TI Journals
International Journal of Economy, Management and Social Sciences
www.tijournals.com
ISSN 2306-7276

A Comparison between Data Mining Methods in Anticipation of Financial Distress and Improvement of Anticipation through Combination of Methods

Adele Amini Salehi *[1], Mahmuod Muosavi Shiri [2], Hoda Majbuori Yazdi [3], Mahdi Filsaraei [4]

[1,3] *Department of Accounting, Mashhad Branch, Islamic Azad University, Mashhad, Iran.*
[2] *Department of Management, Economics and Accounting, Payame Noor University, Mashhad, Iran.*
[4] *Department of Accounting, Bojnourd Branch, Islamic Azad University, Bojnourd, Iran.*

ARTICLE INFO

Keywords:
Financial distress
Artificial Neural Networks
Support Vector Machine
Decision Tree
Majority voting combination

ABSTRACT

Nowadays, many studies have been conducted on corporate financial distress anticipation using data mining techniques. Artificial Neural Networks, Support Vector Machine and Decision Tree Algorithms are three current methods for data mining to anticipate corporate financial distress. This study compares the anticipation accuracy of these three methods in anticipation of corporate financial distress. The effect of combining these three methods is also studied through relative majority voting method to improve anticipation of corporate financial distress.

Statistical population of this study includes 100 sound companies and 100 distressed companies, active in Tehran Stock Exchange Market between 2005 and 2011, which were studied for the two years of "t" and "t-1". Findings of the study show that decision tree with 95.95% accuracy for the year "t" and artificial neural network with 89.92% accuracy for the year "t-1" have the highest level of efficiency to anticipate corporate financial distress. The results also show that combination of relative majority voting with 93.89% accuracy in the year "t" and 89.89% accuracy in the yea "t-1" is able to anticipate corporate financial distress.

1. Introduction

Every day, we are witnessing multiple crises in the business world. Ever-increasing competition between economic institutions has restricted accessibility to resources and increased probability of financial distress. In consequence of the financial crisis it is always too late for many creditors to withdraw their loans, as well as for investors to sell their own stocks, futures, or options [3].

This makes capital holders and other stakeholders worried, and excites them to seek some methods to anticipate financial crises in time, in order to prevent loss of their principal investment and relevant interest [7].

Financial failure occurs when a firm suffers chronic and serious losses or when the firm becomes insolvent with liabilities that are disproportionate to its assets [9]. Common causes and symptoms of financial failure include lack of financial knowledge, failure to set capital plans, poor debt management, inadequate protection against unforeseen events and difficulties in adhering to proper operating discipline in the financial market [4].

In-time anticipation of financial distress is one of the methods which can help appropriate benefiting from investment opportunities and better allocation of resources. Thus, it will be possible at first to make the company conscious about occurrence of bankruptcy by giving necessary warnings. Secondly, investors and credit-holders can invest their resources in suitable opportunities by means of these models of anticipation [14].

2. Literature review

2.1 Decision tree

Decision trees are powerful classification algorithms that are becoming increasingly more popular due to their intuitive explain ability characteristics [13]. Popular decision tree algorithms include Quinlan's ID3, C4.5, C5, and Bremen et al.'s CART (Classification and Regression Trees), Best First Decision Tree and AD Decision Tree. In this study we used CART decision tree algorithm.

2.2 Support vector machine

Support Vector Machine (SVM) belongs to a family of generalized linear models which achieves a classification or regression decision based on the value of the linear combination of features. The mapping function in SVM can be either a classification function (used to categorize

* Corresponding author.
Email address: adele_salehi@yahoo.com

International Journal of Economy, Management and Social Sciences, 2(9) September 2013

the data, as is the case in this study) or a regression function (used to estimate the numerical value of the desired output). For classification, nonlinear kernel functions are often used to transform the input data (inherently representing highly complex nonlinear relationships) to a high dimensional feature space in which the input data becomes more separable (i.e., linearly separable) compared to the original input space. Then, the maximum-margin hyperplanes are constructed to optimally separate the classes in the training data. Two parallel hyperplanes are constructed on each side of the hyperplane that separates the data by maximizing the distance between the two parallel hyperplanes. An assumption is made that the larger the margin or distance between these parallel hyperplanes the lower the generalization error of the classifier will be [13].

