NN, ID3
Martin-del-Brio and Serrano-Cinca, 1993	SOM	N/A
Serrano-Cinca, 1996	SOM	N/A
Serrano-Cinca, 1997	BPN	DA, LR
Altman et al., 1994	BPN	DA
Wilson and Sharda, 1994	BPN	DA
Boritz and Kennedy, 1995	BPN	DA, LR, Probit
Boritz et al., 1995	BPN	DA, k-NN, LR, Probit
Jo and Han, 1996	BPN	DA, k-NN
Lee et al., 1996	BPN	LR, DA
Jo et al., 1997	BPN	DA, CBR
Kiviluoto, 1998	SOM, RBF-SOM, LVQ	DA, k-NN
Zhang et al., 1999	BPN	LR
Shin and Lee, 2002	GA	N/A
Hong and Shin, 2003	GA	ANN
Kim, 2004	Hybrid GA & ANN	LR, CBR, BPN
Shin et al., 2005	SVM	BPN
Kim and Kim, 2007	Modified bagging	ANN, DT, Bagging
Ahn and Kim, 2009	Hybrid GA & CBR	CBR
Nanni and Lumini, 2009	Bagging	ANN, SVM, k-NN
Ok and Kim, 2009	GA	LR, DT, ANN, k-NN
Heo and Yang, 2014	Adaboost	ANN, SVM, DT, Z-score
Wang et al., 2014	FS-boosting	LR, NB, DT, BPN, SVM, Bagging, Boosting
Tsai et al., 2014	bagging, boosting	ANN, SVM, DT
Kim and Ahn, 2015	k-RNN+OCSVM	Simple Random Under-sample
Lopez and Sanz, 2015	Hybrid BPN & SOM	DA, LR, RF, BPN, SVM
Kwon et al., 2017	RNN	DA, GLM, SVM, ANN
Barboza et al., 2017	RF	SVM, Boosting, Bagging, BPN, LR, DA

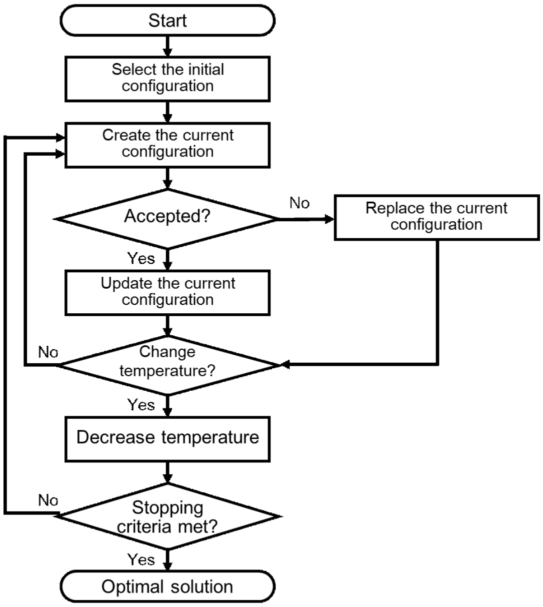
ANN: artificial neural networks, CBR: case-based reasoning, NB: naïve Bayes, BPN: Backpropagation neural networks, SOM: Self-organizing map, RF: random forest, RBF: Radial basis function, LVQ: Learning vector quantization, DA: Discriminant analysis, DT: decision tree, FS: feature selection, GLM: Generalized linear model, LR: Logistic regression, k-NN: k-nearest neighbor, k-RNN: k-Reverse nearest neighbor, OCSVM: One-class support vector machine, PNN: Probabilistic neural networks, GA: Genetic algorithm, SVM: Support vector machines, N/A: Not Applicable novel financial and credit risk assessment models. Ohlson (1980) proposed a logistic regression model for prediction of financial defaults, and Wilson and Sharda (1994) compared the predictive power of multivariate discriminant analysis with artificial neural networks. Besides, Jo and Han (1996) proposed a model that integrates discriminant analysis, artificial neural network and case-based reasoning for bankruptcy prediction. In the study by Shin et al. (2005), SVM was superior to BPN (back-propagation neural network) in corporate bankruptcy prediction problem.

