INTERNATIONAL JOURNAL OF SUSTAINABLE CONSTRUCTION ENGINEERING AND TECHNOLOGY
VOL. 13 NO. 3 (2022) 68-79

IJSCET

http://publisher.uthm.edu.my/ojs/index.php/ijscet

ISSN : 2180-3242 e-ISSN : 2600-7959

International Journal of Sustainable Construction Engineering and Technology

A Back Propagation Neural Network Model with the Synthetic Minority Over-Sampling Technique for Construction Company Bankruptcy Prediction

Ngo Thanh-Long[1*], Tran-Minh[2], Le Hong-Chuong[1]

[1]Department of Construction Mechanical Engineering,
Hanoi University of Civil Engineering, 55 Giai Phong Road, Hai Ba Trung District, Hanoi, 10000, VIETNAM

[2]Joint stock Commercial Bank for Investment and Development of Viet Nam,
35 Hang Voi Road, Hoan Kiem District, Hanoi, 10000, VIETNAM

*Corresponding Author

DOI: https://doi.org/10.30880/ijscet.2022.13.03.007
Received 24 February 2022; Accepted 3 July 2022; Available online 10 December 2022

Abstract: Improving model accuracy is one of the most frequently addressed issues in bankruptcy prediction. Several previous studies employed artificial neural networks (ANNs) to enhance the accuracy at which construction company bankruptcy can be predicted. However, most of these studies use the sample-matching technique and available company quarters or company years in the dataset, resulting in sample selection biases and between-class imbalances. This study integrates a back propagation neural network (BPNN) with the synthetic minority over-sampling technique (SMOTE) and the use of all of the available company-year samples during the sample period to enhance the accuracy at which bankruptcy in construction companies can be predicted. In addition to eliminating sample selection biases during the sample matching and between-class imbalance, these methods also achieve the high accuracy rates. Furthermore, the approach used in this study shows optimal over-sampling times, neurons of the hidden layer, and learning rate, all of which are major parameters in the BPNN and SMOTE-BPNN models. The traditional BPNN model is brought as a benchmark for evaluating the predictive abilities of the SMOTE-BPNN model. The experiential results of this paper indicate that the SMOTE-BPNN model outperforms the traditional BPNN.

Keywords: Bankruptcy prediction, artificial neural networks, back propagation neural network, receiver operating characteristics, minority class (MC)

1. Introduction

Improving model accuracy is one of the most frequently addressed issues in bankruptcy prediction (Jardin, 2010). There are some ways in which the accuracy of the prediction model can be improved, such as the use of various methods, role evaluation of variables, and so on. The use of various methods was developed using statistical and artificial intelligent techniques, such as univariate analysis, multivariate discriminant analysis (MDA), the logistic regression (LR) model, artificial neural networks (ANNs), the neuro-fuzzy approach, and support vector machines (SVM). Thus, there are many leading researchers that have investigated this topic, including Beaver (1966); Meyer & Pifer (1970); Ohlson (1980); Theodossiou (1993); Chen, Huang, & Lin (2009); Chen (2011), and Jeong, Min, & Kim (2012), and so on. However,

*Corresponding author: longnt@huce.edu.vn

these studies only focus on non-construction industries, and they cannot be applied to the construction industry due to different characteristics between industries. Chava & Jarrow (2004) stated that different industries own different accounting conventions and levels of competition; therefore, the probability of bankruptcy can be different.

Tserng et. al., (2011); Tserng et. al., (2011); Chen (2012); Chang (2011); Chen (2009), and Kangari, Farid, & Elgharib (1992) have indicated that the special characteristics and financial risks of the construction industry are significantly different. First, because the construction is a project-based industry, projects dominate most of the company's operation; thus, projects have a large influence on a company's financial performance. Second, construction companies often have to deal with large-size projects, and their values may include total company assets. As a result, the capital structures of construction companies are rather different from those in other types of industries. Third, construction companies face a high degree of uncertainty and high operational risks due to technical, human, and natural factors, such as floods and earthquakes. Fourth, the construction industry is easily influenced by the current economic situation, governmental regulations, public policy issues, and the business cycle. Fifth, construction companies may suffer severe financial loss. Sixth, as inventory cannot be realized into cash due to contract disputes, construction companies suffer from insufficient liquidity due to the combination of the construction industry's risky behavior and its excessive optimism in revenue recognition. Seventh, the capital resources in construction companies are relatively unstable, while the interest rate payments are fairly high.

Accordingly, there are many studies have focused on improving bankruptcy prediction models for the construction industry by developing MDA and LR models (Mason & Harris, 1979; Kangari, Farid, & Elgharib, 1992; Russell & Jaselskis, 1992; Severson, Russell, & Jaselskis, 1994; Abidali & Harris, 1995; Russell & Zhai, 1996). However, these models need strict assumptions, for example normality, linearity, a pre-existing functional form, and independence among predictor variables relating to the criterion and predictor variables limit their application in reality (Neves & Vieira, 2006; Hua et. al., 2007).

Thus, some scholars have begun applying non-parametric models, such as ANNs, in the construction industry (Pompe & Feelders, 1997; Al-Sobiei, Arditi, & Polat, 2005; Chen, 2012) and enforced support vector machines (Tserng, Lin, Tsai, & Chen, 211). In these techniques, the back propagation neural network (BPNN) is the most common type of ANN for predicting bankruptcy in non-construction industries because it is the simplest and most reliable classifier and achieves a high level of accuracy; thus, it has attracted a numerous number of researchers (Neves & Vieira, 2006; Boyacioglu, Kara, & Baykan, 2009; Lin, 2009; Jardin, 2010; Jeong, Min, & Kim, 2012; Chen, 2012). The models using a BPNN outperform the other existing models, such as MDA, LR, SVM, and data mining (DM) models (Boyacioglu, Kara, & Baykan, 2009; Jeong, Min, & Kim, 2012). In contrast, there are only a few studies that have used a BPNN to predict bankruptcy in the construction industry, and when a BPNN is used, studies employ the sample-matching technique and all of the available company quarters or company years in the dataset to construct their model, leading to sample selection biases and between-class imbalance (Zmijewski, 1984; Tserng et. al., 2011). Thus, this study uses the synthetic minority over-sampling technique (SMOTE) and all of the available company-year samples in the BPNN to resolve these limitations and thus improve the performance of the model.

