Available online at www.sciencedirect.com
ScienceDirect
European Journal of Operational Research 180 (2007) 738–753
EUROPEAN JOURNAL OF OPERATIONAL RESEARCH
www.elsevier.com/locate/ejor

Computing, Artificial Intelligence and Information Management

Using Bayesian networks for bankruptcy prediction: Some methodological issues

Lili Sun [a,*], Prakash P. Shenoy [b]

[a] Accounting and Information Systems, Rutgers, The State University of New Jersey, 180 University Ave, Newark, NJ 07102-1897, USA
[b] School of Business, University of Kansas, 1300 Sunnyside Ave, Summerfield Hall, Lawrence, KS 66045-7585, USA

Received 2 March 2005; accepted 19 April 2006
Available online 12 June 2006

Abstract

This study provides operational guidance for building naïve Bayes Bayesian network (BN) models for bankruptcy prediction. First, we suggest a heuristic method that guides the selection of bankruptcy predictors. Based on the correlations and partial correlations among variables, the method aims at eliminating redundant and less relevant variables. A naïve Bayes model is developed using the proposed heuristic method and is found to perform well based on a 10-fold validation analysis. The developed naïve Bayes model consists of eight first-order variables, six of which are continuous. We also provide guidance on building a cascaded model by selecting second-order variables to compensate for missing values of first-order variables. Second, we analyze whether the number of states into which the six continuous variables are discretized has an impact on the model's performance. Our results show that the model's performance is the best when the number of states for discretization is either two or three. Starting from four states, the performance starts to deteriorate, probably due to over-fitting. Finally, we experiment whether modeling continuous variables with continuous distributions instead of discretizing them can improve the model's performance. Our finding suggests that this is not true. One possible reason is that continuous distributions tested by the study do not represent well the underlying distributions of empirical data. Finally, the results of this study could also be applicable to business decision-making contexts other than bankruptcy prediction.

Keywords: Bankruptcy prediction; Bayesian networks; Naïve Bayes; Variable selection; Discretization of continuous variables

1. Introduction

In today's dynamic economic environment, the number and the magnitude of bankruptcy filings are increasing significantly. Even auditors, who have good knowledge of firms' situations, often fail to make an accurate judgment on firms' going-concern conditions (e.g., Hopwood et al., 1994; McKee, 1998, 2003). Therefore, bankruptcy prediction models have become important decision aids for organizations' stakeholders, including auditors, creditors, and stockholders.

Techniques employed to develop bankruptcy prediction models have evolved from the simple univariate analysis (Beaver, 1966) and multiple discriminant analysis (MDA) (Altman, 1968), to

* Corresponding author. Tel.: +1 973 353 5762.
E-mail address: sunlili@rbsmail.rutgers.edu (L. Sun).

0377-2217/$ - see front matter
doi:10.1016/j.ejor.2006.04.019

logit and probit models (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 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.

In order to allow a formal Bayesian model to become useful decision aids, adequate operational guidance needs to be provided (Senetti, 1995). Although some prior work (e.g., Sarkar and Sriram, 2001; Kotsiantis et al., 2005) have introduced BNs to bankruptcy prediction, there is still a lack of proper guidance in the selection of variables and the discretization of continuous variables. This study attempts to fill this void. This study focuses on one type of BN models: naïve Bayes, which are simple to implement and have been shown to perform well in bankruptcy prediction (Sarkar and Sriram, 2001). Specifically, the study addresses the following research questions. First, there exists a large pool of potential bankruptcy predictors, including various financial ratios, stock market information, industry level factors, etc. A method is needed to guide the selection of variables that can be used to develop a well-performing naïve Bayes BN for bankruptcy prediction. This work proposes such a heuristic method based on the assumption of linear dependence as measured by correlations between variables. Grounded on existent feature selection literature (e.g., Koller and Sahami, 1996), the proposed method aims at identifying key predictors and eliminating redundant or irrelevant ones. Secondly, BN models generally use discrete-valued variables. Through discretization, continuous variables are converted into discrete variables with several states. It is unclear whether and how the number of states into which continuous variables are discretized have an impact on BN models' performance. This study explores this issue. The study further examines whether modeling continuous variables with continuous distributions instead of discretizing these variables can improve the model's performance.

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

2. Literature review

In this section, we briefly review some techniques employed to develop bankruptcy prediction models in prior research and discuss the advantages of BN as a classification tool.