Moreover, Nanni and Lumini (2009) proposed the ensemble method to calculate the credit score required for bankruptcy prediction. A recent study, Barboza et al. (2017), suggested an improved model combining the existing statistical models of Altman (1968) and Ohlson (1980) with machine learning models including boosting, bagging, and RF. <Table 1> shows typical studies of this research field.

As shown in <Table 1>, recent studies have been conducted to combine several models using the genetic algorithm, bagging, boosting, and so on. In general, it is possible to improve the prediction performance in the case of the bagging or boosting which combines the prediction results after learning each learning technique, but the degree of improvement may be limited because it is a method of combining the results of the predictors already learned. On the other hand, in the case of the hybrid model using genetic algorithms, it is a combination model of the wrapper method that optimizes the objective function of the genetic algorithm to maximize the prediction performance and optimizes the selection of the parameters of each model. Therefore, there is a possibility that the degree of improvement of prediction performance becomes more substantial than the former method of merely combining prediction results. In particular, it is known that genetic algorithms lead to improvement of prediction performance in the wrapper-based combination model through many previous studies. In this paper, we propose a hybrid model based on simulated annealing, which is one of the global optimization methods similar to genetic algorithms.

2.2 Simulated Annealing

This research suggests how to combine simulated annealing with machine learning for prediction of corporate bankruptcy, so it should be explained specifically in this context. Simulated annealing (SA) is an algorithm created by Kirkpatrick et al. (1983), which is also referred to as a "quenching technique". SA is known as a probabilistic technique for approximating the global optimum of a specified function. In especial, it is a metaheuristic to approximate global optimization in a large search space. For problems where finding an approximate global optimum is more important than finding a precise local optimum in a fixed amount of time, simulated annealing may be preferable to alternatives such as gradient descent. <Figure 1> shows the basic process of the SA.


Start
Select the initial configuration
Create the current configuration
Accepted?
No
Replace the current configuration
Yes
Update the current configuration
No
Change temperature?
Yes
Decrease temperature
No
Stopping criteria met?
Yes
Optimal solution

〈Figure 1〉 Basic process of SA

The followings are the simple explanation of how to work of the simulated annealing algorithms. At each step, it randomly chooses a solution similar to the current one, tests its quality, and makes a decision to move to it or not based on either one of two probabilities between which it chooses according to the fact that the new solution is better or worse than the current one. During this search, the temperature is gradually decreased from an initial value and affects the two probabilities.

