

Cost-sensitive AdaBoost Selective Ensemble for Financial Distress Prediction

Hongbao Wang*

College of Applied English, Heilongjiang University, No. 74 Xuefu Road, Harbin,
Heilongjiang Province, China, 150080
Lodgewang@163.com

Abstract

Financial distress prediction (FDP) models are effective tools to prevent stakeholders from suffering economic loss. In the process of FDP, the misclassification cost of type I error of the model is much higher than that of type II error. Some FDP models based on single classifiers take the asymmetric costs into consideration, but the study on cost-sensitive ensemble approach for FDP is rarely explored. This paper constructs cost-sensitive AdaBoost selective ensemble FDP model for minimizing misclassification cost so that the loss of users of the model will suffer less. On the initial sample of 180 Chinese listed companies and 30 financial ratios, 8 times of holdout experiments are carried out for FDP respectively two years and three years in advance. The experimental results suggest that the proposed approach helps to reduce total misclassification costs compared with FDP model based on cost-sensitive C4.5 decision tree and that based on C4.5 decision tree.

Keywords: *Financial distress prediction (FDP), Cost-sensitive AdaBoost selective ensemble, Cost-sensitive C4.5 decision tree, Misclassification cost, Genetic algorithm*

1. Introduction

Facing a far more complex internal and external economic environment than before, many economic entities with management deficiency and lack of innovation suffer great damages or even become insolvent. Financial distress prediction (FDP) has a key influence on healthy development of enterprises, the interests of credit institutions and securities investors, even the economic security of a country [1]. For banks, FDP has profound impact on the credit ranking because they should pay close attention to the current and future financial status of their enterprise customers. Additionally, FDP helps shareholders to detect the financially distressed condition of a company in advance such that they will withdraw capital before suffering financial damage [2]. Therefore, the current economic background has lifted the importance of FDP to an unprecedented position.

Statistical techniques were firstly employed in FDP, such as single ratio analysis [3], multivariate discriminate analysis (MDA) [4] and logistic regression (Logit) [5]. After that, FDP entered a booming stage because of the rapid progress of artificial intelligent techniques [6]. Neural network (NN) has become one of the most widely used machine learning techniques in FDP because it outperforms traditional statistic techniques [7]. In addition, other artificial intelligent techniques were also used in FDP, such as decision tree [8], genetic algorithm [9], rough sets and etc. [10]. Developed on the basis of statistical learning theory, support vector machine (SVM) is a relatively new machine learning technique, which has been widely applied in many fields, such as classification, data mining and time series forecasting [11-13]. SVM is superior to other algorithms for FDP due to risk minimization, avoidance of local optimization and good generalization ability

Corresponding author: Hongbao Wang, [Tel:+86 13936501851](tel:+8613936501851)

and classification performance for relatively small sample [14-15].

The research above in FDP mainly employs single classifiers, whose performance depends too much on the characteristics of samples, and thus neglects the advantages of classifier ensemble approach. Recently, more research on classifier ensemble arose, due to their lower variance of estimated error and higher recognition performance [16-18].

However, most of the studies in FDP do not incorporate the misclassification costs of the two errors, namely type I error and type II. The former stands for classifying a financially distressed company as a healthy one, which results in the cost of losing principal and profit. While the latter stands for classifying a healthy company as a financially distressed one, which results in the cost of losing profit. Therefore, the asymmetric cost information should be taken into consideration in FDP so that different stakeholders, such as stockholders, creditors, auditors, clients, employees, and *etc.*, could select their favorable models based on their cost preference. Chen, Ribeiro, Vieira, Duarte and Neves [19] proposed cost-sensitive learning vector quantization approach in FDP. To our knowledge, few FDP models based on classifier ensemble have taken misclassification costs into consideration.

As a result, the main motivation of this paper is to employ cost-sensitive AdaBoost selective ensemble to construct FDP models in order to minimize misclassification costs. The main objectives of this paper are to (1) incorporate cost information into AdaBoost selective ensemble in FDP, (2) exclude the missing and outlier data in the initial data preprocessing stage and use statistical methods to screen financial ratios in order to improve the prediction accuracy of FDP model, (3) compare cost-sensitive AdaBoost selective ensemble approach with cost-sensitive C4.5 decision tree approach and C4.5 decision tree approach in the aspect of cost of misclassification and prediction performance, and (4) expand cost-sensitive ensemble approach so that it will provide stakeholders with evidence in model selection.