Different methods have been implemented in developing bankruptcy prediction models. Beaver (1966) used univariate analysis to compare patterns of 29 ratios in the five years preceding bankruptcy, for a sample of failed firms, with a control group of firms that did not fail. During the late 1960s and throughout the 1970s, multiple discriminant analysis (MDA) was used to develop bankruptcy prediction models. Two of the well-known bankruptcy prediction models, Altman's Z-score (Altman, 1968) and ZETA (Altman et al., 1977) were developed using MDA. Beginning in the 1980s more advanced estimation methods, such as logit (Ohlson, 1980) and probit (Zmijewski, 1984), were employed.

During the 1990s, the neural network (NN) model was introduced into bankruptcy prediction. Research has shown contradictory results regarding NN's superiority over linear models (Altman et al., 1994; Tam and Kiang, 1992). 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 field of bankruptcy prediction.

Prior research has shown that BNs perform well as a classification and prediction tool in different domains (see e.g. Clark and Niblett, 1989; Langley et al., 1992; Pazzani et al., 1996; Sarkar and Sriram, 2001; Anderson et al., 2004). Unlike most regression techniques, BNs do not have any requirements on the underlying distributions of variables. BNs can easily model complex relationships among variables including partial mediators and "interaction effects". BNs do not require complete information for observations. Observations that have some

missing variables can still be used to train or test BN models. This feature is valuable for bankruptcy studies because bankruptcy samples are usually small and bankrupt firms tend to have missing information. BNs are dynamic and interactive. They can easily be updated with new information as it is learned. Subjective human knowledge can easily be incorporated into models. Compared to other machine learning techniques, such as neural networks, BN models are more transparent and intuitive because relationships among variables are explicitly represented by the direct acyclic graph. Users report that BNs' representations are quite intuitive and easy to understand (Kononenko, 1990).

3. Bayesian network models

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). The probabilistic part of the BN is represented by a set of conditional probabilities. Next, we discuss the basic concepts of BN models in the context of bankruptcy prediction.

3.1. Bayes rule

Bayes rule can be expressed as follows:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \tag{1}$$

In a bankruptcy prediction context, this can be interpreted as follows. Suppose we are interested in event A, which represents a company's bankruptcy filing. We start with a prior probability $P(A)$, representing our belief about A before observing any relevant evidence. For instance, $P(A)$ can be measured as the mean percentage of firms in the whole population that have declared bankruptcy in the past. $P(B|A)$ represents the likelihood for bankruptcy based on observing a bankruptcy predictor B such as a late 10-K filing. $P(B)$, the probability of a firm filing its 10-K late, is just a normalizing constant. Suppose we observe B. By Eq. (1), our revised belief for the probability of bankruptcy, the posterior probability $P(A|B)$, is obtained by multiplying the prior probability of bankruptcy $P(A)$ by the likelihood $P(B|A)$ and then normalizing the result by dividing by the constant $P(B)$.

Eq. (1) can be rearranged into Eq. (2), which states that the posterior odds for A equals the prior odds for A multiplied by the likelihood ratio for A from evidence B, i.e.,

$$\frac{P(A|B)}{P(\sim A|B)} = \frac{P(A)}{P(\sim A)} \frac{P(B|A)}{P(B|\sim A)}, \tag{2}$$

where $\frac{P(B|A)}{P(B|\sim A)}$ represents the likelihood ratio for A from evidence B.

Based on the graphical structure of a BN model, it can be classified as a naïve Bayes, a tree augmented naïve Bayes, a general BN, etc. The present study focuses on the naïve Bayes model because it is simple to implement and have been shown to perform well in bankruptcy prediction (Sarkar and Sriram, 2001). Next, we further discuss the naïve Bayes model.

3.2. A naïve Bayes Bayesian network model

The naïve Bayes BN model is named by Titterington et al. (1981) because of its simplicity. Fig. 1 presents a graphical representation of a naïve Bayesian network 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 Fig. 1, A represents the bankruptcy variable. $B_1, B_2, \ldots, B_n$ represent n bankruptcy predictor variables. The naïve Bayes model assumes the following conditional independence:

$$B_i \perp \{B_1, B_2, \ldots, B_{i-1}, B_{i+1}, \ldots, B_n\}|A \quad \text{for } i = 1, 2, \ldots, n.$$

The above assumption says that predictors, $B_1, B_2 \ldots, B_n$ are conditionally mutually independent given the state of bankruptcy. Based on this

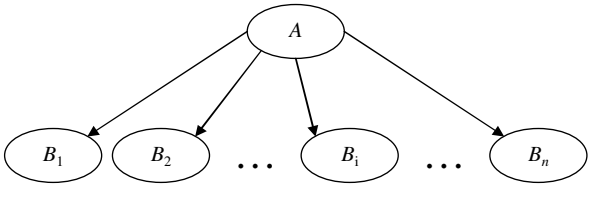

Fig. 1. A naïve Bayes BN model.