The contributions of this research are as follows. First, this study integrates a BPNN with the SMOTE and uses all of the available company-year samples during the sample period to build a new model (SMOTE-BPNN) for predicting the probability of bankruptcy in construction companies. Second, the approach employed in the model may resolve existing approaches' sample selection biases and between-class imbalance. Third, the accuracy rates of the BPNN and SMOTE-BPNN models are examined and compared, and the optimal over-sampling times, learning rate, and neurons of the hidden layer, all of which are major parameters in the BPNN and SMOTE-BPNN models, are shown. These results could allow users to reduce a sizable amount of unnecessary calculations when predicting bankruptcy. Fourth, the research shows that the SMOTE-BPNN model outperforms the BPNN model. Finally, the proposed SMOTE-BPNN model could be used to investors, assist managers, auditors, and the government in the US for predicting the probability of construction company bankruptcy, and it may be used as a reference for non-construction industries.

The rest of this paper includes five sections as follows. Section 2 reviews the literature on the techniques and models related to bankruptcy prediction. Section 3 introduces the methodology of the research, including the BPNN and SMOTE. Section 4 describes the collected data and selected variables. Section 5 provides empirical results and important discussions. Finally, section 6 presents the conclusions drawn from this research.

2. Literature Review

Numerous studies have used various methods to improve the prediction models of company bankruptcy in non-construction industries, including the univariate analysis model (Beaver, 1966), linear probability model (LPM) (Meyer & Pifer, 1970), LR model (Ohlson, 1980), MDA (Deakin, 1976; Taffler, 1982), probit model, cumulative sums (CUSUM) procedure (Theodossiou, 1993), neuro-fuzzy approach (Chen, Huang, & Lin, 2009), SVM (Kim & Sohn 2010; Chen, 2011), ANN models (Neves & Vieira, 2006; Boyacioglu, Kara, & Baykan, 2009; Lin, 2009; Kim & Kang, 2010; Jardin, 2010; Jeong, Min, & Kim, 2012), and so on. In these models, the BPNN of the ANN is the most frequently utilized technique for classifying and predicting bankruptcy (Neves & Vieira, 2006; Boyacioglu, Kara, & Baykan, 2009; Lin, 2009; Chen & Lin, 2009). However, these studies may lack the ability to assess the construction industry accurately due to the differences between the construction industry and non-construction industries.

Although company bankruptcy prediction models for the construction industry have been around since the late 1970s, only a few current studies have focused on this topic due to the limited number of bankruptcy samples and the construction industry's unique characteristics (Chen, 2009; Tserng et. al., 2011; Tserng et. al., 2011; Chen, 2012). Mason and Harris (1979) developed a six-variable Z-score model that classified a construction firm as being long-term solvent if it had a positive Z-score and potentially insolvent if it had a negative Z-score. Abidali (1990) developed a seven-variable Z-score model and suggested that a score of 2.94 was the minimum value for indicating long-term solvency. Kangari, Farid, & Elgharib (1992) presented a six-variable model for predicting a firm's financial performance using multiple-regression analysis. This model graded the construction firm based on the characteristics of various construction trades and the affection of the firm size. To improve the prediction probability performance, Russell & Jaselskis (1992) proposed a model for predicting the failure probability of a given construction project before contracting award based on project characteristics and economic-related factors. However, many of the factors introduced in this model are qualitative and depend on human judgment. Severson, Russell, & Jaselskis (1994) used LR to develop models for predicting claim and non-claim contracts and reported that there was a misclassification rate of greater than 30% when using corporate financial variables. Abidali & Harris (1995) presented an A-score model that included both conventional financial variables and trend measurement variables. The model linked the A-score and Z-score values, thus making it possible to predict the probability of construction contractor bankruptcy more accurately. Nevertheless, these managerial performance variables are subjective and qualitative. Russell & Zhai (1996) proposed a contractor failure prediction model with financial variables and the stochastic dynamics of economic. Their model found a misclassification rate of 15.5% for a testing sample and 22% for a validation sample. These researchers used mathematical statistics with its aforementioned limitations to create an MDA or LR models. As mentioned previously, some of these statistics require strict assumptions, such as normality, linearity, the pre-existing functional form, and independence among predictor variables relating to the criterion and predictor variables limit their application in reality (Hua et. al., 2007). Some methods are very sensitive to exceptions, which are common in bankruptcy, and most conclusions in non-linear models have an implicit Gaussian distribution, which is often inappropriate (Neves & Vieira 2006). Chen (2009) described a model for predicting financial performance with thirty-six variables, including both economic and financial variables, but the accurate prediction rate of the model only reaches 78.9%. Tserng et. al., (2011) offered a model for predicting bankruptcy in construction contractors by basing enforced support vector machine. They compared this model with the traditional LR model and found that its accurate prediction rate of 80.31% using seven variables was more exact than the LR model. Tserng et. al., (2015) used the grey system theory to predict financial performance.

Several scholars recently used the BPNN of an ANN to propose models for predicting construction company bankruptcy. Pompe & Feelders (1997) made a comparison between the performances of neural networks, linear discriminant analysis, and classification tree models in ten experiments. The study's result indicated that the neural network performs more better than the other two models, with achieving a 70% accuracy. The study uses ten variables and 576 annual reports of Belgian construction companies selected from 175,000 construction firms. Al-Sobiei, Arditi, & Polat (2005) used an ANN to predict contractor bankruptcy and achieved a 75% accuracy with using 102 non-bankruptcy and 78 bankruptcy contractors from the US. Chen (2012) presented a new model that combines fuzzy, neural networks, and self-organizing feature map optimization. The model achieves an accuracy of 85.1% for predicting construction financial distress in Taiwan. However, these models also use all of the available company quarters or company years in the dataset, and most of them have low accuracy rates.

