Australian Journal of Basic and Applied Sciences, 5(10): 629-634, 2011
ISSN 1991-8178

Comparing the Ability of Bayesian networks and Adaboost for Predicting Financial Distress of Firms listed on Tehran Stock Rxchange (TSE)

[1]Seyedhossein Naslmosavi, [2] Arezoo Aghaei Chadegani and [3]Mohammadghorban Mehri

[1]Department of Accounting, Qaemshahr Branch, Islamic Azad University, Qaemshahr, Iran.
[2]Department of Accounting, Mobarakeh Branch, Islamic Azad University, Mobarakeh, Isfahan, Iran
[3]Ghorveh Branch, Islamic Azad university, Ghorveh, Iran.

Abstract: Financial distress and bankruptcy of companies may cause the resources to be wasted and the investment opportunities to be faded. Bankruptcy prediction by providing necessary warnings can make the companies aware of this problem. The aim of this study is to compare the ability of Bayesian networks and adaboost for predicting financial distress of firms listed on Tehran Stock Exchange (TSE). Two naïve bayes models were developed based upon conditional correlation between variables and conditional likelihood. The accuracy in predicting bankruptcy of the first naïve bayes model's performance that is based upon conditional correlation is 90% and the accuracy of the second naïve bayes model is 93% and finally the accuracy of the adaboost that was built to compare with naïve bayes models is 88%. Collectively the results show that it is possible to predict financial distress using Bayesian and Adaboost models. But, Bayesian networks are more capable to predict financial distress of companies listed on TSE compare to Adaboost. With respect to the variables in developed models in this research we find that firms with lower profitability and more long term liabilities and lower liquidity are more in Risk of financial distress. To reduce financial distress risk, firms should use more conservative methods which lead to decrease in debts and reduce their costs.

Key words: Bankruptcy predictors; Financial distress; Bayesian networks; naïve Bayes; Adaboost; discretization of continuous variables; companies listed on Tehran Stock Exchange(TSE).

INTRODUCTION

The prediction of financial distress is an important and challenging issue that has attracted a lot of attention of academic studies. Bankruptcy prediction is an important concern for the various stakeholders in a firm, owners, managers, investors, creditors and business partners, as well as government institutions responsible for maintaining the stability of financial markets and general economic prosperity. The recent financial scandals such as Enron and WorldCom, in the United States once again highlighted the economic importance of this task. Techniques employed to develop bankruptcy prediction models have evolved from the simple univariate analysis (Beaver 1966) and multiple discriminant analysis (MDA) (Altman 1968) in the 1960s, to logit and probit models in the 1980s (Ohlson 1980, Zmijewski 1984), to neural network models (NN) (Tam and Kiang 1992), rough set theory (McKee 1998), discrete hazard models (Shumway 2001), Bayesian network (BN) models (Sarkar and Sriram 2001), and genetic programming (McKee and Lensberg 2002). Among these techniques, BN models and adaboost have many attractive features. They are easy to interpret, perform well as a classification tool, have no restriction on variables' underlying distributions, and have no requirement of complete information (Sun and Shenoy 2006).

In this study, we compare the ability of Bayesian networks and Adaboost models for predicting financial distress of companies listed on Tehran Stock Exchange. The sample consists of 72 bankrupt firms and 72 non bankrupt ones from 1997 to 2007 and bankrupt firms are those firms that subject to business law par. 141. In order to develop a bankruptcy prediction model, we consider 20 predictor variables including liquidity ratios, leverage ratios, profitability ratios and other factors like firm's size and auditor's opinion and then we use two methods for choosing variables. The first method is based upon conditional correlation between variables and the second method based upon conditional likelihood. Then three models for predicting financial distress are developed using naïve bayes model and adaboost model and the result of three models are compared.

The remainder of this paper is organized as follows. Section 2 provides a literature review on bankruptcy prediction techniques. In section 3, we discuss the probabilistic concepts underlying BN models and Adaboost. In section 4, we describe our sample and data. Section 5 describes research process and present results. Section 6 summarizes and concludes the paper.