Table 1
Definitions of potential predictor variables

Construct	Name	Definition
Financial-accounting factors		
Size	*TA*	Natural log of (total assets/GNP implicit price deflator index). The index assumes a base value of 100 for 1968.
Liquidity	*W*	(current assets − current liabilities)/total assets
	CR	Current assets/current liabilities
	OF	Operating cash flows/total liabilities
	LM	$(L+\mu)/\sigma$. L = cash + short-term marketable securities, μ = mean, σ = standard deviation of quarter-to-quarter change in L over prior 12 quarters
	CA	Current assets/total assets
	CH	Cash/total assets
Leverage	*TL*	(total liabilities/total assets) × 100%
	LTD	Long term debts/total assets
Turnover	*S*	Sales/total assets
	CS	Current assets/sales
Profitability	*E*	Earnings before interest and taxes/total assets
	NT	Net income/total assets
	IT	One if net income was negative for the last two years, else zero
	RE	Retained earnings/total assets
	CHN	(net income in year t − net income in $t-1$)/(absolute net income in year t + absolute net income in year $t-1$)
Market-based factors	*M*	Natural log of each firm's size relative to the CRSP NYSE/AMEX/NASDAQ market capitalization index
	R	The firm's stock return in year $t-1$ minus the value-weighted CRSP NYSE/AMEX/NASDAQ index return in year $t-1$
Other factors	*AU*	AU = zero if Compustat codes auditors' opinions as "1. unqualified"; AU = one if Compustat codings are "0. unaudited"; "2. qualified"; "3. no opinion"; "4. unqualified with additional language"; "5. adverse opinion"
	IFR	Industry failure rate, calculated as the average bankruptcy rate in the past two years, where bankruptcy rate = (the number of bankruptcies in a two-digit SIC industry/the total number of firms in the same industry) × 100%

conditional independence assumption, the posterior odds of A can be expressed as

$$\frac{P(A|B)}{P(\sim A|B)} = \frac{P(A)}{P(\sim A)} \times \prod_{i=1}^{n} \frac{P(B_i|A)}{P(B_i|\sim A)}. \quad (3)$$

Eq. (3), B represents a vector of observations $(B_1,\ldots,B_n)$. If only k of n predictors were observed, then the posterior odds for A is given by an equation similar to (3) where only the likelihood ratios from the k predictors are used (instead of all n predictors as in (3)). The predictors that are not observed have no effect on the posterior odds for A.

4. Sample and data

Sample firms used in this study are publicly traded firms on major stock exchanges (NASDAQ, the New York and American Stock exchanges) across various industries during the period 1989–2002. We do not impose any selection restriction on the size or industry characteristics when forming bankrupt and non-bankrupt samples. The following steps are used to identify bankrupt[1] firms. First, bankrupt firms are identified through Compustat and Lexis–Nexis Bankruptcy Report databases. Next, bankruptcy filing dates are identified through searching the Lexis–Nexis Bankruptcy Report library, Lexis–Nexis News, and firms' Form 8-K reports. Firms without available bankruptcy filing dates are eliminated. For each bankrupt firm, the most recent annual report filed prior to its bankruptcy filing date is identified. The lag between the

[1] The bankrupt sample in this study consists of firms that file bankruptcy petitions under both Chapters 11 and 7.

