Genetic Programming Decision Tree for Bankruptcy Prediction

Wo-Chiang Lee

Department of Finance and Banking
Aletheia University
E-mail: wclee@email.au.edu.tw
32 ,Chen Li Street ,Tamsui, Taipei County ,Taiwan

Abstract

In this paper, we apply the CART ,C5.0 , GP decision tree classifiers and compares with logic model and ANN model for Taiwan listed electronic companies bankruptcy prediction. Results reveal that the GP decision tree can outperform all the classifiers either in overall percentage of correct or k-fold cross validation test in out sample. That is to say, GP decision tree model have the highest accuracy and lowest expected misclassification costs. It can provide an efficient alternative to discriminates financial distress problems in Taiwan.

Keywords: Financial distress model, decision tree, GP decision tree

1. Motivation and Introduction

Measuring the credit risk accurately also allows banks to engineer future lending transactions, so as to achieve targeted return/risk characteristics. Hence, the bankruptcy prediction is an important and widely studied topic since it can have significant impact on bank lending decisions and profitability.

As my best knowledge, the traditional approach for credit risk of banks is to produce internal rating, which takes into account various quantitative as well as subjective factors, such as leverage, earnings, reputation, etc., through a scoring system. The problem with this approach is of course the subjective aspect of the prediction, which makes it difficult to make consistent estimates. Some banks, especially smaller ones, use the ratings issued by the standard credit rating agencies, such as Moody's and Standard & Poor's. The problem with these ratings is that they tend to be reactive rather than predictive[1]. Therefore, to develop a fairly accurate quantitative prediction models that can serve as very early warning signals for counterparty defaults.

Beaver(1966),Altman(1968) and Ohlson(1980) are the pioneers of the financial distress empirical approach. Beaver, in particular, was one of the first researcher to study the prediction of bankruptcy using financial statement data. However, his analysis is very simple in that it is based on studying one financial ratio at a time and on developing a cutoff threshold for each ratio. The approaches by Altman(1968) and Ohlson(1980) are essentially linear models that classify between healthy and bankrupt firms using financial ratios as inputs. Altman(1968) used the classical multivariate discriminate analysis technique (hence **MDA)**. Both the MDA model and the linear regression model (hereafter **LR)** have been widely used in practice and in many academic studies. They have been standard benchmarks for the loan default prediction problem. Whereas research studies on using artificial neural network (hence **ANN**) for bankruptcy prediction started in 1990, and are still active now.
It is worthy to attention, the decision tree has became a very popular data mining technique and commonly used for classification[2], but can also be used for regression. So, the focus of this article is on the empirical approach, especially the use of the decision tree model. Besides, we will compare different decision tree algorithms with two main approaches for bankruptcy prediction. The first approach is the logistic regression model. The second approach is artificial neural networks model.

This paper is organized as follows. Section 2 gives a brief introduction to GP decision tree algorithms. Section 3 shows the data description and explain our experiment design and results analysis followed by a few concluding remarks in Section 4.

2. Brief Review of GP Decision Tree Algorithms

1. For the agencies to change a rating of a debt, they usually wait until they have a considerably high confidence/evidence to support their decision.

2. Predicting what group a case belongs to.

The decision tree method encompasses a number of specific algorithms which including Classification and Regression Trees (hereafter **CART**), Chi-squared Automatic Interaction Detection (hence **CHAID**), **C4.5** (Quinlan, 1993)**, C5.0** (Quinlan,2003) and integration of C4.5 with genetic programming(hence **GP decision tree**).

In this paper, we use the CART,C5.0 and GP decision tree algorithms except traditional logistic regression and ANN models. The reason is that the previous two decision trees algorithms had been founded to be quite effective for creating decision rules which perform as well or better than rules developed using more traditional methods. While the GP decision tree is a new hybrid algorithms. We will take a brief review of GP decision tree models.

Genetic programming, a branch of genetic algorithms (**GA)**, is a technique that applies the Darwinian theory of evolution to develop efficient computer programs(Koza ,1992).The main difference between **GP** and **GA** lies in the representation of the solution. **GP** creates computer programs in the lisp or scheme computer languages as the solution. **GA** creates a string of numbers that represent the solution.