3. Methodology

The framework of this study's methodology is displayed in Fig. 1. The process of predicting construction company bankruptcy is as follows. First, the sample set and variables for the BPNN and SMOTE-BPNN models are selected. Then, the sample set is put into the BPNN and SMOTE-BPNN models, and the SMOTE is executed for the SMOTE-BPNN model. Finally, the results obtained from these models are compared and analyzed. The BPNN and SMOTE-BPNN models are introduced in the following sections.

3.1 Back Propagation Neural Network (BPNN)

3.1.1 Brief Concept of the BPNN

ANN models have occupied an important position in bankruptcy prediction research due to certain features. ANNs are sets of algorithms based on the function of the human brain and good processing capabilities and do not require before understanding of the problem. Classifying data can be considered as a regression problem in which a function that draws an input into the corresponding class is found while minimizing the misclassification rate. ANNs hold inside non-linear regression capabilities that make them highly competitive for difficult classification problems (Bishop, 1995). In addition, ANNs are large parallel processing systems, and their ability to perform high-speed calculations and tolerate mistakes allow them to filter noise from the training data (Lin, Lin, & Chang, 2002).

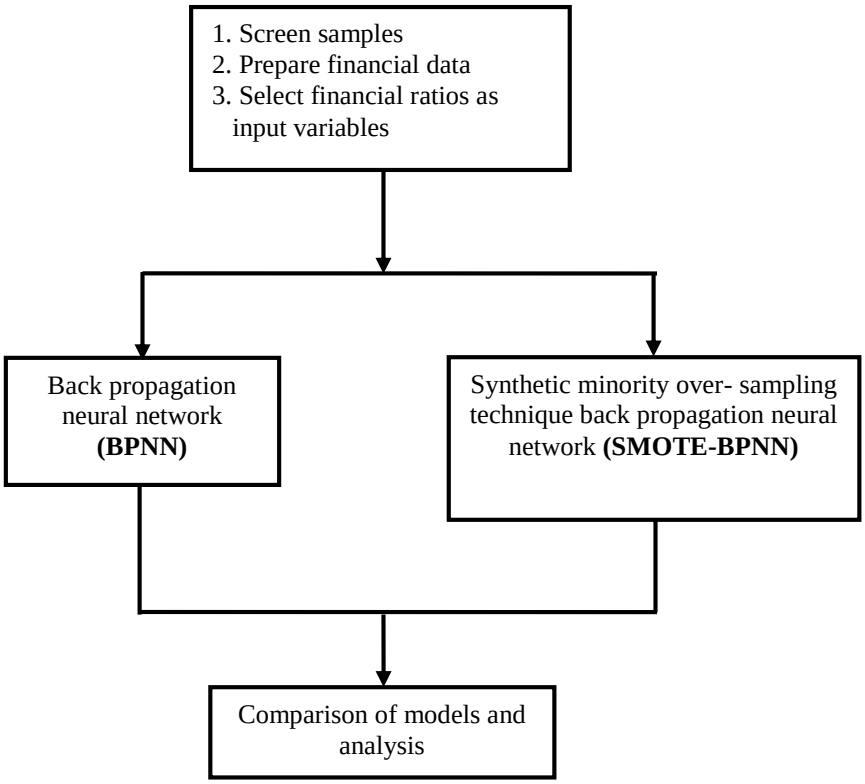

Fig. 1 - Methodology framework

BPNNs use a monitored learning method and the feed-forward architecture of ANNs. Fig. 2 shows the most common BPNN, which includes one input layer, a single hidden layer, and one output layer. It has l input neurons, m hidden neurons, and n output neurons. The information from the input layer is gone through the network via connecting weights to the hidden layer and then the output layer. The weights connecting input elements i to hidden neurons j are denoted as Wji, while the weighted connecting hidden neurons j to output neurons k are denoted as Wkj. Each neuron calculates its output based on the amount of simulation it receives as an input. A neuron's net input is calculated as the weighed sum of its inputs, and the output of the neuron is based on a sigmoid function and depends on the magnitude of this net input.

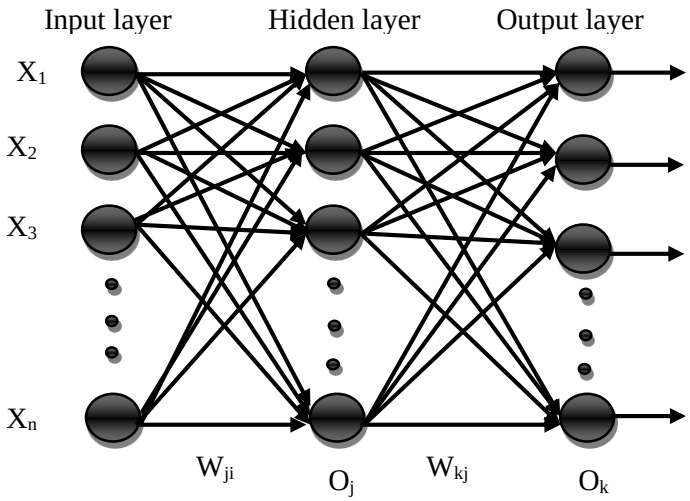

Fig. 2 - Back propagation neural network architecture

Put $I_c = (I_{c1}, I_{c2}, \ldots, I_{cl})$, $c = 1, 2, \ldots, N$ be the *cth* pattern between *N* input patterns, where W_{ji}, Wkj are the connection weights among the *ith* input neuron and the *jth* hidden neuron, and the *jth* hidden neuron and the *kth* output neuron, respectively.

Output from a neuron in the input layer is

$$O_{ci}=I_{ci},\ i=1, 2, ..., I \quad (1)$$

Output from a neuron in the hidden layer is

$$O_{cj} = f(NET_{cj}) = f\left(\sum_{i=0}^{I} W_{ji}\, O_{ci}\right), j = 1, 2, \ldots, m \quad (2)$$

Output from a neuron in the output layer is

$$O_{ck} = f(NET_{ck}) = f\left(\sum_{j=0}^{m} W_{kj}\, O_{cj}\right), k = 1, 2, \ldots, n \quad (3)$$

Where $f()$ is the sigmoid function given by $f(x) = 1/(1+e^{-x})$.