3. Research Process and Experiments

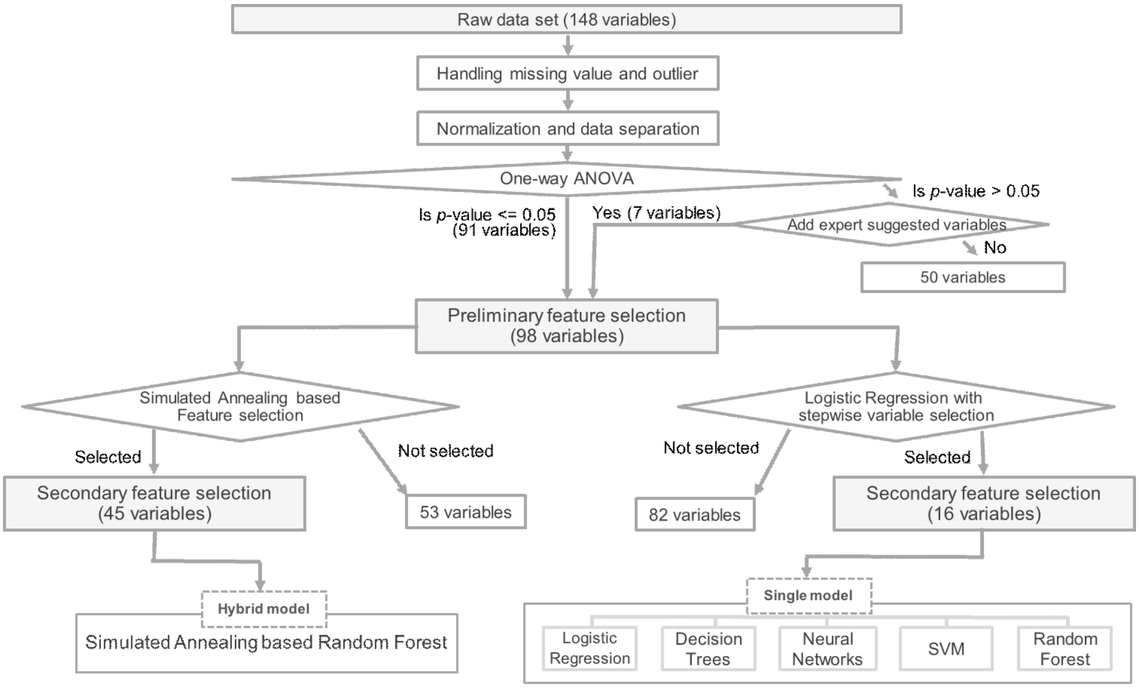
As described above, this study proposes a model that combines a machine learning technique and simulated annealing, which is an optimization technique, for predicting corporate bankruptcy. Simulated annealing is known to have comparable optimization performance to genetic algorithms. It is meaningful to confirm the usefulness of the proposed model because there are few studies on prediction and classification in business decision-making problems using the SA. The following section briefly describes the working process of SA to be used in this study.

In this study, we use the combined model of simulated annealing and machine learning to optimize the input features of the bankruptcy prediction model. A machine learning techniques to be used in the suggested model is the random forest, and the reasons for the selection are known to have the highest forecasting and classification performance in recent studies. There are many ways to combine optimization and machine learning techniques, and typical types are feature selection, feature weighting, and instance selection. In this study, we propose a model that combines machine learning and optimization technique for feature selection, which has been studied the most. Specifically, it can be said that the classification performance of the random forest is a fitness function, and the feature subsets to be used in the random forest is optimized.

To verify the usefulness of the proposed model, we evaluated the model by using the real-world financial data of the Korean heavy industry companies. The research data are used initially in Ok and Kim (2009), but the sampling and splitting of the data sets are configured differently for our research purposes. The research data include 1,426 normal and 1,426 financially bankrupt companies and a total of 2,852 company data are used. The initial data consists of 148 features related to financial ratios and sizes of company.

In this study, the performance of the proposed model is evaluated by comparing the popular machine learning techniques to confirm the usefulness of the results. Machine learning techniques used for comparison are decision tree (CART), artificial neural network, SVM, and logistic regression analysis. On the other hand, we compared traditional random forest models that are not combined with SA to confirm the performance of the combined model. In the case of the proposed model, the iteration of the SA is optimized by increasing the number in the range of 100 to 500 by 100. <Figure 2> shows the research process of this study.

In this study, R-based packages were used for most of the analyses. The decision tree was analyzed using "rpart", the artificial neural network using "nnet", the SVM using "ksvm", and the logistic regression using "glm" and the basic parameters used were the default values in the packages. In the case of Random Forest, which is a part of the combined model, "RandomForest" package was used. For the fairness of the analysis, the maximum number of tree extensions (=500) and the number of used variables (=2) were set


Raw data set (148 variables)
Handling missing value and outlier
Normalization and data separation
One-way ANOVA
Is p-value > 0.05
Is p-value <= 0.05 (91 variables)
Yes (7 variables)
Add expert suggested variables
No
50 variables
Preliminary feature selection (98 variables)
Simulated Annealing based Feature selection
Logistic Regression with stepwise variable selection
Selected
Not selected
Not selected
Selected
Secondary feature selection (45 variables)
53 variables
82 variables
Secondary feature selection (16 variables)
Hybrid model
Simulated Annealing based Random Forest
Single model
Logistic Regression
Decision Trees
Neural Networks
SVM
Random Forest

〈Figure 2〉 Research process in this study equal to the parameters of the proposed model.