2.3 Artificial neural networks

Artificial neural networks (ANN) are biologically inspired analytical techniques, capable of modeling extremely complex non-linear functions [8]. In this study we used two popular neural network architectures, multi-layer perceptron (MLP) with a back-propagation. This supervised learning algorithm is strong function approximators for prediction as well as classification type prediction problems.

2.4 The method of majority voting for combination

Multiple classifier combination is a technique to combine decisions of different classifiers to generate a joint decision. All single classifiers are trained to solve the same problem, which is financial distress prediction in this research. There have been several combining schemes on outputs of individual classifiers, among which the family of majority voting is a most widely used method [12].

The family of majority voting methods is a simple yet effective method of combination. It chooses the class label which is supported by the majority of multiple classifiers.

In the method of majority voting, each classifier has one vote that can be casted for any class, and the class which gets major vote is selected finally [16]. Pure majority voting methods treat classifiers equally in the combination for prediction. It is believed sometimes that classifiers with high predictive performance should be given more rights in combination approach. By considering their differences in predictive performance, weighted majority voting methods put classifiers different weights according to their performance.

3. Research background

Numerous studies have been conducted on financial distress in different areas some of which are being reviewed as follows:
David Olson et al compared the accuracy of four methods of neural networks, logit regression, support vector machine, and decision tree algorithms. In this study, 100 bankrupted American companies between 2006 and 2009 were considered and 19 financial ratios were used. The results show that decision tree algorithm has the best efficiency in anticipation of bankruptcy with 94.8% accuracy [13].

Mu-Yen Chen compared neural network method, dynamic numerical technique and decision tree algorithm in anticipation of financial distress, and concluded that decision tree algorithms have had the best performance in anticipation of imminent bankruptcies [2].

In a study, Ji San and Hiu Li compared multiple discriminant analysis methods, logit regression, artificial neural network, decision tree, support vector machine and a combination of these five methods. Results of this study showed that the combination of five methods with 85.14% accuracy is capable to perform a higher anticipation than each of the classifiers [15].

By means of support vector machine, Kiang et al suggested a model to anticipate bankruptcy and compared the model with the function of artificial neural networks. Results of this study showed that support vector machine has a better performance than artificial neural network in both generalizability and accuracy [10].

Cielen et al performed a study to anticipate bankruptcy by three patterns of minimum sum of deviation, data envelopment analysis and decision tree. Their studies showed that these three methods provide 78.9%, 86.4% and 85.5% accuracy of classification accordingly [5].

Mohammad Esmail Fadaei-Nejad and Rasool Eskandari compared error propagation models, genetics algorithm and cumulative optimization of particles in a study and also compared the effects of market data and financial ratios in the accuracy of model anticipation. The results show that using genetics algorithm is more effective in anticipation of bankruptcy than error propagation model. It is also specified that referring to market data to anticipate bankruptcy is more effective than financial data [6].

In a study, Mohammadreza Nikbakht and Maryam Sharifi compared accuracy of neural networks and multiple discriminant analysis in anticipation of bankruptcy. Results show that artificial neural networks (ANN) with 98.3% accuracy provide better anticipation of bankruptcy in comparison with multiple- diagnostic analysis with 93.1% accuracy [11].

4. Methodology

This research is an applicable study. Stakeholders of some organizations such as banks, economic analysts, shareholders and credit-holders are users of the results of this study. Historical data of sound and distressed companies are used in this study. Statistical population includes all companies listed in Tehran Stock Exchange Market. Article 141 of Iranian Law of Commerce is considered as a criterion to classify companies into two sound and distressed groups.

Data is divided into two training and testing parts among which, training data is used to create the model and testing data is used to assess the accuracy of created models.
In splitting the data into training and testing dataset one can choose to make a single split (e.g., half of the data for training and other half of the data for testing) or multiple splits, which is commonly referred to as k-fold cross validation. The idea behind k-fold cross validation is to minimize the bias associated with the random sampling of the training and holdout data samples. Specifically, in k-fold cross validation the complete data set is randomly split into k mutually exclusive subsets of approximately equal size. Each prediction model is trained and tested k times using exactly the same k data sets (i.e., folds). Each time, the model is trained on all but one folds and tested on the remaining single fold. The cross validation estimate of the overall accuracy of a model is calculated by averaging the k individual accuracy measures as shown in the following equation

$$OA = \frac{1}{K}\sum_{i=1}^{k} A_i$$

Where OA stands for overall cross validation accuracy, k is the number of folds used, and A is the accuracy measure of each fold in this study, to estimate the performance of predictors a stratified 10-fold cross validation approach is used. Empirical studies showed that 10 seem to be an "optimal" number of folds (that balances the time it takes to complete the test and the bias and variance associated with the validation process [1].