3.1.2 Main Parameters of the Proposed BPNN

Previous studies have used the same numbers of neurons in the hidden layers and the same input layers (Odom & Sharda, 199; Lee, Han, & Kwon, 1996; Lee & To, 2010; Jeong, Min, & Kim, 2012). The proposed BPNN includes 20 neurons for both the input and hidden layers. However, according to Lee & To (2010)'s suggestion, the paper also uses the hidden layer with 14 neurons for a comparison with the aforementioned case. The output layer uses only one neuron with a response of 0 displaying bankruptcy and a response of 1 displaying non-bankruptcy. The learning rate in this study ranges from 0.4 to 0.8 referred to previous studies (Odom & Sharda, 1990; Lee & To, 2010; Chen & Lin, 2009; Jardin, 2010) in which these studies have not used the range of 0.4-0.8. The maximum number of interactions in the training model is set at 10,000.

3.2 Over-Sampling Technique

3.2.1 Problem with Imbalance Data and The Proposed Models

Most of the previous studies used the "sample-matching" method, in which a set of bankrupt companies is matched with the same number or some multiple of non-bankrupt companies due to the limited number of bankrupt companies. According to Zmijewski (1984), this method results in biases, and he argued that if a model could not be constructed using the entire population, the estimated coefficient would likely be inadequate and thus the prediction outcome would be untrustworthy. To avoid this problem and thus enhance a model's prediction performance, many recent studies used all of the available company quarters or company years in the dataset to construct their model (Brockman & Turtle, 2003; Reisz & Perlich, 2007; Hillegeist et. al., 2004; Gharghori, Chan, & Faff, 2006; Agarwal & Taffler, 2008; Chen, 2012).

Thus, in this research, we used all of the available company-year data to construct the BPNN and SMOTE-BPNN models for predicting the probability of bankruptcy in construction companies. After inputting all of the company year data, there was a significant difference in the sample dimensions of bankrupt and non-bankrupt companies, with the number of non-bankrupt samples exceeding the number of bankrupt samples. This form of imbalance is referred to as between-class imbalance (Gao et. al., 2011). Weiss & Provst (2003); Estabrooks, Jo, & Japkowicz (2004) have shown that balanced datasets provided better classification performance than those that are imbalanced. Similar to other ANN models, the BPN only represented the distribution of the major parts of the input points, while it overlooked the small parts of the input points (Tserng et. al., 2011). To tackle this problem, certain important information should be emphasized through certain techniques (Chang, Chang, & Wang, 2007). In particular, SMOTE suggested by Chawla et. al., (2002) provides a convenient and effective way for treating with imbalanced learning problems and has been applied in numerous studies.

3.2.2 Synthetic Minority Over-Sampling Technique (SMOTE)

Chawla et. al., (2002) showed an over-sampling approach referred to as SMOTE. The SMOTE's minority class (MC) is over-sampled by generating "synthetic" samples. The MC is over-sampled by showing synthetic examples into this each class sample through the line fragments connecting any or all of the k minority class' nearest neighbors. Depending on the number of over-sampling needed, neighbors from the k nearest neighbors are randomly taken. Next, the dissimilarity between the original minority sample under consideration and its nearest neighbor are computed. This difference is multiplied by a random value between [0, 1] and supplemented to the original sample under consideration, resulting in a new "synthetic" sample. This technique helps handle the constraints mentioned by simple oversampling and augments the original dataset in a manner that significantly improves learning; thus, this technique has shown many promising benefits. SMOTE can raise the accuracy of classifiers for a MC. Because it creates more related minority class samples to learn from, thus allowing a learner to carve broader decision regions, leading to more coverage of the MC.

4. Data and Variable Selection

4.1 Data

The empirical investigation of this study considers a fairly large cross section of construction companies in the US. This research collects data from the Center for Research in Securities Prices (CRSP) and the Compustat Industrial File-Quarterly data (Wharton Research Data Services, 2009). Sample construction companies are on the New York Stock Exchange (NYSE), NASDAQ, and American Exchange (AMEX) and cover the period from 1970 to 2008. This research limits its attention to construction companies with December fiscal year-ends by selecting companies with standard industrial classification (SIC) codes between 1500 and 1799. Similar to the studies of Severson, Russell, & Jaselskis (1994); Russell & Zhai (1996); Tserng et. al., (2011); Tserng et. al., (2011), the sample firms have three construction categories: (1) building construction (SIC codes 1500 to 1599), (2) heavy construction (SIC codes 1600 to 1699), and (3) special trade construction (SIC codes 1700 to 1799).

According to Tserng et. al., (2011), the sample selection meets the following three criteria. First, data must be present in the CRSP for at least two years before the time of bankruptcy for data completeness. Next, construction companies that do not have financial reports for at least two years are removed from the sample. Third, the delisting code assigned by the CRSP is used to define bankruptcy samples in this study. The research follows the definition of bankrupt companies proposed by Dichev (1998) and Brockman & Turtle (2003), and this study defines bankruptcy samples using the delisting codes of 400 and 550 to 585, which represent companies delisted due to liquidation, bankruptcy, or poor performance. Forty-four bankrupt construction companies were identified based on this methodology. Because this paper aims to predict bankruptcy within one year, the financial statements of the years immediately prior to bankruptcy are used as the bankruptcy samples. Prior research typically involves "selecting" a group of non-bankrupt companies on which to perform the analysis. However, this method may produce sample selection biases. To avoid sample selection biases, following Tserng et. al., (2011), the analysis in this paper uses every company year for which data are available. The final combined sample consists of 1,262 company-year observations, including 44 bankrupt and 1,218 non-bankrupt samples from 155 construction companies.

4.2 Input Variable Selection

In this study, we selected twenty financial ratios of companies as input variables for the bankruptcy prediction analysis, as shown in Table 1. These variables were selected for the following two reasons. First, all of the variables selected were among those most commonly used in prior studies that used construction firm bankruptcy prediction models (Abidali & Harris, 1995; Kangari, Farid, & Elgharib, 1992; Mason & Harris, 1979; Russell & Zhai, 1996; Severson, Russell, & Jaselskis, 1994; Severson, Jaselskis, & Russell, 1993; Kangari & Bakheet, 2001; Tserng et. al., 2011). Second, these variables include a broad cross section of accounting ratios that describe a company's profitability, liquidity, leverage, and activity. As a group, these ratios capture the financial characteristics and performance of the construction industry.