In this study, some features were selected by SA and compared them with features selected by a statistical selection process in order to examine the usefulness of the proposed feature selection method. So, we proceeded two steps for the feature selection process. In the first step, we selected only those with a missing value of less than 5% out of 148 features. At the same time, features with the p-value less than 0.05 was selected by one-way ANOVA. As a result, 98 features were selected. Then, in the proposed model, 45 features were selected according to the SA feature selection. For the comparative model, 16 features were selected through the conditional forward stepwise method of logistic regression analysis. Through the above preprocessing and feature selection process, data of 2,852 firms and 45 and 16 feature subsets were finally selected.

In order to confirm the usefulness of the proposed feature selection method, we construct three feature subsets "Features after the preliminary selection", "Features selected by LR" and "Features selected by SA". Then, we apply them to the proposed model and the comparative models to confirm classification performance. "Features after the preliminary selection" is a feature subset selected through one-way ANOVA and it consists of 96 features as mentioned earlier. "Feature selected by LR" and "Feature selected by SA" are derived from the secondary feature selection by applying LR and SA to the feature subset from "Features after the preliminary selection", respectively. The former consists of 16 features, the latter consists of 45 features.

Besides, the data set consisted of 2,282 for training and 570 for validation of the proposed model. We have constructed the same ratio of healthy and insolvent companies in each data set. The selected features are shown in <Table 2> below.

〈Table 2〉 Selected features in this study

(a) Selected features of the proposed model (“Feature selected by SA”)

Feature code	Feature name	Feature code	Feature name
G1	Growth rate of sales	S14	Expected price / Total assets
G2	Operating income growth	S15	Reserve rate
G3	Net income growth rate	S17	Non-Current Liabilities ratio
G4	Total asset growth rate	S20	Accounts payable inventory turnover ratio
G5	Net worth growth rate	S24	Net Working capital to total assets ratio
G7	Growth rate of current assets	S25	Current liabilities ratio
B5	Net Income	S26	Ratio of total reserve assets
B9	Material cost / Sales	S28	Total cash flow loan proportion
B10	Labor costs / Sales	S39	Non-Current Assets Turnover ratio
B11	Overhead costs / Sales	S48	Cash operating profit / Total borrowing

Feature code	Feature name	Feature code	Feature name
B12	Income before income taxes to total assets	D1	Receivables turnover
B13	Net interest expenses to sales	D7	Payables turnover period
B16	Depreciation rate to total expenses	D20	DSO (Days Sales Outstanding)
B18	Interest expenses to total liabilities	D11	Net Working capital turnover period
B19	Financial expenses to total expenses	D14	Stockholders’ equity turnover
B22	Dividends to capital stock	D15	Capital stock turnover
B23	Accumulated earnings ratio to Stockholders’ equity to total assets	D16	Inventories turnover period
B25	Net gain on foreign currency transactions and translation to sales	P1	Gross value-added to machinery and equipment
B26	Research and development costs to sales	P3	GVA (Gross Value Added)
B29	Equity capital dividend rate	P17	GVA ratio
B38	Operating income to business capital	P15	Gross value-added to total assets or productivity of capital
S7	Interest expenses to sales	K1	Professional ability (years of establishment)
S13	Borrowings / BIS capital ratio		

(b) Selected features of the comparative models (“Feature selected by LR”)

Feature code	Feature name
B9	Material cost / sales
B13	Interest expenses to sales
B15	Depreciation rate
B18	Interest expenses to total liabilities
B19	Interest expenses to total expenses
S3	EBITDA / Total borrowings
S4	Borrowings to total assets ratio
S14	Expected price / Total assets
S23	Corporate tax rate
S24	Net Working capital to total assets ratio
S26	Ratio of total reserve assets
S28	Total cash flow loan proportion
S30	Reserve amount paid-in capital rate
D7	Payables turnover period
D15	Capital stock turnover
K1	Professional ability (years of establishment)