5. Statistical population and sampling

Statistical Population:
Statistical population includes all companies listed in Tehran Stock Exchange Market.

Statistical Sample Size:
To select statistical sample size, paired sampling method is used as follows:
Selection of financially distressed companies: A total of 100 companies, included in article 141 of Iranian Law of Commerce (criterion for financial distress in this study) between 2005 and 2011, were selected as financially distressed companies. So, a sound company was selected too in lieu of each distressed company. It is attempted to make conformity between two groups of company from the perspective of size, while selecting sound companies. Total value of assets has been the criterion for company sizes; thus, 100 successful companies were selected too.

Data related to theoretical bases, background of the study and literature are collected by desk study method including review of papers, theses and magazines about financial studies, some of which are obtained by browsing the web and referring to scientific and academic centers. Required data for consideration and test of hypotheses is extracted by means of financial statements of listed companies in Tehran Stock Exchange Market and Rahavard Novin Software.

6. Research Variables

Financial distress of companies is considered as a dependent variable in this study.

Financial ratios are independent variables in this study. To define independent variables, those financial ratios, having been more referred in previous studies, are considered as primary variables in following table:

Table 1. List of selected independent variables

No	Independent variables	No	Independent variables
1	Working capital to equity	16	Sales to fixed assets
2	Current liabilities to total assets	17	Sales to working capital
3	Current liabilities to total equity	18	Sales to equity
4	Long-term liabilities to equity	19	Working capital to total assets
5	Long-term liabilities to total assets	20	Current liabilities to total liabilities
6	Sales to total assets	21	Net profit to working capital
7	Total liabilities to total assets	22	Net profit to total liabilities
8	Total liabilities to equity	23	Net profit to sales
9	Current assets to sales	24	Net profit to assets
10	Current assets to current liabilities	25	Net profit to equity
11	Current assets to total assets	26	EBIT[1] to total assets
12	Fixed assets to total assets	27	EBIT to total liabilities
13	Equity to total liabilities	28	EBIT to financial expenses
14	Equity to total assets	29	EBIT to fixed assets
15	Long-term liabilities to sum of long-term liabilities and equity	30	EBIT to sales

[1] Earnings before Interest and Tax

International Journal of Economy, Management and Social Sciences, 2(9) September 2013

After selection of primary ratios, it is necessary to eliminate those variables lacking high functionality to separate distressed and sound companies. We use step-by-step diagnostic analytical method to select effective variables. Following tables represent test results for "t" and "t-1" years

Table 2. Diagnostic Analysis in Step-by-Step method – "t" Year

Step	Variables in the Analysis	Tolerance	F to Remove	Wilks' Lambda
1	Net profit to total liabilities	1.000	175.090	
2	Net profit to total liabilities	.475	18.420	.540
	Equity to total liabilities	.475	10.482	.519
3	Net profit to total liabilities	.408	26.610	.534
	Equity to total liabilities	.453	14.269	.503
	EBIT to total liabilities	.581	9.567	.492
4	Net profit to total liabilities	.405	27.708	.520
	Equity to total liabilities	.446	16.095	.492
	EBIT to total liabilities	.406	15.865	.491
	EBIT to total assets	.660	6.244	.468
5	Net profit to total liabilities	.403	24.444	.498
	Equity to total liabilities	.444	16.808	.480
	EBIT to total liabilities	.403	16.882	.480
	EBIT to total assets	.660	5.712	.453
	Sales to total assets	.967	5.343	.453

Results of above table indicate that the ratio of net profit to total liabilities is the most effective variable among all selected variables in previous stage, taking part in the model during the first step. Finally, after doing statistical calculations, it was revealed that variables being preserved in the model during the fifth step had a significant effect on dependent variable, i.e. financial distress, or in other words, the distressed and sound companies. So, variables of net profit to total liabilities, equity to total liabilities, EBIT to total liabilities, EBIT to total assets and sales to total assets are selected as effective variables during the year "t".