Table 1 - Number and characteristics of the financial variables selected in this paper

Liquidity	Leverage	Activity	Profitability
1. Current ratio	5. Total liabilities to net worth	9. Revenues to net working capital	17. Return on assets (ROA)
2. Quick ratio	6. Retained earnings to sales	10. Accounts receivable turnover	18. Return on equity (ROE)
3. Net working capital to total assets	7. Debt ratio	11. Accounts payable turnover	19. Return on sales (ROS)
4. Current assets to net assets	8. Times interest earned	12. Sales to net worth	20. Profits to net working capital
		13. Quality of inventory	
		14. Fixed assets to net worth	
		15. Turnover of total assets	
		16. Revenues to fixed assets	

5. Results and Discussion

5.1 Discriminatory Power

This paper uses discriminatory power to evaluate and compare the accuracy rate of bankruptcy probability prediction cases; discriminatory power also indicates which case has the best predictive performance for construction company bankruptcy risk. The discriminatory power measures the extent to which the model can differentiate companies that are more likely to bankrupt from companies that are less likely to bankrupt. In a perfectly discriminating model, all of the companies that actually file for bankruptcy are assigned a larger probability of bankruptcy than any surviving company. The receiver operating characteristics (ROC) curve has been widely performed in the field of medicine for testing the efficaciousness of various treatments and diagnostic techniques. It has also been a common technique for evaluating the discriminatory power of various credit scoring and rating cases (Agarwal & Taffler, 2008; Stein, 2007; Tserng, Liao, Tsai, & Chen, 2011; Tserng, Lin, Tsai, & Chen, 2011; Chawla, Bowyer, Hall, & Kegelmeyer, 2002).

The ROC curve is created by scoring all of the credits, arranging the non-bankruptcies from riskiest to safest on the x-axis, and then plotting the percentage of bankruptcies excluded at each level on the y-axis. Thus, the y-axis is constructed by combining each score on the x-axis with the cumulative percentage of bankruptcies with a score equal to or worse than that score in the test data. In other words, the ROC curve plots the Type II error against one minus the Type I error. In the case of bankruptcy prediction, it shows the percentage of non-bankrupt companies that must be involuntarily refused credit (Type II) to avoid lending to a special percentage of (1–Type I) when using a special bankruptcy model (Stein, 2007). The ROC curve presents different relative performances across all of the possible cut-off points associated with the costs of each type of classification error and shows a form of cost-benefit analysis for decision makers.

The ROC curve of a completely random prediction that is the main diagonal, while a perfect model will be a ROC curve that draws straight up from (0, 0) to (0, 1) and then across to (1, 1). In two models, the model with a better ranking will draw a ROC curve that is further to the top left than the other model. The area below the curve (AUC) is usually used as a summary statistic for the quality of a ranking. A model with a perfect ranking receives an AUC of 1, while a model with constant or random predictions receives an AUC of 0.5 (Reisz & Perlich, 2007). The general rule is as follows: if $AUC \geq 0.9$, the model has an outstanding discrimination; if $0.8 \leq AUC < 0.9$, the model has an excellent discrimination; if $0.7 \leq AUC < 0.8$, the model has an acceptable discrimination; if $AUC = 0.5$, the model has no discriminatory power (Hosmer & Lemeshow, 2000).

5.2 Validation Process

The quality of a classifier is measured by its generalization capabilities and robustness. It is very essential to avoid over-fitting, which is a common problem with neural networks. The dataset is separated into two groups: a training set and a test set. The BPNN and SMOTE-BPNN models are trained within the training set, and their performance is tested on the test set. The predictive performance in the dataset will be evaluated by the ROC curve. Cross-validation is the most adequate procedure for evaluating the generalization capabilities of the network. In this research, we used a five-fold cross-validation. This validation procedure consists of dividing the dataset into 5 sets (Test1, ..., Test5), using four (Test2, ..., Test5) for training and the remaining Test1 for testing. When training is completed, the test error e1 is recorded and the process is repeated: training with (Test1, Test3, Test4, Test 5), testing with Test2, and test error e2 is recorded. After completing the 5 cycles, the generalization error, or cross-validation error, is calculated as the average of the test set errors (average AUC value). As a result, the proportions of the training dataset and test dataset are 80% and 20%, respectively. The numbers of bankruptcy and non-bankruptcy companies of each group are presented in Table 2. The final combined sample consists of 1,262 company-year observations, including 44 bankrupt and 1,218 non-bankrupt samples from 155 construction companies. Moreover, the dataset will be divided into 5 sets in which 4 sets used for training and one set used for testing. Consequently, the 1262 firm-years will be separated into 5 sets with the number of bankrupt and non-bankrupt in each group, as shown in table 2.

Table 2 - The number of test sets

Test set	Non-bankruptcy sample	Bankruptcy sample
Test set 1	243	9
Test set 2	243	9
Test set 3	243	9
Test set 4	244	9
Test set 5	245	8

5.3 Validation Result

Table 3 shows the accuracy rate of cases in which there are 14 and 20 neurons in the hidden layer corresponding to learning rates between 0.4 and 0.8. Both cases have a relatively high accuracy ($0.774210 \leq AUC \leq 0.798689$), and their accuracy rates are similar. These results indicate that when a user takes the hidden layer with14 neurons into the BPNN, the degree of accuracy is very high while the calculations are relatively simple. With different learning rates, the accuracy rate of both cases changes slightly, implying that the user should select a learning rate ranging from 0.4 to 0.8. When using 14 neurons, the best accuracy is 79.8445% for a learning rate of 0.4, while the best accuracy for the 20-neuron case is 79.8689% for a learning rate of 0.7. As a result, the 20-neuron case corresponding to a learning rate of 0.7 achieves a better accuracy than the 14-neuron case. Accordingly, this study selects two major parameters for the SMOTE-BPNN model, including the hidden layer has 20 neurons and the learning rate is 0.7. In this case, the numbers of input variables and neurons in the hidden layer are the same, and it is suitable to results from some previous studies (Odom & Sharda, 1990; Lee, Han, & Kwon, 1996; Lee & To, 2010; Jeong, Min, & Kim, 2012).