2. Methodology

2.1 Framework of Cost-sensitive AdaBoost Selective Ensemble Method for FDP

This paper improves the traditional AdaBoost ensemble, which incorporates costs in the two stages of FDP: construction of single classifiers and integration of ensemble's results. It proposes a business FDP model based on cost-sensitive AdaBoost selective ensemble, whose modeling process is shown in Figure 1. In the stage of construction of single classifier, the method of cost-sensitive weight adjustment is adopted to adjust the weights of each training data in order to make single classifier cost-sensitive. In the stage of integration of ensemble's results, genetic algorithm is adopted to screen the collection of single classifiers and then integrate the results. The two stages make misclassification costs minimum.

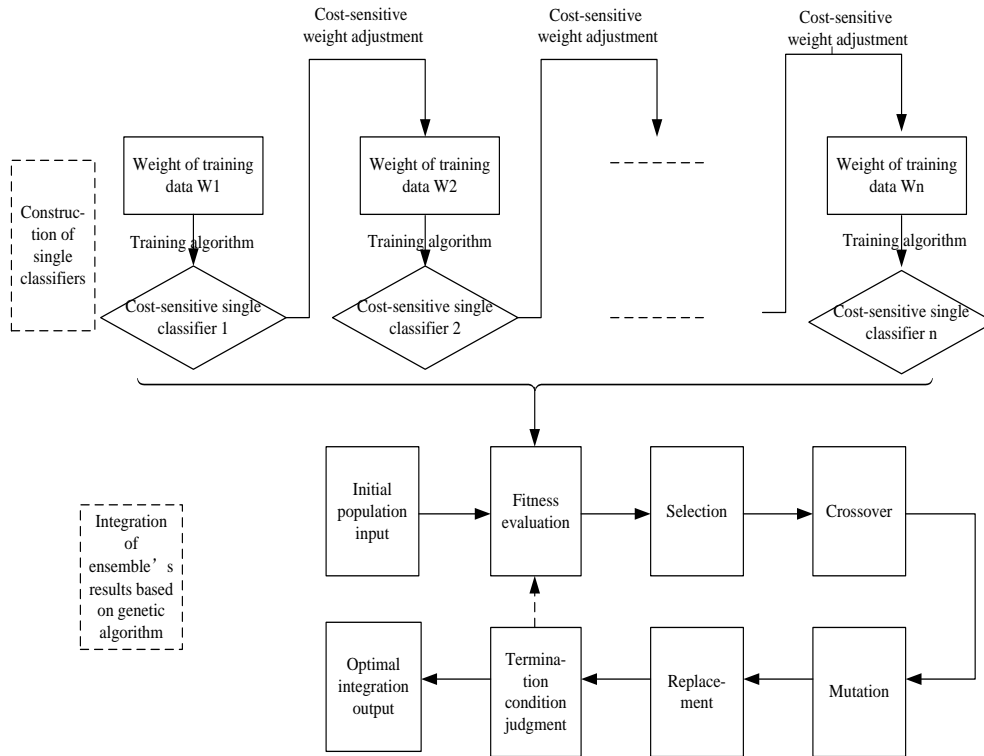


Figure 1. Business Financial Distress Prediction Modeling Process based on Cost-sensitive AdaBoost Selective Ensemble

2.2 Cost-sensitive AdaBoost Ensemble

AdaBoost is an adaptive boosting algorithm and most widely employed boosting algorithm [20]. AdaBoost sequentially trains single classifiers in order to construct a composite classifier. The core of AdaBoost algorithm is weight adjustment. In the each training round, the weights of the correctly classified data are decreased while the weights of the misclassified data are increased. After the rounds of training, the composite classifier is constructed by combining the single classifiers by weighted or simple voting schemes.