Literature Review:

William Beaver (1966) conducted a very comprehensive study using a variety of financial ratios to run a model for bankruptcy prediction. His conclusion was that the cash flow to debt ratio was the single best

Corresponding Author: Seyedhossien Naslmosavi, Department of Accounting, Qaemshahr Branch, Islamic Azad University, Qaemshahr, Iran.
E-mail: nseyedhossoein@live.utm.my

predictor of bankruptcy. Beaver's univariate analysis led the way to a multivariate analysis by Edward Altman (1986), who used multiple discriminant analysis (MDA) in his effort to find a new bankruptcy prediction model. He selected 33 publicly-traded manufacturing bankrupt companies between 1946 and 1965 and matched them to 33 firms on a random basis for a stratified sample (assets and industry). The results of the MDA exercise yielded an equation; he called the Z-score, which correctly classified 94% of the bankrupt companies and 97% of the non-bankrupt companies one year prior to bankruptcy. The ratios used in the Altman model are: working capital over total assets, retained earnings over total assets, earnings before interest and taxes over total assets, market value of equity over book value of total liabilities, and sales over total assets (Altman, 1986).

Beginning in the 1980s more advanced estimation methods, such as logit (Ohlson 1980) and probit (Zmijewski 1984), were employed. In 1990, the NNs technique has jumped into the field of corporate bankruptcy prediction and Odom and Sharda (1990) were the first to apply NNs to bankruptcy prediction problem. The NNs technique dominates the literature on business failure in 1990s, and still most frequently used in corporate bankruptcy prediction (Heui-Yeong 2004). Later on, Sarkar and Sriram (2001) developed Bayesian network (BN) models for early warning of bank failures. They found that both a naïve BN model and a composite attribute BN model have comparable performance to the well-known induced decision tree classification algorithm. Some other techniques, such as rough set theory (McKee 1998), discrete hazard models (Shumway 2001), and genetic programming (McKee and Lensberg 2002), have also been introduced to the bankruptcy prediction area (Sun and Shenoy 2006). Some studies try to compare the ability of these methods to find which model has more capability to predict financial distress. For example, Alfaro *et al.,* (2008) compare adaboost with neural networks. They find that adaboost decreases the generalization error by about thirty percent with respect to the error produced with neural networks. Regarding the high ability of Bayesian networks for predicting bankruptcy we motivated to compare Bayesian networks performance with adaboost algorithm. In next part Bayesian networks and adaboost algorithm will be discussed.

Bayesian Network Models:

Bayesian Networks are gaining an increasing popularity as modeling tools for complex problems involving probabilistic reasoning under certainty. Bayesian networks (BN) are probabilistic graphical models that represent a set of random variables for a given problem, and the probabilistic relationships between them. The structure of a BN is represented by a direct acyclic graph (DAG), in which the nodes represent variables and the edges express the dependencies between variables (Pearl 1988).

Underlying Concept and Theory:

Bayesian networks are based upon probability theory and the basic measure of our belief in a proposition (say A) will be the function P (A). The basic concept in the Bayesian treatment of certainties in Bayesian Network; is conditional probability which gives a measure of how our beliefs in certain propositions are changed by the introduction of related knowledge.
Bayes rule can be expressed as follows:

$P(A\backslash B) \; P(B\backslash A)\, P(A) / P(B)$

Advantages and Disadvantages of Bayesian Networks:

The major advantage of Bayesian Network is that the output is explicitly a probability, which can be easily interpreted. A decade ago the calculations needed to prograte the probabilities through the nodes of complex Bayesian Networks were prohibitively time consuming but the increased capacity of modern computers has attracted interest in the implementation of Bayesian Networks. An important advantage of a Bayesian Network is the availability of a graphical model framework of a problem, which is useful for both people and computers. One problem of Bayesian Network centre's on the quality and extends of the prior beliefs used in Bayesian inference processing. A Bayesian Network is only as useful as this prior knowledge is reliable since the quality of these prior beliefs will distort to the entire network and invalidate the results. Perhaps the most significant disadvantage of an approach involving Bayesian Networks is the fact that there is no universally accepted method for constructing a network from data (Kyprianidou 2002).

A Naïve Bayes Bayesian Network Model:

A naive Bayesian network is a very simple structure in which all random variables representing observable data have a single, common parent node-the class variable. The naive Bayesian classifier has been used extensively for classification because of its simplicity, and because it embodies the strong independence assumption that, given the value of the class, the attributes are independent of each other.