Table 3 - AUC for each learning rate and hidden neuron combination

Case	Learning rate				
	0.4	0.5	0.6	0.7	0.8
14 neurons	0.79845	0.79667	0.77426	0.78134	0.78443
20 neurons	0.79393	0.79355	0.79203	0.79869	0.78335

After applying the SMOTE through several over-sampling times, the new datasets were created. Then, these new datasets were trained by the BPNN model. Finally, throughout the validation process, the ROC curves were drawn, and the average AUC values were calculated, as shown in Table 4. The SMOTE-BPNN model's accuracy rates corresponding to over-sampling times fluctuate slightly. The model achieves relatively high accuracy rates from 79.55% to 84.10%, and most cases attain accuracy rates of greater than 80%. The best accuracy rate corresponded to 6 over-sampling times, while the worse accuracy rate of 79.55% corresponded to 10 over-sampling times. It can be found that in the SMOTE, the optimal over-sampling times are very important due to their effect on the degree of accuracy. Fig. 3 shows the accuracy rates in five tests corresponding to 6 over-sampling times. The accuracy rates fluctuate significantly between 92.91% and 76.50%. The best accuracy rate of 92.91% is achieved in Test 1 corresponding to outstanding discrimination, while the worse accuracy rate of 76.50% is achieved in Test 4 corresponding to acceptable discrimination. The Test 2, 3 and 5 have accuracy rates of 80.11%, 88.84% and 82.14%, respectively. They have an excellent discrimination. Thus, the reliable accuracy rate of the SMOTE-BPNN model must be based on the average value of many tests, and test datasets must be selected randomly.

Table 4 - Average AUC for varying over-sampling times

Over-sampling time	SMOTE
1 time	0.8246
2 times	0.8248
3 times	0.8251
4 times	0.8042
5 times	0.8005
6 times	0.841
7 times	0.8019
8 times	0.8064
9 times	0.8111
10 times	0.7955
11 times	0.8088
12 times	0.8046

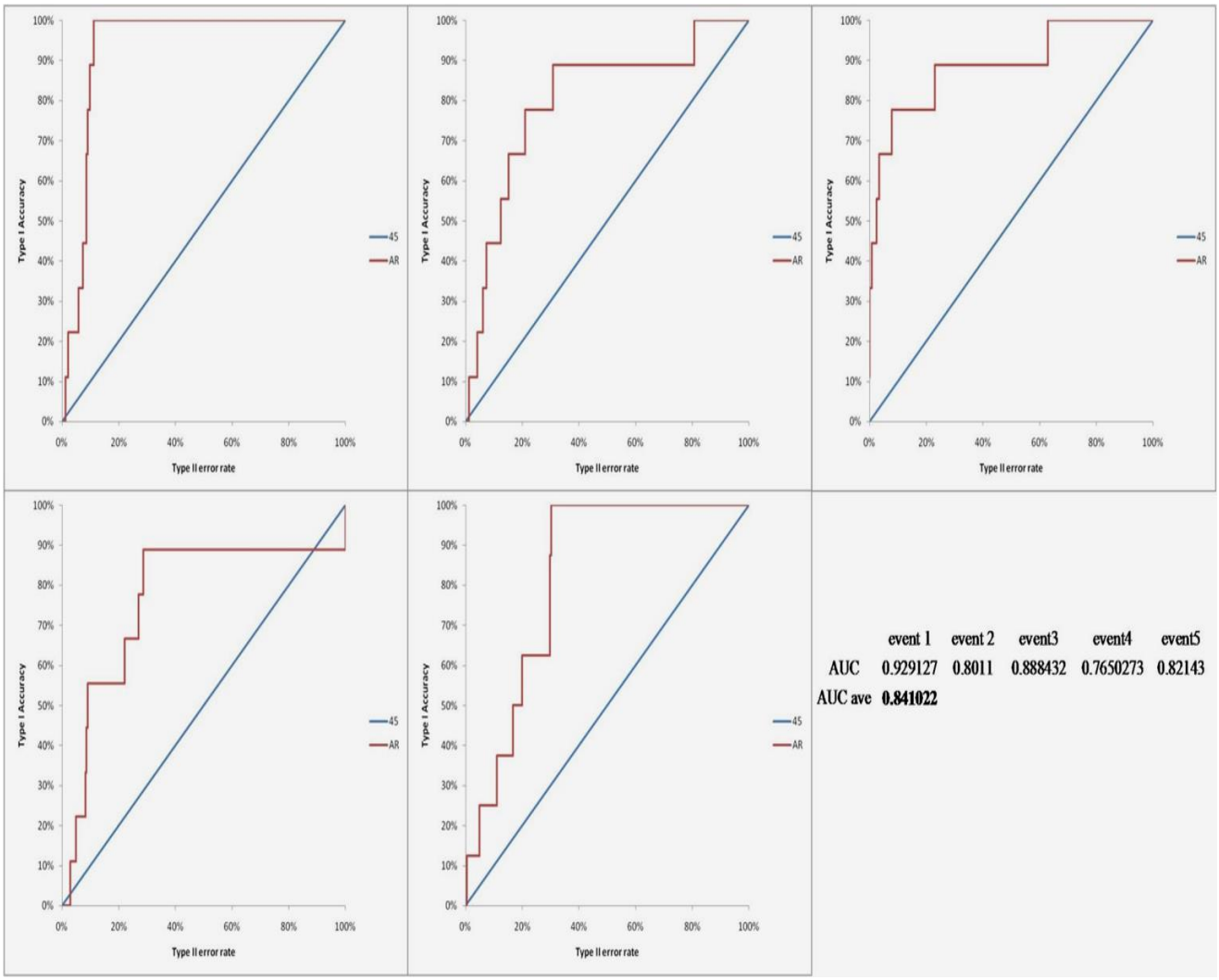

Fig. 3 - AUC values in 6 over-sampling times

Table 5 - Performance of the BPNN and SMOTE-BPNN models

	BPNN	SMOTE-BPNN
Average AUC	0.799	0.841

Table 5 displays the different accuracy rates between the BPNN and SMOTE-BPNN models. The SMOTE-BPNN model outperforms the BPNN model in terms of predictive performance. This result indicates that in construction company bankruptcy prediction, which is a highly complicated and non-linear problem, the BPNN using the SMOTE is able to grasp information within financial data and correctly analyze these data, resulting in more accurate prediction results. The BPNN model only achieves an accuracy rate of 79.9%, while the SMOTE-BPNN model improves and achieves accuracy rate of 84.1%. Thus, the between-class imbalance issue is resolved by the SMOTE. In addition, this result is better than those achieved by the most recent papers in the construction industry, such as those by Al-Sobiei, Arditi, & Polat (2005); Chen (2009); Tserng, et. al., (2011), and Huang & Tserng (2018).