In the case of SA, we tried to optimize the search by adjusting the two parameters of the SA (iteration and improvement). Iteration means the number of iterations for the SA and improvement means an integer defining how many iterations should pass without an improvement in fitness before the current subset is reset to the last known improvement. First, we fixed the iteration value to 1000 and checked the change of prediction performance according to the change of the improvement value. For this, the improvement value is increased by 50 from 100 to 300. As a result, the performance of the training set was best when the improvement was 150. Then, we fixed the improvement value to 150 and checked the change of prediction performance according to the change of the iteration number. Iteration values were increased by 500 from 2000 to 3500 in order to see changes in the prediction performances. <Table 3> shows the results of the prediction accuracy depending on the changes in the improvement and the iteration value.

When the improvement value is 3000, and the iteration value is 150, the training set shows the best performance. As a result of applying it to the holdout set using this value, the prediction accuracy is 85.96%, and it is shown that the best prediction performance is obtained. In order to verify the classification performance of the proposed model as described above, the comparative models were analyzed and the results are shown in <Table 4>.

As shown in <Table 4>, in the "Features after the preliminary selection", RF was the best in the

〈Table 3〉 Results of the prediction accuracy with “Feature selected by SA”

Iteration	Improvement	Training set	Holdout set	
		Accuracy	Hit #	Accuracy
1000	100	87.99%	483	84.74%
	150	88.08%	486	85.26%
	200	88.04%	486	85.26%
	250	87.73%	483	84.74%
	300	87.90%	486	85.26%

Improvement	Iteration	Training set	Holdout set	
		Accuracy	Hit #	Accuracy
150	2000	88.12%	490	85.96%
	2500	87.73%	487	85.44%
	3000	88.17%	490	85.96%
	3500	88.03%	484	84.91%

〈Table 4〉 Experimental results with the comparative models

	Features after the preliminary selection			
	Training set		Holdout set	
	Hit #	Accuracy	Hit #	Accuracy
DT	1927	84.44%	452	79.30%
RF	2199	96.36%	478	83.86%
SVM	1990	87.20%	456	80.00%
LR	1861	81.55%	461	80.88%
ANN	2114	92.64%	426	74.74%
	Features selected by LR			
	Training set		Holdout set	
	Hit #	Accuracy	Hit #	Accuracy
DT	1966	86.15%	440	77.19%
RF	2222	97.37%	465	81.58%
SVM	1970	86.33%	469	82.28%
LR	1858	81.42%	462	81.05%
ANN	2040	89.40%	444	77.89%
	Features selected by SA			
	Training set		Holdout set	
	Hit #	Accuracy	Hit #	Accuracy
DT	2014	88.26%	481	84.39%
RF	2237	98.03%	483	84.74%
SVM	1964	86.06%	447	78.42%
LR	1614	70.73%	378	66.32%
ANN	2068	90.62%	430	75.44%

holdout set at 83.86%. In the "Features selected by LR", SVM showed the best classification performance at 82.28%. Finally, in the "Features selected by SA", RF showed the highest classification performance with 84.74%.

The results of <Table 3> and <Table 4> showed that the predictive accuracy of the proposed model is better than that of the simple random forest model. Notably, the performance is significantly improved as compared with the traditional decision tree, artificial neural network, SVM, and logistic regression analysis. In addition, we conducted the two-sample test for proportions to confirm the statistical significance of the differences in the performance of the models. The results are shown in <Table 5>.