In order to make construction of single classifier cost-sensitive, the paper proposes a new weight adjustment method in the construction of single classifiers, which takes misclassification costs into consideration. The algorithm is shown as follows:

Algorithm name: Cost-sensitive AdaBoost ensemble

Input: $D = \{(x_1, y_1, c_1), (x_2, y_2, c_2), \dots, (x_n, y_n, c_n)\}$, where $c_i \in R^+$ and $y_i \in \{-1, 1\}$.

Output: $h_1(x), h_2(x), \dots, h_T(x)$

$$\text{Weights of initial data } w_1^i = c_i / \sum_{j=1}^n c_j \quad (1)$$

Do for $t=1, 2, \dots, T$, where t is training rounds, and T is the quantity of single classifiers. Single classifier is constructed by training methods of single classifier, training dataset D and weights:

$$h_t(x) : x \rightarrow \{-1, 1\} \quad (2)$$

$$\text{The error rate of } h_t(x) \text{ is calculated as: } \xi_t = \sum_{i=1}^n w_i^t [h_t(x_i) \neq y_i] \quad (3)$$

$$\text{Weights are adjusted as: } w_i^{t+1} = \frac{w_i^t}{Z_t} \begin{cases} e^{\frac{1}{2} \log(\frac{1-\xi_i}{\xi_i}) B(i)} & \text{if } h_t(x_i) \neq y_i \\ e^{-\frac{1}{2} \log(\frac{1-\xi_i}{\xi_i}) B(i)} & \text{if } h_t(x_i) = y_i \end{cases} \quad (4)$$

End for

Z_t in the formula (4) is normalization factor. $B(i) = B(\text{sign}(y_i h_t(x_i)), c_i)$ is a function of x_i and c_i , where x_i is classified data and c_i is misclassification cost. The function of $B(i)$ is used to make weight adjustment of data cost-sensitive. Therefore, when data with larger misclassification costs are misclassified, their weights are increased in magnitude. When data is correctly classified, their weights are decreased in magnitude.

2.3 Cost-sensitive AdaBoost Selective Ensemble

Classification efficiency of ensemble is determined by classification accuracy of single classifiers and discrepancy among single classifiers. Classification performance of ensemble becomes better when classification accuracy of single classifiers is higher. Classification efficiency of ensemble is higher when discrepancy among single classifiers is larger. Therefore, in the stage of integration of ensemble's results, the classification performance is better when single classifiers with higher classification accuracy and larger discrepancy participate in the stage rather than all the single classifiers.

The rationale for cost-sensitive AdaBoost selective ensemble is the same, which is shown as follows:

Suppose training dataset is $D = \{(x_1, y_1, c_1), (x_2, y_2, c_2), \dots, (x_n, y_n, c_n)\}$, $c_i \in R^+$ is the misclassification cost and $y_i \in \{-1, 1\}$. Cost-sensitive AdaBoost selective ensemble is adopted to train dataset D , which is $E(x) = \{f_1(x), f_2(x), \dots, f_m(x)\}$. When single classifier $f_i(x)$ participates in the stage, its weight is w_i . f_i denotes the actual output of single classifier i , $f_i = [f_{i1}, f_{i2}, \dots, f_{in}]^T$, where f_{ik} is the actual output of $f_i(x_k)$. Therefore, the average misclassification cost of single classifier $f_j(x)$ on training dataset D is:

$$Err_j = \frac{1}{n} \sum_{i=1}^n L(f_{ji}, y_i, c_i) \quad (5)$$

$L(\bullet)$ is a loss function. If $f_{ji} \neq y_i$, $L(\bullet)$ is c_i . If not, $L(\bullet)$ is 0. When cost-sensitive AdaBoost ensemble adopts weighted voting to integrate results, the average misclassification cost of $E(x)$ on training dataset D is:

$$Err_E = \frac{1}{n} \sum_{i=1}^n L(E_{ji}, y_i, c_i) \quad (6)$$

Where $E_{ji} = \arg \max_a (W_a)$, $W_a = \sum_{j=1}^m \{\prod (f_{ji} = a) w_i\}$ and $a \in \{1, 2, \dots, k\}$. $\prod(\bullet)$ is an exponential function. When the condition is false, the value is 0. When the condition is true, the value is 1.