Table 2
Descriptive statistics

Variable	N		Mean		Median		Std. dev.		Minimum		Maximum		Test of means (proportions) difference
	NB	*B*	*NB*	B	*NB*	*B*	*NB*	*B*	*NB*	*B*	*NB*	*B*	*T*-test
Continuous variables													
TA	5887	871	−1.322	−1.192	−1.448	−1.181	2.516	1.932	−12.200	−7.142	7.707	5.522	−1.46
W	4877	767	−0.098	−0.145	0.227	0.018	6.583	0.890	−272.000	−15.332	0.995	0.849	0.20
CR	4886	790	3.300	1.432	1.873	1.071	7.618	1.559	0.000	0.012	239.333	17.728	6.87***
OF	4782	854	−0.137	−0.318	0.091	−0.043	2.428	1.258	−91.333	−25.397	55.730	1.279	2.13*
LM	1884	593	1.730	0.612	1.186	0.345	2.862	1.300	−3.191	−3.515	63.987	10.059	9.21***
CA	4888	793	0.530	0.469	0.548	0.460	0.265	0.245	0.000	0.000	1.000	1.000	6.02***
CH	5869	867	0.175	0.101	0.075	0.034	0.225	0.162	−0.012	0.000	1.000	0.991	9.39***
TL	5882	867	0.947	1.002	0.551	0.831	7.999	1.223	0.000	0.000	331.429	24.027	−0.20
LTD	5864	866	0.236	0.304	0.088	0.187	2.215	0.388	0.000	0.000	114.286	4.297	−0.90
S	5844	857	1.027	1.307	0.817	1.040	1.085	1.779	−0.081	−0.930	27.355	39.912	−6.40***
CS	4748	780	3.729	1.997	0.445	0.373	47.075	13.714	0.000	0.000	1818.000	305.919	1.02
E	5842	748	−0.321	−0.309	0.046	−0.071	9.164	1.000	−590.125	−13.486	3.734	0.342	−0.03
NT	5860	868	−0.445	−0.521	0.017	−0.165	10.209	1.546	−602.500	−23.993	42.478	0.354	0.22
RE	5740	814	−3.629	−1.787	0.035	−0.358	100.491	8.831	−6625.500	−206.975	1.717	0.581	−0.52
CHN	5298	845	0.003	−0.315	0.048	−0.372	0.565	0.583	−1.000	−1.000	1.000	1.000	15.13***
M	4776	627	−11.133	−13.086	−11.267	−13.062	2.058	1.670	−18.818	−18.331	−4.022	−7.188	22.79***
R	4808	683	0.028	−0.582	−0.086	−0.686	0.825	0.421	−0.996	−1.000	20.395	3.274	18.96***
IFR	5997	861	0.725	1.545	0.486	1.136	0.978	1.508	0.000	0.000	12.500	16.667	−21.24***
			Proportion										*Z*-test
Dichotomous variables													
IT	5299	845	0.244	0.634	0.000	1.000	0.429	0.482	0.000	0.000	1.000	1.000	−23.06***
AU	5424	861	0.251	0.559	0.000	1.000	0.434	0.497	0.000	0.000	1.000	1.000	−18.42***

*Significant at p-value < 0.05; **Significant at p-value < 0.01; ***Significant at p-value < 0.001.

Table 3
Average annual percentage of bankruptcies by industry during the entire study period

Industry	Primary SIC code	Average industry failure rate (%)
0. Agriculture, forestry, and fisheries	0100–0999	0.78
1. Mining and construction	1000–1999 except for 1300–1399	1.10
2. Food	2000–2111	0.81
3. Textiles, printing and publishing	2200–2799	1.45
4. Chemicals	2800–2824 and 2840–2899	0.52
5. Pharmaceuticals	2830–2836	0.18
6. Extractive industries	2900–2999 and 1300–1399	0.68
7. Durable manufacturers	3000–3999, except 3570–3579, and 3670–3679	0.82
8. Computers	7370–7379, 3570–3579, and 3670–3679	0.80
9. Transportation	4000–4899	1.58
10. Utilities	4900–4999	0.74
11. Retail	5000–5999	1.94
12. Financial institutions	6000–6411	0.33
13. Insurance and real estate	6500–6999	0.18
14. Services	7000–8999, except 7370–7379	1.14
15. Other	>9000	1.13

fiscal year-end of the most recently filed annual report and bankruptcy filing date must be less than 2 years.[2] The above procedure results in 890 bankrupt firms. The non-bankrupt sample is formed as described below. First, we identify all active[3] firms available in Compustat for each sample year of 1989–2002. Then we randomly select 500 firms from the identified active-firm-pool for each sample year. Once a non-bankrupt firm is selected for a year, it is excluded from selection in later years. Thus, for 14 sample years (1989–2002), we end up with 7000 active firms as the initial non-bankrupt sample. Among these 7000 firms, 63 firms have missing information on all 20 potential predictors and are deleted. Therefore, 6937 firms are used to examine the correlations among variables. Further, another five firms have missing information on all the eight variables selected. Therefore 6932 active firms are used to train and test the developed naïve BN models.