In the recent years, genetic programming **(GP)** has been successfully applied to solve different optimization search problems. It can be used as long as the solution can be encoded in tree structure. Hence, a more active way of overcoming the limitations of standard greedy decision tree induction algorithms can be the usage of genetic programming**.**

Integration with GP and decision tree, each individual of the population in GP can be a decision tree. The functions to be used in the GP are the attributes of the decision tree and classes form the terminal set. Further to say,C4.5 is one of the tools for designing decision trees from training examples, In most cases ,C4.5 can generate near optimal decision tree when the training data are given all together. However, if the training data are given incrementally. C4.5 cannot be used. In this case, genetic programming (GP) might be a better choice. Actually, GP can be considered as a decision tree breeder in which good decision tree can be generated automatically through evolution. In GP based decision tree design the training examples can be given all together.

In this paper, we try to integrate C4.5 and GP in such a way that each individual is initialized by C4.5 using part of the training examples. By so doing, we can have relatively good decision tree from the very beginning and use them while waiting for better decision tree to emerge.

To design the decision tree using GP, each individual is defined as a decision tree, which represents both the genotype and the phenotype. The design process is still an evaluation process containing two phases,

- Select part of the training examples at random from the whole training set, and design a decision tree using C4.5 ,Repeat this for all trees in the initial population.
- Evolve the tree using GP.

To test the effectiveness of integrating C4.5 and GP, we conducted some experiments with a financial distress data set and compares with other classifiers.

3. Empirical Results and Analysis

3.1 Variables description

Table 1 Variables description

Variables	Description
X1	Return on total assets(ROA)
X2	Current ratio
X3	Ratio of stock price to cash flow
X4	Fixed asset turnover ratio
X5	Holding ratio of major Stockholders
X6	Earning after taxes(EAT)
X7	Coverage ratios
X8	Distance-to-default(DD)
Y	Classification output:{1:financial distress; 0:normal }

We obtained the empirical data from *Taiwan Economic Journal* (hereafter **TEJ**) databank. Some listed electronic companies, which had been occurred financial distress, are included in the sample except financial security corporations.

Data derives from year 1999 to 2003.During the five years, total 55 listed electronic companies occur financial distress. According to Beaver (1966) and Altman (1968), we make a match for scale and size as a 1:2 ratio. This is to say, total 110 normal and 55 financial distress companies are contained in the databank.

The next step, we choice the previous year financial and non-financial ratio of above listed companies as input variables[3]. Most of studies used the **CAMELS** as indexes. In the paper, we follow Lee (2004) and considered a pool of about 31 financial variables and 3 non-financial variables. In addition, we also consider the distance to default **(DD)** variables which is calculated form **KMV** model[4]. We apply

[3] The reason of using one year ahead financial ratio is that most researches had provided that two year ahead financial ratio is worse than one year ahead.

[4] Crosbie, P.J.(1999).Modeling default risk. KMVCorporation,12-January.

factor analysis to narrow down the choice and extracted eight input variables which are defined as table 1.

These popular financial ratios will be used as inputs. Even for decision tree and other nonlinear models. In our study , the observed dependent variable is determined by whether exceeds a threshold value 0.5.Futhermore,we put the test sample into the logit equation,then we can obtain the logit test sampleresults.Whereas in our ANN model uses the eight variables as inputs. The model is presented as equation (1).

$$Y_{ANN} = f(X1, X2, X3, X4, X5, X6, X7, X8) \qquad (1)$$

$$Y = \begin{cases} 1 & ,if \quad Y_{ANN} \geq 0.5 \\ 0 & ,if \quad Y_{ANN} < 0.5 \end{cases} \qquad (2)$$

The output Y_{ANN} is in between 0 and 1.Equation (2) is called as ANN decision rule.The parameters in ANN are as follows:

- Hidden unit is 10,15,20
- Transfer function is hyperbolic.
- Learning algorithms is backpropagation.