6. Conclusions

The SMOTE-BPNN method is used to bankruptcy prediction for construction companies to improve the classification performance of the ANN. The method uses all of the available firm-year samples during the sample period and the SMOTE to resolve sample selection biases in the sample-matching method and between-class imbalance, which is a significant limitation for the BPNN method and ANNs.

Several conclusions can be shown from the results of this paper. First, the SMOTE-BPNN model has the best performance with major parameters, including 20 neurons of the hidden layer, a learning rate of 0.7, and 6 over-sampling

times. These results could help users cut a significant amount of unnecessary calculations. Second, due to the application of the SMOTE procedure, the SMOTE-BPNN model clearly outperforms the BPNN model in bankruptcy prediction. As a result, the accuracy rate is improved from 79.9% to 84.1%. Third, the results show that the SMOTE might resolve the between-class imbalance in construction company bankruptcy prediction models using the BPNN. Finally, this model could be used to help managers, investors, auditors, and the government in the US to predict construction company bankruptcy probability, and the approach can also be used as a reference for non-construction industries.

However, several problems remain to be addressed by further research. First, this research has not addressed many cases corresponding to changes in the amount of neurons in the hidden layer. The amount of hidden nodes is the factor, since too many hidden nodes can cause an overfitting problem. There are various rules of thumb were suggested to determine the optimal amount of hidden nodes. Second, further improvements in the SMOTE-BPNN model could be achieved by using techniques to input variable selection, such as decision-tree-based method, connection-weights-based method. Third, the research has not addressed the construction firm's characteristics, such as the size of assets, size of net worth, and type of work, all of which affect the results of the proposed model.

Acknowledgment

The authors would like to thank the Hanoi University of Civil Engineering for giving me the opportunity to conduct this research.

References

Abidali, A. F., & Harris, F. C. (1995). A methodology for predicting company failure in the construction industry, *Construction Management and Economics*, 13(3), 189-196

Agarwal, V., & Taffler, R. (2008). Comparing the performance of market-based and accounting-based bankruptcy prediction models, *Journal of Banking Finance*, 32(8), 1541-1551

Al-Sobiei, O. S., Arditi, D., & Polat, G. (2005). Managing owner's risk of contractor default, *Journal of Construction Engineering and Management*, 131(9), 973-978

Beaver, W. H. (1966). Financial ratios as predictors of failure, *Journal of Accounting Research*, 4, 71-111

Bishop, C. M. (1995), *Neural Networks for Pattern Recognition*. Oxford University Press

Boyacioglu, M. A., Kara, Y., & Baykan, Ö.K. (2009). Predicting bank financial failures using neural networks, support vector machines and multivariate statistical methods: a comparative analysis in the sample of savings deposit insurance fund (SDIF) transferred banks in Turkey, *Expert Systems with Applications*, 36(2), 3355-3366

Brockman, P., & Turtle, H. J. (2003). A barrier option framework for corporate security valuation, *Journal of Financial Econmics*, 67(3), 511-529

Chang, A. S., & Lee, K. P. (2011). Cost/schedule data needs and processes for construction ERP, *Journal of the Chinese Institute of Engineers*, 34(6), 721-731

Chang, F. J., Chang, L. C., & Wang, Y. S. (2007). Enforced self-organizing map neural networks for river flood forecasting, *Hydrological Processes*, 21(6), 741-749

Chava, S., & Jarrow, R. A. (2004). Bankruptcy prediction with industry effects, *Review of Finance*, 8(4), 537-569.

Chawla, N. V., Bowyer, K. W., Hall. L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique, *Journal of Artificial Intelligence Research*, 16, 341-378

Chen, H. J., Huang, S. Y., & Lin, C. S. (2009). Alternative diagnosis of corporate bankruptcy: a neuro fuzzy approach, *Expert Systems with Applications*, 36(4), 7710-7720

Chen, H. L. (2009). Model for predicting financial performance of development and construction corporations, *Journal of Construction Engineering and Management*, 135(11), 1190-1200

Chen, J. H. (2012). Developing SFNN models to predict financial distress of construction companies, *Expert Systems with Applications*, 39(1), 823-827

Chen, M. Y. (2011). Bankruptcy prediction in firms with statistical and intelligent techniques and a comparison of evolutionary computation approaches, *Computers and Mathematics with Applications*, 62(12), 4514-4524

Chen, T., & Lin, Y. C. (2009). A fuzzy back propagation network ensemble with example classification for lot output time prediction in a wafer fab, *Applied Soft Computing*, 9(2), 658-666

Deakin, E. B. (1976). Distributions of financial accounting ratios: some empirical evidence, *The Accounting Review*, 51(1), 90-96

Dichev, I. D. (1998). Is the risk of bankruptcy a systematic risk? *The Journal of Finance*, 53(3), 1131-1147

Estabrooks, A., Jo, T., & Japkowicz, N. (2004). A multiple resampling method for learning from imbalanced data sets, *Computational Intelligence*, 20(1), 18-36

Gao, M., Hong, X., Chen, S., & Harris, C. (2011). A combined SMOTE and PSO based RBF classifier for two-class imbalanced problems, *Neurocomputing*, 74(17), 3456-3466

Gharghori, P., Chan, H., & Faff, R. (2006). Investigating the performance of alternative default-risk models: option-based versus accounting-based approaches, *Australian Journal of Management*, 31(2), 207-234