Figure 1 presents a graphical representation of a naïve Bayesian network model.

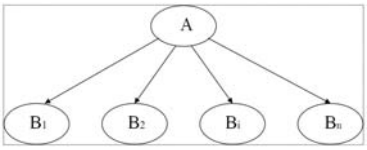

Fig. 1: A Naïve Bayes BN Model.

In a naïve Bayes model, the node of interest has to be the root node, which means, it has no parent nodes. In a bankruptcy prediction context, in Figure 1, *A* represents the bankruptcy variable. *B*1, *B*2 …, *Bn* represent *n* bankruptcy predictor variables. The naïve Bayes model assumes the following conditional independence: $Bi \perp \{B1, B2, ..., Bi\tilde{1}, Bi1, ...Bn\} | A$, for $i = 1, 2, \ldots, n$.

The above assumption says that predictors, *B*1, *B*2 …, *Bn* are conditionally mutually independent given the state of bankruptcy.

Adaboost Algorithm:

AdaBoost is a new classification method which is used extensively for prediction in classification issues. It is also a learning algorithm used to generate multiple classifiers from which the best classifier is selected (Schapire, 1999).

Schapire (2008) states that Adaboost algorithm is defined as follow:

input: $(x_1, y_1), \ldots, (x_m, y_m)$, where $x_i \in \mathcal{X}$, and $y_i \in \{-1, +1\}$;
initialization: $D_1(i) = 1/m$;
for t **from** 1 **to** T **do**
  run A on D_t and get $h_t : \mathcal{X} \to \{-1, +1\}, h_t \in \mathcal{H}$;
  $D_{t+1}(i) = \frac{D_t(i)}{Z_t} e^{-\alpha_t y_i h_t(x_i)}$;
  where $\alpha_t = \frac{1}{2} \ln \left(\frac{1-\epsilon_t}{\epsilon_t} \right), \epsilon_t = \mathrm{err}_{D_t}(h_t)$, and Z_t is normalization factor.
end
output: final/combined hypothesis: $H(x) = \mathrm{sign}(\sum_{t=1}^{T} \alpha_t h_t(x))$.

Suppose Sn = {(x1, y1), (x2, y2), . . . , (xn, yn)} is a set of training samples and yi (i = 1, 2, . . . , n) ε {-1, +1}, which represents only two classes for simplification purpose, in case of bankruptcy (bankrupt and non bankrupt). The weight distribution over these samples at the t boosting iteration is denoted as $W_t = \{W_t^1, W_t^2, \ldots, W_t^n\}$ (t = 1,2…., T) which is initially set uniformly. It means the It means the weight w_t^i (i = 1, 2, . . . , n) is given a value of 1/n at the first iteration when t = 1, and will be updated adaptively at later iterations. At iteration t, AdaBoost builds a new training data set by sampling from the initial training data set with the weight distribution of Wt, and calls the Weak Leaner to construct a base classifier, represented as f_t, on this new training data set. F_t should then be applied to classifying the samples in the initial data set (Sun *et al.*, 2011).

Sample and Data:

Research sample are companies listed on Tehran Stocks Exchange (TSE) across various industries during the period 1997 - 2007. We do not impose any selection restriction on the size or industry characteristics when forming bankrupt and non-bankrupt samples. At first we identified bankrupt firms: the firms that included in 141 article of trade law. The procedure of finding bankrupt firms results in 72 bankrupt firms during the period 1997 - 2007. Then 72 non bankrupt firms are selected during the period 1997 - 2007 randomly.

Through our own analysis and reviewing past research 20 variables are identified as potential bankruptcy predictors. These variables are included financial ratios measuring firm's liquidity, leverage, turnover, profitability and firm's size and other factors like auditors' opinions. All variables for bankrupt firms are calculated in the year that firms included in 141 article of trade law. And all variables for non bankrupt firms are calculated in the base year. The variables in this study are shown in table 1.