〈Table 5〉 Statistical significance for the proposed and the comparative models

(a) Features after the preliminary selection

	RF	SVM	LR	ANN	SA
DT	1.9859**	0.2935	0.668	1.8298**	2.9678***
RF		1.6936**	1.3202*	3.8001***	0.9904
SVM			0.3745	2.1222**	2.6773***
LR				2.4946***	2.306**
ANN					4.7670***

(b) Features selected by LR

	RF	SVM	LR	ANN	SA
DT	1.832**	2.1377**	1.6033*	0.2832	3.8189***
RF		0.3071	0.4091	1.5497*	2.0054**
SVM			0.5366	1.8558**	1.6998**
LR				1.3207*	2.2334**
ANN					3.5404***

(c) Features selected by SA

	RF	SVM	LR	ANN	SA
DT	0.1635	2.5904***	7.0788***	3.7713***	0.7459
RF		2.7523***	7.2333***	3.9317***	0.5825
SVM			4.5681***	1.1942	3.327***
LR				3.3889***	7.779***
ANN					4.5001***

As a result, in the "Features after the preliminary selection", the proposed model showed a significant difference at 1% significance level for DT, SVM and ANN, and a significant difference at 5% significance level for LR. Next, in the "Features selected by LR", the proposed model showed a significant difference at 1% significance level for DT and ANN, and at 5% significance level for RF, SVM and LR. Finally, in the "Features selected by SA", the proposed model showed a significant difference in performance at 1% significance level for SVM, LR, and ANN, but no statistically significant difference in performance for DT and RF. Finally, they showed that the proposed model was superior to most comparative models with several feature subsets, and the differences were also statistically significant.

4. Conclusions

In this study, we propose a model that combines simulated annealing and random forest and apply it to corporate bankruptcy prediction problem. Although the simulated annealing is known to perform similar to the genetic algorithm, it is rarely used in a field of business studies. Therefore, this study tried to confirm its usefulness. As a result, it has resulted in a significant improvement in predictive classification performance for simple machine learning techniques, and it can lead to a slight improvement in performance compared to the simple random forest model, which is part of the hybrid model. Thus, it is concluded that the proposed model can be usefully employed for corporate bankruptcy prediction problem.

On the other hand, this study has several issues to be supplemented. From the viewpoint of model integration, the genetic algorithm-based hybrid model has already been used in combination with machine learning techniques in various fields such as feature weighting and instance selection as well as feature selection. In contrast, this study limited the usage of SA to feature selection only. This suggests that a more advanced model could be proposed in future studies. Besides, the analysis of a single model was not conducted under various conditions as diverse as the SA. We used the basic parameter conditions of each classification technique in this study. In-depth analyses in future studies should complement it.

As described in this paper, the purpose of this study was not to demonstrate that the SA-based feature selection method is superior to the GA-based method but to confirm the usefulness of the SA, since the SA has rarely been used in the field of business analytics. Furthermore, since the GA-based feature selection method has already proved its usefulness in the analysis of business data through prior studies, this study did not analyze the usefulness of the GA, but it can be addressed through further research. Finally, the proposed model showed the statistically significant difference in the performance compared to SVM, LR, and ANN, but did not show the significant difference in the performance compared to DT and RF. This is expected to be solved by collecting enough experimental data and needs to be supplemented in future studies.

References

Ahn, H., and K. Kim, “Bankruptcy prediction modeling with hybrid case-based reasoning and genetic algorithms approach”, *Applied Soft Computing*, Vol.9, No.2(2009), 599-607.

Altman, E. I., “Financial ratios, discriminant analysis and the prediction of corporate bankruptcy”, *The Journal of Finance*, Vol.23 No.4(1968), 589-609

Altman, E. I., G. Marco, and F. Varetto, "Corporate distress diagnosis: comparisons using linear discriminant analysis and neural networks (the Italian experience)", *Journal of Banking and Finance,* Vol.18, No.3(1994), 505-529.

Barboza, F., H. Kimura, and E. Altman, “Machine learning models and bankruptcy prediction”, *Expert Systems with Applications,* Vol.83(2017), 405-417.

Boritz, J. E., and D. B. Kennedy, "Effectiveness of neural network types for prediction of business failure", *Expert Systems with Applications,* Vol.9, No.4(1995), 503-512.

Boritz, J. E., D. B. Kennedy, and A. D. M. E. Albuquerque, "Predicting corporate failure using a neural network approach", *Intelligent Systems in Accounting, Finance and Management*, Vol.4, No.2(1995), 95-111.