If the single classifier t is deducted from cost-sensitive ensemble $E(x)$, the new cost-sensitive ensemble, $E(x)$, is obtained. Then the average misclassification cost of $E(x)$ on dataset D is:

$$Err_E = \frac{1}{n} \sum_{i=1}^n L(E_{ji}, y_i, c_i) \quad (7)$$

where $E_{ji} = \arg \max_a (W_a)$ and $W_a = \sum_{j=1}^m \{\mathbb{I}(f_{ji} = a)w_i - \mathbb{I}(f_{ji} = a)w_i\}$.

If $Err_E - Err_E = \frac{1}{n} \sum_{i=1}^n \{L(E_{ji}, y_i, c_i) - L(E_{ji}, y_i, c_i)\} > 0$, the deduction of the single classifier t leads to the fact that the average misclassification cost of the cost-sensitive ensemble on dataset D is effectively reduced.

2.4 Cost-sensitive Selective Ensemble based on Genetic Algorithm

Section 2.3 provides a selection method of single classifier subset in ensemble. However, the efficiency of the method is rather low. Especially when cost-sensitive ensemble contains many single classifiers, it is quite impossible to find an optimal single classifier subset. Therefore, genetic algorithm is adopted to effectively select single classifier subset.

Genetic algorithm has been an effective way to solve optimization problems. Chromosome coding, fitness function and genetic operator are conducted in a specific way in the optimization problem. Cost-sensitive selective ensemble based on genetic algorithm is shown as follows:

2.4.1 Chromosome Coding: Suppose cost-sensitive ensemble contains M single classifiers, each single classifier collection, named as individual in genetic algorithm, corresponds with M binary digit. When single classifier is selected, its binary digit is coded as 1. If not, its binary digit is coded as 0.

2.4.2 Fitness Function: Fitness function is to measure excellence degree of each individual in a group. Analogous to the rule of survival of the fittest, the individual in higher excellence degree is highly possible to be transmitted to future generation, and vice versa. Suppose the average misclassification cost of single classifiers is e in this paper, $\text{Fitness} = 1 - e$.

2.4.3 Genetic Operators: In genetic algorithm, a new population is generated from current population by genetic operators, which is composed of reproduction, crossover and mutation. This paper adopts roulette wheel to select individuals. The two matching individuals are exchanged for their corresponding binary-encoding strings in order of certain probability. And they become two new individuals. In individual mutation, binary bit selects 1 bit or several bits. The value becomes 1 if the value of home position is 0. And the value becomes 0 if the value of home position is 1. Then a new individual is generated.

In order to keep excellent chromosomes, the individual ranked as the first one in fitness function is prohibited to participate in the crossover and mutation, which effectively avoids degeneration in the genetic algorithm.

3. Empirical Experiment

3.1 Experiment Data

3.1.1 Initial Data Collection and Preprocessing: Financially distressed companies refers to those who have had negative net profit in consecutive two years, or their net capital per share is lower than the face value per share for the reason of one year's substantive loss. Thus, healthy companies are selected from those that have never been specially treated (ST).

Since the financial ratios of ST companies have been deteriorating even before ST, the

adoption of financial data one year before ST leads to overestimation of the prediction performance of the model. In this paper, the financial data from two years to three years before ST is selected, namely $U_{(t-2)}$ and $U_{(t-3)}$. The experimental data are collected from RESSET Financial Database. The initial samples consist of totally 180 companies listed in Shenzhen Stock Exchange and Shanghai Stock Exchange. 30 financial ratios are selected as initial features, covering debt ability, growth ability, capital structure, activity ability, profitability and indicators per share. In order to eliminate outlier data and missing data, companies with financial ratios deviating from the mean value as much as four times of standard deviation are discarded and companies missing at least one financial ratio are also discarded. The final number of sample data is 170.

3.1.2 Experimental Data Sets: The empirical experiment aims to validate whether FDP model based on cost-sensitive AdaBoost selective ensemble can minimize the cost of prediction error. 60 pairs of financially distressed companies and healthy companies are selected to form training data set and the rest 25 pairs are used to form testing data set.

3.1.3 Feature Selection: In FDP, many financial indicators are collected in order to describe an accurate financial condition of companies. However, some financial indicators cannot precisely distinguish financially distressed companies from healthy ones. Therefore, feature selection aims to address the problem by removing irrelevant and redundant features, improving the accuracy of the model, decreasing the computational effort and facilitating the use of the model.