Through our own analysis and reviewing past research (e.g., Emery and Cogger, 1982; Hopwood et al., 1989; Altman, 1968; Ohlson, 1980; Hopwood et al., 1994; Shumway, 2001; McKee and Lensberg, 2002), 20 variables[4] are identified as potential bankruptcy predictors. These variables are financial-accounting factors measuring firms' size, liquidity, leverage, turnover, and profitability, market-based factors including market capitalization and abnormal stock returns, and other factors including auditors' opinions and industry failure rate. All variables for bankrupt firms are calculated based upon the most recent available data prior to firms' bankruptcy filings. Table 1 provides definitions of all variables. Table 2 provides descriptive statistics and univariate analysis of variables. Table 3 describes the average annual industry failure rate[5] during the study period. The categorization of industries is based on Barth et al. (1998).

5. Research process and research results

5.1. A heuristic 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. Koller and Sahami (1996) elaborate the importance of feature (variable) selection. First, the computation time grows dramatically as the number of features increases. Secondly, over-fitting problems occurs when we attempt to apply a large number of features to limited data available. Thirdly, irrelevant and redundant features may confuse the learning algorithm and obscure the predictability of truly effective variables. Therefore, a small number of predictive variables are preferred over a very large number of variables including irrelevant and redundant ones.

[2] Similar to Begley et al. (1996), we use this requirement to ensure the data used for prediction are reasonably current.

[3] Compustat considers a firm as active as of the end of the year if it has a closing market price for December of the year.

[4] The 20 variables are not exhaustive and there are other useful bankruptcy variables we are not able to incorporate in this study. The proposed heuristic method of variable selection is applicable to any number of potential variables.

[5] When calculating annual industry failure rate, we assume that the instances of bankruptcies identified in this study represent the number of bankruptcies in the real population, and the number of active firms in Compustat represents the number of non-bankruptcies in the real population.

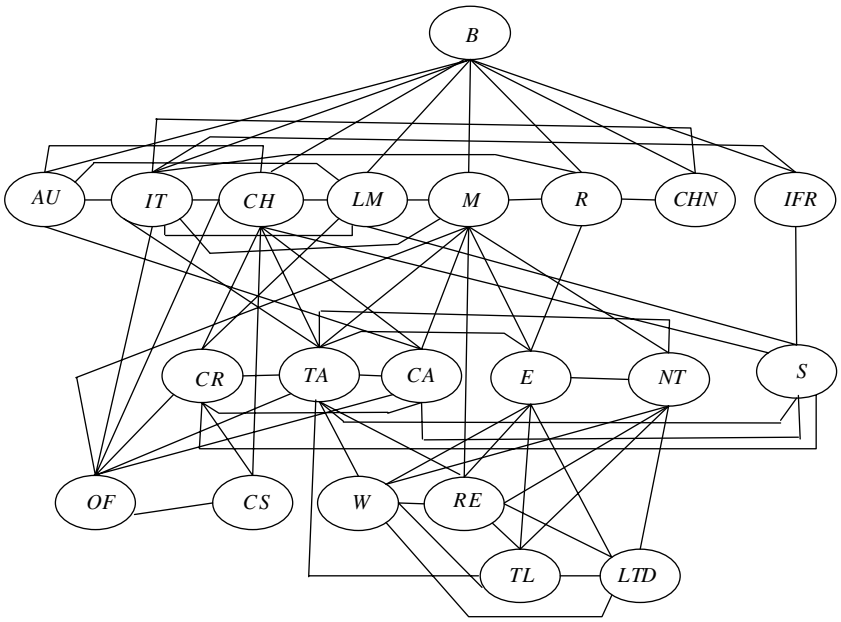

Fig. 2. Dependencies among the variables.

One purpose of this paper is to provide a heuristic method to guide the selection of variables in naïve Bayes models. Grounded on prior feature selection literature (e.g., Koller and Sahami, 1996), the goal is to eliminate variables that provide little or no additional information beyond that subsumed by the remaining variables. To achieve the goal, the proposed heuristic relies on correlations and partial correlations among variables. This heuristic is based on the assumption that the dependence between every pair of variables is linear[6] and measured by the correlation coefficient.