Research Process and Research Results:

First Method for Variable Selection in Naïve Bayes Models:

There exists a large pool of bankruptcy predictors. An appropriate selection of a subset of variables is necessary for developing a useful naïve Bayes model. Variable selection is really important on account of irrelevant and redundant features may confuse the learning algorithm and obscure the predictability of truly effective variables. So, a small number of predictive variables are preferred over a very large number of

variables including irrelevant and redundant ones. One purpose of this paper is to provide a good method to guide the selection of variables in naïve Bayes models. Hence, we compare two different methods for selecting the variables.

Table 1: Definitions of Potential Predictor Variables.

Name	Definition
X1	Natural log of (Total Assets/ GNP Index)
X2	(Current Assets - Current Liabilities)/Total Assets
X3	Current Assets/ Current Liabilities
X4	Operating Cash Flows /Total Liabilities
X5	Current Assets/Total Assets
X6	Cash/Total Assets
X7	Total Liabilities/Total Assets
X8	Long Term Debts/Total Assets
X9	Sales/Total Assets
X10	Current Assets/Sales
X11	Earnings before Interest and Taxes/Total Assets
X12	Net income/Total assets
X13	One if net income was negative for the last two years, else zero
X14	Retained Earnings/Total Assets
X15	(Net income in year t - Net income in t-1)/(Absolute net income in year t + Absolute net income in year t-1)
X16	Natural log of total assets
X17	Zero if auditors' opinions is unqualified otherwise one
X18	Net income/ sales
X19	Retained Earnings/ total owner's equity
X20	Quick assets / total assets

The first method is depending upon correlations and conditional correlations among variables.

First, we obtain the correlations among all variables, including 20 potential predictors and the variable of interest, firm's bankruptcy status. Variables that have significant correlations are assumed to be dependent and therefore connected. The correlations are obtained using the entire sample of 144 firms, including 72 non-bankruptcies and 72 bankruptcies. At first stage four predictors (x5, x8, x11, x12) are connected with B, since they have dependency with B. these variables are first - order variables. Then we identified second-order variables to compensate for the missing information among first-order variables. Conceptually, second-order variables are those that have significant correlations with first-order variables and therefore are expected to provide information on the missing values of first-order variables. To select a given first-order variable's second-order variables, we follow the similar method used to select first-order variables. The major difference is that now we consider each first-order variable instead of *B* as a root variable. For example the conditional correlation between x5 and x17, x7, x9, x19, x4 was significant so they connected to x5 in the model.

After obtaining the conditional correlation between all variables and selecting the first - order and second - order variables, x2, x3, x10, x15, x18 and x20 are eliminated. Figure 2 shows the first naïve bayes model that we obtained.

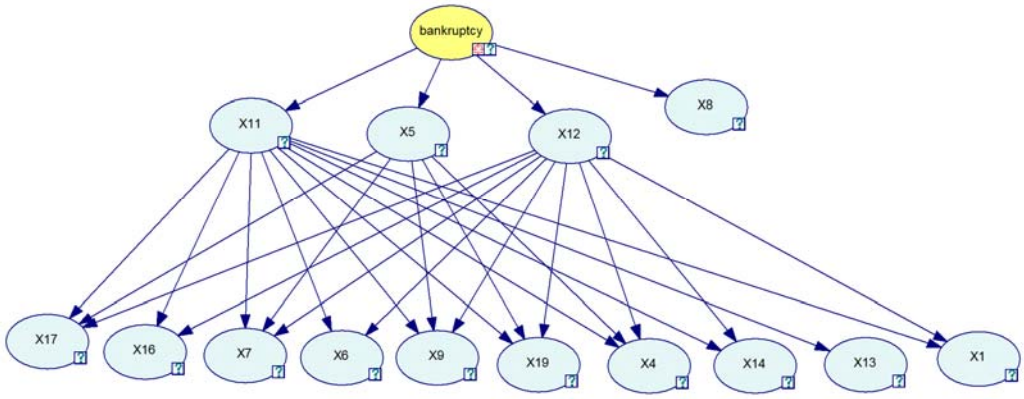

Fig. 2: The Structure of the first Naïve Bayes Model.