3.1.3.1 Statistical Analysis: In this paper, statistical techniques are employed to screen financial ratios. Firstly, Kolmogrov-Simironov test is used to examine whether each financial ratio meets normal distribution. Then, if financial ratios meet normal distribution, T test is used to validate whether the financial ratios are significant. If financial ratios do not meet normal distribution, Mann-Whitney test is used to validate whether the financial ratios are significant, as indicated in Table 1.

Table 1. Significance Test of Financial Ratios in Year t-3

Variables	T test		Kolmogrov-Simironov test		Mann-Whitney test	
	T-Statistic	Prob.	KS-Statistic	Prob.	MW-Statistic	Prob.
Return on Equity			2.097	**0.000	1296.000	**0.000
Return on Assets			2.199	**0.000	1871.000	**0.000
Return on Invested Capital			3.874	**0.000	1548.000	**0.000
Net Profit Margin			4.098	**0.000	1890.000	**0.000
Cost Profit Margin			3.257	**0.000	1841.000	**0.000
Current Ratio			2.727	**0.000	1196.000	**0.000
Quick Ratio			2.932	**0.000	1369.000	**0.000
Equity Ratio			3.807	**0.000	1314.000	**0.000
Debt to Asset Ratio			4.654	**0.000	1323.000	**0.000
Debt to Tangible Asset Ratio			4.911	**0.000	3554.000	0.855
Operating Cash Flow/Total Liability			1.896	**0.002	2509.000	**0.001
Operating Income Growth Rate			2.450	**0.000	2631.000	*0.000
Net Profit Growth Rate			4.820	**0.000	2490.000	**0.000

Total Asset Growth Rate			4.562	**0.000	1698.000	**0.000
Turnover Rate of Accounts Receivable			5.566	**0.000	3213.000	0.213
Turnover Rate of Accounts Payable			2.962	**0.000	2603.000	**0.002
Turnover Rate of Current Assets			1.673	**0.007	3155.500	0.154
Turnover Rate of Fixed Assets			6.202	**0.000	1964.000	**0.000
Turnover Rate of Equity			2.819	**0.000	3196.000	0.194
Turnover Rate of Total Assets			1.834	**0.002	2355.500	**0.000
Earnings Per Share			1.859	**0.002	1684.500	**0.000
Net Asset Value Per Share	5.798	**0.000	1.342	0.054		
Operating Revenue Per Share			2.240	**0.000	2494.000	**0.000
Gross Profit Margin			1.764	**0.004	2419.000	**0.000
Net Return on Assets			3.326	**0.000	1736.000	**0.000
Fixed Assets Ratio	-3.178	**0.002	1.093	0.184		
Equity Ratio			4.654	**0.000	1323.000	**0.000
Operating Profit Growth			5.092	**0.000	2884.000	*0.023
Earning Per Share Growth			3.924	**0.000	2515.000	**0.001
Every Dividend Profit before Tax			2.132	**0.000	1814.000	**0.000

Note: *Significant at 5%; **Significant at 1%.

3.1.3.2 Analysis on Significance Test of Financial Ratios: As shown in Table 1, only two variables pass Kolmogrov-Simironov test, which is consistent with the previous research conclusion that most financial ratios do not meet normal distribution. Additionally, debt to tangible asset ratio, operating cash flow/total liability, net profit growth rate, total asset growth rate, turnover rate of accounts receivable, turnover rate of current assets, turnover rate of equity, operating revenue per share, gross profit margin, net return on assets, fixed assets ratio, equity ratio, operating profit growth, earning per share growth and every dividend profit before tax do not pass the significance test in year t-3. Therefore, these 15 ratios are excluded, by which healthy companies cannot be identified from distressed companies.

3.2 Experimental Design and Parameter Setting

3.2.1 Setting of Cost of Misclassification: In the empirical research, we suppose that the user of FDP model is banks, which can use the prediction result of FDP model to determine whether to make loans to enterprises in order to prevent bad debts. Therefore, when the model has type II error, the cost of error of banks is the loss of loan interest. When the model has type I error, the cost of error of banks is the loss of a complete loan. Since the loan interests of commercial banks to a company in China range from 5% to 30% above personal loan interest rate, 30% above personal loan interest rate is taken into calculation for the sake of unified computing. The cost matrix is shown in the Table 2.