Table 3. Diagnostic Analysis in Step-by-Step method – "t-1" Year

Step	Variables in the Analysis	Tolerance	F to Remove	Wilks' Lambda
1	Net profit to total liabilities	1.000	170.842	
2	Net profit to total liabilities	.841	79.692	.721
	Current Assets to Current liabilities	.841	6.876	.525
3	Net profit to total liabilities	.044	15.959	.536
	Current Assets to Current liabilities	.834	5.590	.508
	EBIT to total assets	.045	5.001	.507
4	Net profit to total liabilities	.044	15.109	.521
	Current Assets to Current liabilities	.826	4.430	.493
	EBIT to total assets	.044	6.391	.498
	EBIT to total liabilities	.557	4.692	.494

Results of above table indicate that the ratio of net profit to total liabilities is the most effective variable among all selected variables in previous stage, taking part in the model during the first step. Finally, after doing statistical calculations, it was revealed that variables being preserved in the model during the forth step had a significant effect on dependent variable, i.e. financial distress, or in other words, the distressed and sound companies. So, variables of net profit to total liabilities, current assets to current liabilities, EBIT to total assets and EBIT to total liabilities are selected as effective variables during the year "t-1". Subsequently, we will assess the effective variables obtained for the years "t" and "t-1" in their algorithms and outputs. To perform algorithms, Matlab Software, version 7.6 is used.

7. Results and Discussion

Results of algorithms are presented in this section.

Table 4. Results of algorithms for "t" year

Fold	Sound Companies	Distressed Companies	Correct classification
CART	97.07	94.76	95.95
MLP	92.64	94.04	92.87
SVM	89.39	95.04	91.84
Combination	92.53	95.76	93.89

Table 4 shows results of algorithms for "t" year. Percentage of sound companies, anticipated correctly by each algorithm is represented in the second column of above table, and percentage of distressed companies is also represented in the third column. In the last column, percentage of sound and distressed companies anticipated correctly by the model is represented. A comparison between algorithms according to total accuracy of model shows that decision tree with 95.95% accuracy has the best performance during "t" year.

Table 5. results of algorithms for "t-1" year

Fold	Sound Companies	Distressed Companies	Correct classification
CART	88.72	83.62	86.84
MLP	85.89	91.94	89.92
SVM	85.04	88.83	87.82
Combination	87.53	90.83	89.89

Table 5 shows algorithm anticipation results during one year previous to financial distress. Percentage of sound companies, anticipated correctly by each algorithm is represented in the second column of this table too, and percentage of distressed companies is represented in the third column. In the last column, percentage of sound and distressed companies is represented. During "t-1" year, Perceptron multi-layer neural network (MLP) has the best performance with 89.92% accuracy.

8. Conclusion

Anticipation of financial distress has been always one of the important studies in the field of financial researches. Banks, credit institutes and actors in economic markets usually use these patterns to make their decision. In-time anticipation of financial distress and then the problem stemming provides the possibility to achieve satisfactory results.

This study compares the anticipation accuracy of three methods (CART, MLP, and SVM) in anticipation of financial distress of companies. The effect of combination of these three methods is also considered by relative majority voting method to improve anticipation of financial distress of companies.

Results of the study showed that all three methods have a high performance in anticipation of corporate financial distress. Comparing results also indicated that CART decision tree with 95.95% accuracy during the year "t" has a better potential than two other algorithms for anticipation, but results obtained during one year prior to financial distress showed that Perceptron multi-layer neural network (MLP) has a better predictive power than two other algorithms with 89.92% accuracy.

Results obtained through combination of the majority voting method show that although this method is subject to more complexity and requires more time in comparison with single algorithms, but does not have higher accuracy in anticipation of corporate financial distress; so that, anticipation accuracy of combination of these three methods for the year "t" is 93.89% and for the year "t-1" is 89.89%. Of course, the combinational method has a higher function comparing with single methods in some other studies.