The naïve Bayes model is typically used with discrete-valued data. Prior research has used bracket median method (Sarkar and Sriram 2001) and extended Pearson-Tukey (EP-T) method (sun and shenoy 2004) for discretization, which divides the continuous cumulative probability distribution into *n* equally probable intervals. We used Uniform Widths method to convert continuous variables into discrete. During the discretization process, one problem that researchers face is to decide the number of states for discretization. We

started from two states to five states and tested the model with all samples in the research. When continuous variables are discretized into 2 states, the model's accuracy in predicting bankruptcy is 83%, and its accuracy in predicting nonbankruptcy is 74%. When the number of discretization states increases to 3, the model's accuracy in predicting bankruptcy is 84% and its accuracy in predicting non-bankruptcy is 84%.When the number of states increases to 4, the model's accuracy in predicting bankruptcy is 90%, which is statistically indifferent to the model's performance with 2 or 3 states and its accuracy in predicting non-bankruptcy is 89%. When we increase the number of states for discretization further, the model's performance continues to drop.

Second Method for Variable Selection in Naïve Bayes Models:

In the second method variables are selected based upon conditional likelihood.

First, we obtain the correlations among all variables, including 20 potential predictors and the variable of interest, firm's bankruptcy status. Variables that have significant correlations are assumed to be dependent and therefore connected. The correlations are obtained using the entire sample of 144 firms, including 72 non-bankruptcies and 72 bankruptcies. At first stage four predictors (x5, x7, x11, x20) are connected with B, since they have dependency with B. these variables are first - order variables. Then we identified second-order variables to compensate for the missing information among first-order variables. To select a given first-order variable's second-order variables, we follow the similar method used to select first-order variables. The major difference is that now we consider each first-order variable instead of *B* as a root variable.

After selecting the first - order and second - order variables, x1, x6, x13, x14, x15, x16, x18 and x19 are eliminated. Figure 3 shows the second naïve bayes model that we obtained.

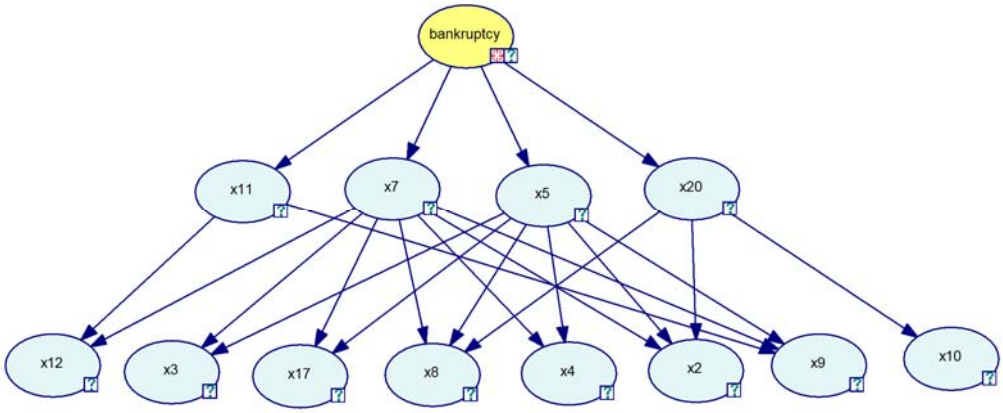

Fig. 3: The Structure of the second Naïve Bayes Model.

When continuous variables are discretized into 2 states, the model's accuracy in predicting bankruptcy is 87%, and its accuracy in predicting non-bankruptcy is 80%. When the number of discretization states increases to 3, the model's accuracy in predicting bankruptcy is 92% and its accuracy in predicting non-bankruptcy is 87%.When the number of states increases to 4, the model's accuracy in predicting bankruptcy is 94%, which is statistically indifferent to the model's performance with 2 or 3 states and its accuracy in predicting non-bankruptcy is 92%. When we increase the number of states for discretization further, the model's performance continues to drop. The rates show that this model's performance is better than the first one.

Naïve Bayes vs. Adaboost Algorithm:

The results of Adaboost algorithm show that this method can predict financial distress of firms listed on Tehran Stock Exchange with 89% accuracy for bankrupt companies and 87% for non bankrupt companies. With respect to the high ability of Bayesian networks and Adaboost algorithm for financial distress prediction, hypotheses 1 and 2 is accepted but hypothesis 3 is rejected because the Adaboost algorithm has not more accuracy rate comparing Bayesian networks.