Table 2. Cost Matrix for FDP

	Actual healthy firms	Actual distressed firms
Predict healthy firms	0.00	1.00
Predict distressed firms	0.09	0.00

3.2.2 Setting of Data Weight Adjustment Formula in the Construction of Single classifiers: The weight adjustment formula is determined by the function $B(i)$, which is the function of classified data and misclassification cost, $\{sign(y_i h_i(x_i)), c_i\}$. Therefore, $B(i)$ is set up as follows:

$$B(i) = \begin{cases} (2 - c_i)/2 & \text{if } sign(y_i h_i(x_i)) = 1 \\ 1 + c_i & \text{if } sign(y_i h_i(x_i)) = -1 \end{cases} \quad (9)$$

where $0 \leq c_i \leq 1$ to all the enterprises. When the classified result of $h_i(x_i)$ is incorrect, $B(i) \geq 1$. If the misclassification cost is larger, $B(i)$ is larger.

3.2.3 Setting of the Type of Single Classifiers and Training Rounds of Cost-sensitive Single Classifiers: With the development of artificial intelligence techniques, many classification techniques have been widely employed in FDP, such as neural network, support vector machine, decision tree and *etc.* However, most of them can only be employed in continuous data, not discrete data. Therefore, the paper sets up C4.5 decision tree as single classifiers, which can be employed in both continuous data and discrete data.

In order to determine the training rounds of cost-sensitive C4.5 decision tree and quantity of single classifiers of cost-sensitive ensemble, dataset $U_{(t-2)}$ and $U_{(t-3)}$ are randomly divided into two disjoint subsets, one of which contains 60 pairs of healthy and unhealthy firms as training dataset. The training rounds are set up as $T=200$ and the fitness function 1-e and evolution curve of average fitness are shown as follows in Figure 2:

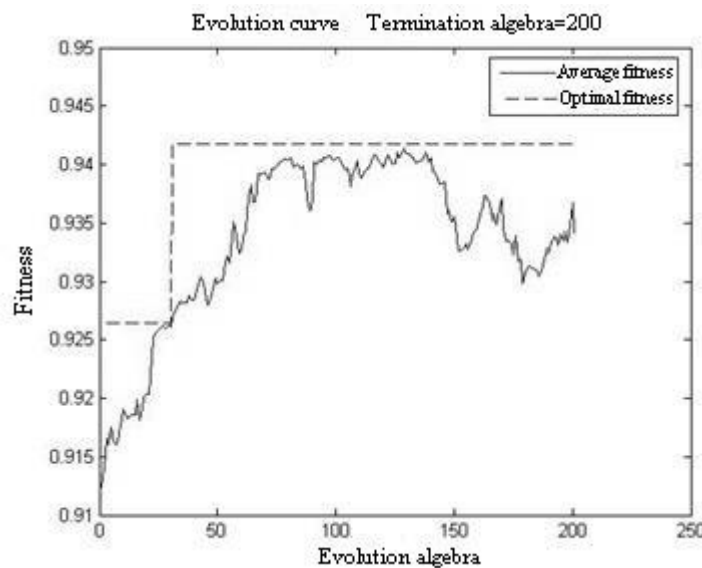


Figure 2. Average Fitness of Dataset $U_{(t-3)}$ and Evolution Curve of the Optimal Fitness

3.2.4 Setting of Genetic Algorithm Parameters: According to the experience of genetic algorithm, operation parameters are set up as follows: population size is 100, length of chromosome is 100, probabilities of crossover and mutation are 0.3 and 0.1, and termination algebra is 200.

3.3 Experiment Results and Analysis

To evaluate the prediction performance of FDP model based on cost-sensitive AdaBoost selective ensemble, the empirical research makes a comparison with the one based on cost-sensitive C 4.5 decision tree and the one based on C4.5 decision tree. Dataset $U_{(t-2)}$ and $U_{(t-3)}$ are formed by repetitively and randomly classifying training sample and testing sample in order to get multiple performance statistics. By 8 times of random sampling without replacement, 60 pairs of financial distressed companies and healthy ones are selected as training data set and the rest 25 pairs are selected as testing data set each time, as shown in Table 3.