Heo, J. Y., and J. Y. Yang, "Bankruptcy forecasting model using AdaBoost: a focus on construction companies", *Journal of Intelligence and Information Systems*, Vol.20, No.1(2014), 35-48.

Hong, S. H., and K. Shin, “Using GA based input selection method for artificial neural network modeling: application to bankruptcy prediction”, *Journal of Intelligence and Information Systems,* Vol.9, No.1(2003), 227-249.

Jo, H., and I. Han, “Integration of case-based forecasting, neural network, and discriminant analysis for bankruptcy prediction”, *Expert Systems with Applications,* Vol.11, No.4(1996), 415-422.

Jo, H., I. Han, and H. Lee, "Bankruptcy prediction using case-based reasoning, neural networks, and discriminant analysis", *Expert Systems with Applications,* Vol.13, No.2(1997), 97-108.

Kim, K., “Data mining using instance selection in artificial neural networks for bankruptcy prediction”, *Journal of Intelligent Information System*, Vol.10, No.1(2004), 109-123.

Kim, S. H., and J. W. Kim, "SOHO bankruptcy prediction using modified bagging predictors", *Journal of Intelligence and Information Systems,* Vol.13, No.2(2007), 15-26.

Kim, T., and H. Ahn, “A hybrid under-sampling approach for better bankruptcy prediction”, *Journal of Intelligent Information System*, Vol.21, No.2(2015), 173-190.

Kirkpatrick, S., C. D. Gelatt, and M. P. Vecchi, “Optimization by simulated annealing”, *Science*, Vol.220, No.4598(1983), 671-680.

Kiviluoto, K., "Predicting bankruptcies with the self-organizing map", *Neurocomputing*, Vol.21(1998), 191-201.

Kwon, H., D. Lee, and M. Shin, “Dynamic forecasts of bankruptcy with recurrent neural network model”, *Journal of Intelligent Information System*, Vol.23, No.3(2017), 139-153.

Lee, K. C., I. Han, and Y. Kwon, "Hybrid neural network models for bankruptcy predictions", *Decision Support Systems*, Vol.18, No.1(1996), 63-72.

López. I. F. J., and I. P. Sanz, "Bankruptcy visualization and prediction using neural networks: A study of US commercial banks", *Expert Systems with Applications,* Vol.42, No.6(2015), 2857-2869.

Martin-del-Brio, B., and C. Serrano-Cinca, "Self-organizing neural networks for the analysis and representation of data: Some financial cases", *Neural Computing & Applications*, Vol.1, No.3(1993), 193-206.

Nanni, L., and A. Lumini, “An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring", *Expert Systems with Applications,* Vol.36, No.2(2009), 3028-3033.

Ohlson, J. A., "Financial ratios and the probabilistic prediction of bankruptcy", *Journal of Accounting Research*, Vol.18, No.1(1980), 109-131.

Ok, J. K. and K. Kim, "Integrated corporate bankruptcy prediction model using genetic algorithms", *Journal of Intelligent Information System*, Vol.15, No.4(2009), 99-120.

Serrano-Cinca, C., "Self-organizing neural networks for financial diagnosis", *Decision Support Systems*, Vol.17, No.3(1996), 227-238.

Serrano-Cinca, C., "Feedforward neural networks in the classification of financial information", *The European Journal of Finance*, Vol.3, No.3(1997), 183-202.

Shin, K., and Y. J. Lee, "A genetic algorithm application in bankruptcy prediction modeling", *Expert Systems with Applications,* Vol.23, No.3(2002), 321-328.

Shin, K., T. S. Lee, and H. J. Kim, "An application of support vector machines in bankruptcy prediction model", *Expert Systems with Applications,* Vol.28, No.1(2005), 127-135.

Tam, K. Y., and M. Y. Kiang, "Managerial applications of neural networks: the case of bank failure predictions", *Management Science*, Vol.38, No.7(1992), 926-947.