Hillegeist, S. A., Keating, E. K., Cram, D. P., & Lundstedt, K. G. (2004). Assessing the probability of bankruptcy, *Review of Accounting Studies*, 9(1) 5-34

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression*: 2. John Wiley & Son, INC

Hua, Z., Wang, Y., Xu, X., Zhang, B., & Liang, L. (2007). Predicting corporate financial distress based on integration of support vector machine and logistic regression, *Expert Systems with Applications*, 33(2), 434-440

Huang, H. T., & Tserng, H. P. (2018). Study of integrating support-vector-machine (SVM) model and market-based model in predicting Taiwan construction contractor default, *KSCE Journal of Civil Engineering*, 22(12), 4750-4759

Jardin P. D. (2010). Predicting bankruptcy using neural networks and other classification methods: the influence of variable selection techniques on model accuracy, *Neurocomputing*, 73(10-12), 2047-2060

Jeong, C. W., Min, J. H., & Kim, M. H. (2012). A tuning method for the architecture of neural network models incorporating GAM and GA as applied to bankruptcy prediction, *Expert Systems with Applications*, 39(3), 3650-3658

Kangari, R., & Bakheet, M. (2001). Construction surety bonding, *Journal of Construction Engineering and Management*, 127(3), 232-238

Kangari, R., Farid, F., & Elgharib, H. M. (1992). Financial performance analysis for construction industry, *Journal of Construction Engineering Management*, 118(2), 349-361

Kim, H. S., & Sohn, S. Y. (2010). Support vector machines for default prediction of based on technology credit, *European Journal of Operational Research*, 201(3), 838-846

Kim, M. J., & Kang, D. K. (2010). Ensemble with neural networks for bankruptcy prediction, *Expert Systems with Applications*, 37(4), 3373-3379

Lee, K. C., Han, I., & Kwon, Y. (1996). Hybrid neural network models for bankruptcy prediction, *Decision Support Systems*, 18(1), 63-72

Lee, M. C., & To, C. (2010). Comparison of support vector machine and back propagation neural network in evaluating the enterprise financial distress, *International Journal of Artificial Intelligence & Applications*, 1(3), 31-43

Lin, T. H. (2009). A cross model study of corporate financial distress prediction in Taiwan: multiple discriminant analysis, logit, probit and neural networks models, *Neurocomputing*, 72(16-18), 3507-3516

Lin, T. K., Lin, C. C. J., & Chang, K.C. (2002). A neural network based methodology for estimating bridge damage after major earthquakes, *Journal of the Chinese Institute of Engineers*, 25(4), 415-424

Mason, R. J., & Harris, F. C. (1979). Predicting company failure in the construction industry. *Proceedings Institution of Civil Engineers*. pp. 301-307

Meyer, P. A., & Pifer, H. W. (1970). Prediction of bank failures, *The Journal of Finance*, 25(4), 853-868

Neves, J. C., & Vieira, A. (2006). Improving bankruptcy prediction with hidden layer learning vector quantization, *European Accounting Review*, 15(2), 253-271

Odom, M. D., & Sharda, R. (1990). A neural network model for bankruptcy prediction, *Proceedings of IJCNN International Joint Conference*, (12). pp. 163-168

Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy, *Journal of Accounting Research*, 18(1), 109-131

Pompe, P. P. M., & Feelders, A. J. (1997). Using machine learning, neural networks, and statistics to predict corporate bankruptcy, *Computer-Aided Civil and Infrastructure Engineering*, 12(4), 267-276

Reisz, A. S., & Perlich, C. (2007). A market-based framework for bankruptcy prediction, *Journal of Financial Stability*, 3(2), 85-131

Russell, J. S., & Jaselskis, E. J. (1992). Predicting construction contractor failure prior to contract award, *Journal of Construction Engineering and Management*, 118(4), 791-811

Russell, J. S., & Zhai, H. (1996). Predicting contractor failure using stochastic dynamics of economic and financial variables, *Journal of Construction Engineering and Management*, 122(2), 183-191

Severson, G. D., Jaselskis, E. J., & Russell, J. S. (1993). Trends in construction contractor financial data, *Journal of Construction Engineering and Management*, 119(4), 854-858

Severson, G. D., Russell, J. S., & Jaselskis, E. J. (1994). Predicting contract surety bond claims using contractor financial data, *Journal of Construction Engineering and Management*, 120(2), 405-420

Stein, R. M. (2007). Benchmarking default prediction models: pitfalls and remedies in model validation, *Journal of Risk Model Validation*, 1(1), 77-113

Taffler, R. J. (1982). Forecasting company failure in the UK using discriminant-analysis and financial ratio data, *Journal of the Royal Statistical Society, Series A (General)*, 14(3), 342-358

Theodossiou, P. S. (1993). Predicting shifts in the mean of a multivariate time series process: an application in predicting business failures, *The Journal of American Statistical Association*, 88(422), 441-449

Tserng, H. P., Liao, H. H., Tsai, L. K., & Chen, P. C. (2011). Predicting construction contractor default with option-based credit models-models' performance and comparison with financial ratio models, *Journal of Construction Engineering and Management*, 137(6), 412-420

Tserng, H. P., Lin, G. F., Tsai, L. K., & Chen, P. C. (2011). An enforced support vector machine model for construction contractor default prediction, *Automation in Construction*, 20(8), 1242-1249

Tserng, H. P., Ngo, T. L., Chen, P. C., & Tran, L. Q. (2015). A grey system theory-based default prediction model for construction firms, *Computer-Aided Civil and Infrastructure Engineering*, 30(2), 120-134

Weiss, G. M., & Provost, F. (2003). Learning when training data are costly: the effect of class distribution on tree induction, *Artificial Intelligence Research*, 19, 315-354

Wharton Research Data Services (2009). *Wharton School of the University of Pennsylvania, Philadelphia*. http://wrds.wharton.upenn.edu/Accessed 1 June 2009

Zmijewski, M. (1984). Methodological issues related to the estimation of financial distress prediction models, *Journal of Accounting Research*, 22, 59-82