Table 3. Experimental Results on Testing Data Set in Year t-3

Data	Cost-sensitive AdaBoost selective ensemble				Cost-sensitive C4.5			C4.5	
	Number of type I error	Number of type II error	Total cost of prediction error	Number of type I error	Number of type II error	Total cost of prediction error	Number of type I error	Number of type II error	Total cost of prediction error
set 1	2	9	2.81	1	9	1.81	9	9	9.81
set 2	1	11	1.99	1	10	1.90	8	8	8.72
set 3	2	7	2.63	4	16	5.44	6	6	6.54
set 4	5	5	5.45	4	14	5.26	12	5	12.45
set 5	3	5	3.45	3	17	4.53	7	4	7.36
set 6	3	4	3.36	1	15	2.35	5	5	5.45
set 7	1	6	1.54	4	9	4.81	10	4	10.36
set 8	7	9	7.81	6	14	7.26	7	7	7.63
Average	3	7	3.63	3	13	4.17	8	6	8.54

As shown in Table 3, the average cost of prediction error of FDP model based on cost-sensitive AdaBoost selective ensemble is 3.63, which is markedly lower than that of FDP model based on single classifier, 8.54. It is also lower than that of FDP model based on cost-sensitive C4.5 decision tree, 4.17. It is mainly because FDP model based on cost-sensitive AdaBoost selective ensemble and the one based on cost-sensitive C4.5 decision tree have less type I errors than the one based on C4.5 decision tree. Even though FDP model based on cost-sensitive AdaBoost selective ensemble has the same type I errors with the one based on cost-sensitive C4.5 decision tree, the former has less type II errors than the latter.

Table 4. Experimental Result Comparison of FDP Models

Dataset	FDP models based on cost-sensitive AdaBoost selective ensemble			FDP models based on cost-sensitive C4.5 decision tree			FDP models based on C4.5 decision tree		
	Type I error rate	Type II error rate	Misclassification cost rate	Type I error rate	Type II error rate	Misclassification cost rate	Type I error rate	Type II error rate	Misclassification cost rate

$U_{(t-2)}$	8%	28%	9.7%	8%	48%	11.3%	28%	8%	26.3%
$U_{(t-3)}$	12%	28%	13.3%	12%	52%	15.3%	32%	24%	31.3%

In Table 4, the empirical results suggest that the misclassification cost rate of all the three models become larger with the selection of earlier testing datasets. It also demonstrates that the prediction performance of FDP models become weaker with the selection of earlier testing datasets.

The experimental results and analysis suggest that FDP models based on cost-sensitive AdaBoost selective ensemble achieves the least misclassification cost rate than that based on cost-sensitive C4.5 decision tree and that based on C4.5 decision tree. With the selection of earlier training data sets, all three models become larger in the misclassification cost, but they do not perform the same way in the prediction performance.

4. Conclusion

Financial distress prediction is extensively studied in the corporate governance field. Few studies incorporate unequal misclassification costs into FDP models based on classifier ensembles. Cost-sensitive classification models, coping with asymmetric costs of type I error and type II error, are of crucial interest to stakeholders' decisions. This paper verifies how the asymmetric costs of two kinds of errors are integrated into FDP models based on cost-sensitive AdaBoost selective ensemble, those based on cost-sensitive single classifier and those based on single classifier. This research takes 85 financial healthy companies and matches them with 85 financially distressed companies. 120 companies are selected as training datasets and the rest 50 are selected as testing datasets. Experimental tests demonstrate that cost-sensitive AdaBoost selective ensemble approach leads to a lower total misclassification cost when compared with cost-sensitive C4.5 decision tree and C4.5 decision tree one.

Acknowledgements

This research is partially supported by Project of Humanities and Social Sciences of Heilongjiang Provincial Department of Education (No. 1254b012).