In our decision tree models, CART is conducted by **MATLAB7.0** and C5.0 is executed by **SEE5** software. Besides, we use the GP-system **Discipulus** developed by Register Machine Learning Technologies Inc.(1998-2004). Discipulus is a general purpose GP-system which can be used for regression and binary classification problems. The software creates small programs with the technique of GP which should solve a question, for example to decide whether a specific sample is malignant or not.

As we used the system only for this kind of classification problems, we call the generated programs classifiers. A classifier is very similar to an assembler program with commands for simple terms of GP these operations establish the functions set.
In this paper, we conduct GP-runs; the data set has been divided into two samples, training set and test set. Discipulus used tournament selection to compare the fitness of the program on the training set. The test set is used to determine how well the best programs generalize. Table 2 shows the parameters setting.

3.2 Result analysis

Table 3 compares the accuracy and performances of all of classifiers in our empirical study. Under the 0.5 cut-off value, the training sample overall percentage of correct of Logit model is 81%. The type I error is 42.42% and type Ⅱ error is 7.46%.The test sample

Table 2 Tableau for Genetic Programming

Population size (N)	500
Number of trees created by complete growth	250
Number of trees created by partial growth	250
Function set	{+,-,*,/,sin,cos, ,log,power}
Terminal set[5]	{1,0}
Criterion of fitness (F)	Sum of squared errors
Number of generations[6] (n)	200

Table 3 Comparison accuracy of classifier

Training sample					
Items	**Logit model**	**ANN model**	**CART**	**C5.0**	**GP-decision tree**
Overall (%)	81	96	91	82	100
Class 1(%)	57.58 (42.42)	87.88 (12.12)	81.81 (18.18)	78.79 (22.21)	100 (0)
Class 0(%)	92.54 (7.46)	100 (0)	95.52 (4.48)	83.58 (17.42	100 (0)
Test sample					
Overall (%)	69.23	73.85	75.38	72.3	92.91
Class 1(%)	41.86 (58.14)	40.91 (59.09)	50.00 (50.00)	77.27 (22.73)	92.91 (7.09)
Class 0(%)	90.91 (9.09)	90.70 (9.30)	88.37 (11.63)	69.76 (31.64)	86.36 (13.64)

Note :The parentheses in Class 1 is Type I Error(%) and Type Ⅱ Error(%) in Class 0,respectively.

overall percentage of correct is 69.23% and type I error is 58.14% ,type Ⅱ error is 9.09%.

The best ANN model shows that the training sample and out sample overall percentage of correct is 96% and 73.85%, respectively. The two values are higher than the Logit model. The type I error is 12.12%,type Ⅱ error is 0 in training sample. But the type I error is 59.09% ,type Ⅱ error is 9.30%.

For CART model, the overall percentage of correct in training sample is 91%, higher than the logit and C5.0 decision tree model, the test sample is 75.38%. The type I error is18.18% ,type Ⅱ error is 4.48% in training sample. They are 50.00% and 11.63% contrast to test sample.

Table 3 also reports the C5.0 decision tree model results. The training sample overall percentage of correct is 82%, the test sample overall percentage of correct is 72.30%.Both the performances are only higher than the logit classified model. The type I error is 22.21% ,type Ⅱ error is 17.42% , The type I error is 22.73% ,type Ⅱ error is 31.64%.

[5] The classification result is 1 and 0,where 1 represents the financial distress and 0 is normal.

[6] When the number of generations (n) is set to 200, the convergence is met.

The last column of table 3 reports the GP decision tree model results. The training sample overall percentage of correct is 100%, the test sample overall percentage of correct is 92.91%. Both the performances are higher than all of the classifiers. The type I error and type Ⅱ error are 0 in training sample. Whereas the type I error is 7.09% ,type Ⅱ error is 13.64% in out sample.

In order to confidently lessen the effects of algorithmic bias, a way of performing repeated training and testing is possible. We conduct the 5-fold cross validation. Table 4 presents the comparison of 5-fold average results. The GP model can outperform all of the classifier, and then is CART model and ANN decision tree model. Logit model is still the worst in training sample. While in test sample, the GP decision tree can beat all of the models, ANN is the worst.