Tsai, C. F., Y. F. Hsu, and D. C. Yen, "A comparative study of classifier ensembles for bankruptcy prediction", *Applied Soft Computing*, Vol.24(2014), 977-984.

Wang, G., J. Ma, and S. Yang, "An improved boosting based on feature selection for corporate bankruptcy prediction", *Expert Systems with Applications,* Vol.41, No.5 (2014), 2353-2361.

Wilson, R. L., and R. Sharda, "Bankruptcy prediction using neural networks", *Decision support systems*, Vol.11, No.5(1994), 545-557.

Zhang, G., M. Y. Hu, B. E. Patuwo, and D. C. Indro, "Artificial neural networks in bankruptcy prediction: General framework and cross-validation analysis", *European journal of operational research*, Vol.116, No.1(1999), 16-32.

국문요약

시뮬레이티드 어니일링 기반의 랜덤 포레스트를 이용한 기업부도예측*

박호연** · 김경재***

기업의 금융 부도를 예측하는 것은 전통적으로 비즈니스 분석에서 가장 중요한 예측문제 중 하나이다. 선행연구에서 예측모델은 통계 및 기계학습 기반의 기법을 적용하거나 결합하는 방식으로 제안되었다. 본 논문에서는 잘 알려진 최적화기법 중 하나인 시뮬레이티드 어니일링에 기반한 새로운 지능형 예측모델을 제안한다. 시뮬레이티드 어니일링은 유전자알고리즘과 유사한 최적화 성능을 가진 것으로 알려져 있다. 그럼에도 불구하고, 시뮬레이티드 어니일링을 사용한 비즈니스 의사결정 문제의 예측과 분류에 관한 연구가 거의 없었기 때문에, 비즈니스 분석에서의 유용성을 확인하는 것은 의미가 있다. 본 연구에서는 시뮬레이티드 어니일링과 기계학습의 결합 모델을 사용하여 부도예측모델의 입력 특징을 선정한다. 최적화 기법과 기계학습기법을 결합하는 대표적인 유형은 특징 선택, 특징 가중치 및 사례 선택이다. 이 연구에서는 선행연구에서 가장 많이 연구된 특징 선택을 위한 결합모델을 제안한다. 제안하는 모델의 우수성을 확인하기 위하여 본 연구에서는 한국 기업의 실제 재무데이터를 이용하여 그 결과를 분석한다. 분석결과는 제안된 모델의 예측 정확도가 단순한 모델의 예측 정확성보다 우수하다는 것을 보여준다. 특히 기존의 의사결정나무, 랜덤포레스트, 인공신경망, SVM 및 로지스틱 회귀분석에 비해 분류성능이 향상되었다.

주제어 : 시뮬레이티드 어니일링, 랜덤 포레스트, 부도예측, 특징선택, 비즈니스 애널리틱스

논문접수일 : 2018년 11월 30일　논문수정일 : 2018년 12월 11일　게재확정일 : 2018년 12월 12일
원고유형 : 일반논문　교신저자 : 김경재

* 이 논문은 2017년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (No. NRF-2017R1D1A1B03032770)

** 동국대학교_서울 일반대학원 경영정보학과
*** 교신저자: 김경재
동국대학교_서울 경영대학 경영정보학과
04620 서울특별시 중구 필동로 1길 30
Tel: +82-2-2260-3324, E-mail: kjkim@dongguk.edu

저자소개

박호연
동국대학교에서 컴퓨터공학을 전공하여 공학사, 경영정보학을 전공하여 경영학석사를 취득하였으며, 현재 동교에서 경영정보학을 전공하여 박사과정을 수료하였다. 주요 관심분야는 빅데이터, 비즈니스 애널리틱스, 텍스트마이닝 등이다.

김경재
현재 동국대학교 경영대학 경영정보학과 교수로 재직 중이다. KAIST에서 경영정보시스템을 전공으로 박사학위를 취득하였으며, 연구관심분야는 비즈니스 애널리틱스, CRM, 추천기술, 빅데이터 등이다.