References

- [1] Z. Xiao, X. L. Yang, Y. Pang and X. Dang, "The Prediction for Listed Companies' Financial Distress by Using Multiple Prediction Methods with Rough Set and Dempster-Shafer Evidence Theory", *Knowledge-Based Systems*, vol. 26, (2012), pp. 196-206.
- [2] P. C. Ko and P. C. Lin, "An Evolution-based Approach with Modularized Evaluations to Forecast Financial Distress", *Knowledge-Based Systems*, vol. 19, no. 1, (2006), 84-91.
- [3] W. Beaver, "Alternative Accounting Measures as Predictors of Failure", *The Accounting Review*, vol. 1, no. 1, (1968), pp. 113-122.
- [4] E. I. Altman, "Financial Ratios Discriminant Analysis and the Prediction of Corporate Bankruptcy", *Journal of Finance*, vol. 23, no. 4, (1968), pp. 589-609.
- [5] J. A. Ohlson, "Financial Ratios and the Probabilistic Prediction of Bankruptcy", *Journal of Accounting Research*, Spring, vol. 19, (1980), pp. 109-131.
- [6] A. Chaudhuri and K. De, "Fuzzy Support Vector Machine for Bankruptcy Prediction", *Applied Soft Computing*, vol. 11, (2011), pp. 2472-2486.
- [7] M. Odom and R. Sharda, "A Neural Networks Model for Bankruptcy Prediction", *Proceedings of the IEEE International Conference on Neural Network*, vol. 2, (1990), pp. 163-168.
- [8] H. Frydman, E.I. Altman and D. Kao, "Introducing Recursive Partitioning for Financial Classification: the Case of Financial Distress", *Journal of Finance*, vol. 40, no. 1, (1985), pp. 269-291.
- [9] T. Lensberg, A. Eilifsen and T. E. McKee, "Bankruptcy Theory Development and Classification via Genetic Programming", *European Journal of Operational Research*, vol. 169, no. 2, (2006), pp. 677-697.

- [10] T. E. McKee, “Developing a Bankruptcy Prediction Model via Rough Sets Theory”, *International Journal of Intelligent Systems in Accounting, Finance and Management*, vol. 9, (2000), pp. 159-173.
- [11] S. Wang, “Support Vector Machines Classification for High-dimensional Dataset”, *Fourth International Conference on Multimedia Information Networking and Security*, (2012), pp. 315-318.
- [12] Q. Ye, Z. Han, J. Jiao and J. Liu, “Human Detection in Images via Piecewise Linear Support Vector Machines”, *IEEE Transactions on Image Processing*, vol. 22, no. 2, (2013), pp. 778-789.
- [13] E. Ceperic, V. Ceperic and A. Baric, “A Strategy for Short-term Load Forecasting by Support Vector Regression Machines”, *IEEE Transactions on Power Systems*, vol. 28, no. 4, (2013), pp. 4356- 4364.
- [14] Y. Ding, X. Song and Y. Zen, “Forecasting Financial Condition of Chinese Listed Companies based on Support Vector Machine”, *Expert Systems with Applications*, vol. 34, (2008), pp. 3081–3089.
- [15] M. Boyacioglu, Y. Kara and O. Baykan, “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 Transferred Banks in Turkey*, *Expert Systems with Applications*, vol. 36, (2009), pp. 3355–3366.
- [16] 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, no. 4, (2007), pp. 1–28.
- [17] L. I. Kuncheva, “Diversity in Multiple Classifier Systems”, *Information Fusion*, vol. 6, (2005), pp. 3–4.
- [18] D. Ruta and B. Gabrys, “Classifier Selection for Majority Voting”, *Information Fusion*, vol. 6, (2005), pp. 63–81.
- [19] N. Chen, B. Ribeiro, A. S. Vieira, J. Duarte and J. C. Neves, “Hybrid Genetic Algorithm and Learning Vector Quantization Modeling for Cost-sensitive Bankruptcy Prediction”, *2nd International Conference on Machine Learning and Computing*, (2010), pp. 213-217.
- [20] Y. Freund, “Boosting a Weak Learning Algorithm by Majority”, *Information and Computation*, vol. 121, no. 2, (1995), pp. 256-285.

Author



Hongbao Wang, Ph.D., Associate Professor, College of Applied Foreign Language, Heilongjiang University, Harbin, China
Scientific activities: data mining, financial distress prediction, business English