Based on the results, it is generally concluded that since anticipation percentage of decision tree algorithm is higher for the year "t" and there is a little difference between this algorithm and other data mining methods during the year "t-1", and regarding advantages of this

algorithm and also disadvantages of other methods especially due to low transparency of neural networks and support vector machine and low generalizability of these models, decision tree can be a suitable substitute for two other algorithms having been used more to anticipate financial distress and bankruptcy, because decision tree algorithm provides generalizability and transparency of model, along with its high accuracy in anticipation of corporate financial distress. As far as the combination method using the majority voting approach is concerned, the results show that in spite of providing higher function in anticipation of financial distress of companies, the use of combination method is not recommended due to its higher complexity and more required time and expenses than single methods, although it is likely to obtain a better results from combination method through participation of a greater number of algorithms in the voting which surely requires more complexity and more time.

References

[1] Breiman, J. H. Fried man, R .A. Olsen, C .J. Stone (1984), "Classification and Regression Trees", Wadsworth & Brooks /Cole Advance d Books & Software, Monterey, CA.

[2] Chen, Mu, Yen. (2011). "Predicting corporate financial distress based on integration of decision tree classification and logistic regression". Expert Systems with Applications, Volume 38, Issue 9, Pages 11261-11272.

[3] Chen, Mu, Yen. (2011). "Bankruptcy prediction in firms with statistical and intelligent techniques and a comparison of evolutionary computation approaches". Computers & Mathematics with Applications, Volume 62, and Issue 12, Pages 4514-4524.

[4] Chen, Wei, Sen., Du, Yin, Kuan. (2009). "Using neural networks and data mining techniques for the financial distress prediction model" Expert Systems with Applications, Volume 36, Issue 2, Pages4075-4086.

[5] Cielen, Anja. Peeters, Ludo., Vanhoof, Koen. (2004). "Bankruptcy prediction using a data envelopment analysis". European Journal of Operational Research, Volume 154, Issue 2, Pages 526-532.

[6] Fadaei-Nejad, Mohammad Esmail; Eskandari, Rasool (2011) "Design and Explanation of Bankruptcy Anticipation Model of Companies in Tehran Stock Exchange Market"; Accounting and Auditing Considerations; Year 3; No. 9; pp: 38-55.

[7] Ghodrati, Hassan; Manavi-Moghaddam, Amir Hadi (2010) "A Consideration of Accuracy in Bankruptcy Anticipation Models (Models of Altman, Shirata, Ohlson, Zemiski, Springit, C. A. Score, Folmer.

[8] Hay kin. (2008), "Neural N et work s an d Learning Machines", 3 r d Ed. Prentice Hall, NewJersey.

[9] Hua, Z., Wang, Y., Xu, X., Zhang, B., Liang, L. (2007). "Prediction corporate financial distress based on integration of support vector machine and logistic regression". Expert system with application, Volume 33, Issue 2, Pages 434-440.

[10] Kyung, Shik, Shin., Taik, Soo, Lee. Hyun, Jung Kim (2005). "An application of support vector machines in bankruptcy prediction model". Expert Systems with Applications, Volume 28, Issue 1, Pages 127-135.

[11] Nikbakht, Mohammadreza; Sharifi, Maryam (2010) "Anticipation of Financial Bankruptcy of Companies Listed in Tehran Stock Exchange Market by Artificial Neural Networks (ANN)"; Industrial Management; series 2; No. 4; pp: 163-180.

[12] Oh, S. -B . (2 0 03). O n the relationship between majority vote accuracy and dependency in multiple classifier systems. Patter n Recognition Letters, 24, 35 9 – 3 63.

[13] Olson, David. Delen, Dursun., Meng, Yanyan. (2012). "Comparative analysis of data mining methods for bankruptcy prediction". Decision Support Systems, Volume52, Issue2, Pages464-473.

[14] Rostami, Mohammadreza; Fallahbakhsh, Mirfeyz; Eskandari, Farzaneh (2011) "Assessment of Financial Distress in Companies Listed in Tehran Stock Exchange Market: A Comparative Study between Data Envelopment Analysis and Logistics Regression"; Management Researches in Iran; Series 15; Number 3; pp: 129-147.

[15] Sun, Jie, Li, Hui. (2008). "Listed companies financial distress prediction based on weighted majority voting combination of multiple classifiers". Expert System with Application, Volume 35, Issue3, Pages 818-827.

[16] Sun, Jie, Li, Hui. (2009). "Majority voting combination of multiple case-based reasoning for financial distress prediction". Expert System with Application. Volume 36, Issue 3, Pages 4363-4673.