Summary and Conclusions:

In this study, we examine several important methodological issues related to the use of naïve Bayes Bayesian network (BN) models and Adaboost algorithm to predict bankruptcy and comparing the ability of these two models. First, we provide two different methods that guide the selection of predictor variables from a pool of potential variables. Under the first method, only variables that have significant correlations with the variable of interest, the status of bankruptcy, are selected. As a result, 4 variables are selected from a pool of 20 potential predictors. Then we selected second order variables. The first naïve BN consisting of these selected

variables have an average prediction accuracy of 90% for the bankruptcy sample and 89% for the non-bankruptcy sample.

Second, we investigate the impact on a naïve Bayes model's performance of the number of states into which continuous variables are discretized. We find that the model's performance is the best with the continuous variables being discretized into 4 states. When the number of states is increased to 5 or more, the model's performance deteriorates. Then we use conditional likelihood method for selecting variables and run the second naïve bayes model. This model has an average prediction accuracy of 94% for the bankruptcy sample and 92% for the non-bankruptcy sample. Finally, we run a Adaboost algorithm and our result show that the naïve bayes model that based upon conditional likelihood has the best performance of predicting bankruptcy among these three models. On the basis of study findings, firms that have lower profitability and obtained their assets by loan are more in danger than the others. Also, liquidity is a factor that has inverse relation with appearance of bankruptcy. To reduce bankruptcy risk, firms should apply prudent growth strategies which lead to decrease in debts and handle further cost control. Future research is useful to do a comparison of other methods with Adaboost algorithm. Various variable selection algorithms have been developed for other bankruptcy prediction techniques, such as genetic algorithms for neural networks. It is interesting future research to explore how these algorithms can be applied into BN models.

In addition, there are other important bankruptcy predictors which are not examined by the study. Finally, this study focuses on only one type of BN models: naïve Bayes. Future research is also needed to explore how to better apply other types of BN models to bankruptcy prediction.

REFERENCES

Alfaro, E., N. Garcia, M. Gamez, D. Elizondo, 2008. Bankruptcy Forecasting: An Empirical comparison of Adaboost and Neural networks. *Decision support systems,* 45: 110-122.

Altman, E.I., 1968. Financial Ratios, Discrimanent analysis and the Prediction of corporate Bankruptcy. *The Journal of Finance.* 23: 589-609.

Beaver, W.H., 1966. Financial Ratios as Predictors of Failure. Journal of Accounting Research, Issue, 4: 71-111.

Kyprianidou. Constantia, 2002. Analysing basic Genetics using Bayesian Networks and the Impact of Genetic Testing on the Insurance Industry. Dissertation in Actuarial Science Cass Business School City University.

McKee, T.E. and T. Lensberg, 2002. Genetic Programming and Rough Sets: A Hybrid Approach to Bankruptcy Classification. European Journal of Operational Research, 138: 436-451.

Odam, M. and R. Sharda, 1990. Neural Network for Bankruptcy Prediction. Probus Publishing Company, pp: 177-185.

Ohlson, J.A., 1980. Financial Ratios and the Probabilistic Prediction of Bankruptcy. Journal of Accounting Research, 18: 109-131.

Pearl, J., 1988. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. *Morgan Kauffman Publishers San Mateo, CA*.

Sarkar, S. and R.S. Sriram, 2001. Bayesian Models for Early Warning of Bank Failures. *Management Science,* pp: 1457-1475.

Schapire, 1999. Improved boosting algorithms using confidence rated predictions. *Machine Learning,* 37 (3): 297-336.

Sun, J., M. Jia and H. Li, 2011. Adaboost ensemble for financial distress prediction: An empirical comparison with data from Chinese listed companies. *Expert systems with applications,* 38: 9305-9312.

Sun, Lili, Shenoy, Prakash, 2006. Using Bayesian Networks for Bankruptcy Prediction. European Journal of Operational Research, 180(2): 738-753.

Yeong, Heui, 2004. Financial Data Modeling and Analysis for Bankruptcy Prediction. Working paper. University of Technology. Sydney.

Zimijewski, M.E., 1984. Methodological Issues Related to the Estimation of Financial Distress prediction models. Journal of Accounting research, pp: 59-82.