Table 4 Comparison of 5-fold cross validation accuracy

Training sample					
Items	Logit model	ANN model	CART	SEE	GP
1	0.9394	0.9697	1.0000	0.9600	1.0000
2	1.0000	1.0000	0.9697	0.9000	1.0000
3	0.8485	0.9091	0.9697	0.9600	0.9900
4	1.0000	1.0000	1.0000	0.9500	0.9697
5	0.7576	0.9697	0.9697	0.8800	1.0000
Average	**0.9091**	**0.9697**	**0.98182**	**0.9300**	**0.9919**
S.E.	**0.1049**	**0.0371**	**0.0165**	**0.0374**	**0.0131**
Test sample					
1	0.6769	0.7692	0.6923	0.7730	0.9000
2	0.7076	0.6462	0.6462	0.7230	0.8000
3	0.7230	0.6769	0.7538	0.7690	0.9242
4	0.6615	0.6308	0.7538	0.6770	0.8769
5	0.7076	0.6615	0.6769	0.7380	0.9300
Average	**0.6953**	**0.6769**	**0.7046**	**0.736**	**0.8862**
S.E.	**0.02525**	**0.0543**	**0.04788**	**0.0390**	**0.0470**

Note: S.E. means the standard error.

In an addition to k-fold validation, we further apply the cumulative accuracy profile (**hence CAP)** index and the receiver operating characteristic **(hence ROC)** to compare the performances of all kind of classifiers. According to Basel Committee on Banking Supervision's working paper (2005), the rating method is better when the closer area of CAP and ROC is to one. Table 5 shows the area and average value of CAP and ROC. It still represents that the GP decision tree is better than other classify models in the out sample, then is ANN ,comes again is Logit and C5.0.The CART is worst.

Table 5 Area of CAP and ROC

Model	Logit	ANN	CART	C5.0	GP
AR-CAP	0.7082	0.7188	0.3234	0.4672	0.7991
AR-ROC	0.854	0.859	0.6617	0.7336	0.8995
average	0.781	0.7889	0.4925	0.6004	0.8493

4. Concluding Remarks

In this paper, we reviews and evaluates five types of classifiers for financial distress prediction. From many studies exiting in the literature, it can be seen that decision tree model and ANN are generally more superior to logic model. It also obtains some evidences in this paper. However, ANN has also being criticized to identify the relative importance of potential independent variables, and certain interpretative difficulties.

We shows that the GP decision tree yields the best classification accuracy though the approximate decision rules inferred are less intuitive and humanly understandable. That is to say, it can provide an efficient alternative for discriminate financial distress problems in Taiwan. Furthermore, the result is feasible to construct the bankruptcy prediction model.

An interesting topic for further research might consider hybrid of genetic learning algorithms and decision tree.

5. References

[1]. Altman, E. I., "Financial Ratios, Discriminate Analysis and the Prediction of Corporate Bankruptcy," *Journal of Finance*, 23(4), pp.578-609 ,1968.

[2]. Basel Committee on Banking Supervision, "Studies on the Validation of Internal Rating Systems," *Working paper* ,No.14.,2005.

[3]. Beaver, R., "Financial Ratios as Predictors of Failure, Empirical Research in Accounting: Selected Studies," *Journal of . Accounting Research*, Vol. 4, pp.71-111,1966.

[4]. Crosbie, P. J., "Modeling Default Risk. KMV Corporation,"12-January,1999.

[5]. Koza, J. R., "Genetic Programming: On the Programming of Computers by Means of Natural Selection," *Cambridge: MIT Press*,1992

[6]. Lee, W. C., "Empirical Study for KMV Model in the Financial Distress(in Chinese)," *SinPac Financial Journal* ,Vol.26, pp.97-138.,2004.

[7]. Ohlson, J., "Financial ratios and the probabilistic prediction of bankruptcy," *Journal of Accounting Research.*, Vol. 18, pp. 109–131.,1980.

[8]. Quinlan, J. R. , "C4.5: Programs for Machine Learning," Morgan Kauffman.,1993

[9]. Quinlan, J. R,"C5.0 Online Tutorial," http://www.rulequest.com.,2003