Next, we describe how the proposed heuristic works. First, we obtain the correlations among all variables, including 20 potential predictors and the variable of interest, firms' bankruptcy status. Variables that have significant correlations (Pearson correlation coefficient $\geqslant 0.10$) are assumed to be dependent and therefore connected. We use the cutoff of 0.10[7] to help identify a small subset of most important predictors while excluding the unimportant ones. Ideally, only the training sample should be used to obtain the correlation coefficient information. However, this study uses a 10-fold analysis that requires 10 training samples. It is too time-consuming to analyze the correlation coefficients among all the 10 training sample. Therefore, the correlations are obtained using the entire (both training and test) sample of 7827 firms, including 6937 non-bankruptcies and 890 bankruptcies. Fig. 2 shows the dependencies among the variables.

In Fig. 2, eight predictors, *LM*, *CHN*, *IT*, *M*, *AU*, *R*, *IFR*, and *CH* are connected with *B* (bankruptcy status), since they have dependency (correlations $\geqslant 0.10$) with *B*. Among these eight predictors, thirteen pairs of variables have dependency (correlations $\geqslant 0.10$) within the pair. To avoid double counting information, we analyze whether one variable is dependent with *B* given the other variable in the pair by examining the partial correlations between that variable and *B*, while controlling the other variable in the pair. These partial correlations are presented in Panel A of Table 4.

In Panel A of Table 4, Pair 1 is *CH* and *LM*. The significant partial correlation between *B* and *CH* (−0.10) given *LM* suggests that *CH* has incremental contribution in predicting *B* beyond *LM*; the significant partial correlation between *B* and *LM* (−0.13) given *CH* suggests that *LM* has incremental contribution in predicting *B* beyond *CH*. Therefore, both *CH* and *LM* are kept in the model. Similarly, within each of the rest 12 pairs, one variable has incremen-

[6] Note that this assumption is imposed by the proposed heuristic method for variable selection, not by the BN model itself.

[7] Since there is no established cutoff, we experimented with cutoffs of 0.05, 0.1, 0.15, and 0.2. The cutoff of 0.1 is the one that leads to the best prediction ability, with the least number of variables. The choice of an optimal cutoff is itself a research issue, which is not covered in this paper.

Table 4
Partial correlations

Pair no.	Partial correlation	
Panel A: Selection of first-order variables		
1	$\mathrm{Corr}(B, CH\|LM) = -0.10$,	$\mathrm{Corr}(B, LM\|CH) = -0.13$;
2	$\mathrm{Corr}(B, CH\|IT) = -0.19$,	$\mathrm{Corr}(B, IT\|CH) = 0.33$;
3	$\mathrm{Corr}(B, CH\|AU) = -0.11$,	$\mathrm{Corr}(B, AU\|CH) = 0.22$;
4	$\mathrm{Corr}(B, LM\|M) = -0.13$,	$\mathrm{Corr}(B, M\|LM) = -0.31$;
5	$\mathrm{Corr}(B, LM\|AU) = -0.13$,	$\mathrm{Corr}(B, AU\|LM) = -0.30$;
6	$\mathrm{Corr}(B, IT\|CHN) = 0.28$,	$\mathrm{Corr}(B, CHN\|IT) = -0.17$;
7	$\mathrm{Corr}(B, IT\|IFR) = 0.28$,	$\mathrm{Corr}(B, IFR\|IT) = 0.22$;
8	$\mathrm{Corr}(B, IT\|R) = 0.29$,	$\mathrm{Corr}(B, R\|IT) = -0.22$;
9	$\mathrm{Corr}(B, IT\|M) = 0.22$,	$\mathrm{Corr}(B, M\|IT) = -0.22$;
10	$\mathrm{Corr}(B, IT\|LM) = 0.32$,	$\mathrm{Corr}(B, LM\|IT) = -0.12$;
11	$\mathrm{Corr}(B, IT\|AU) = 0.25$,	$\mathrm{Corr}(B, AU\|IT) = 0.19$;
12	$\mathrm{Corr}(B, M\|R) = -0.26$,	$\mathrm{Corr}(B, R\|M) = -0.19$;
13	$\mathrm{Corr}(B, CHN\|R) = -0.22$,	$\mathrm{Corr}(B, CHN\|R) = -0.14$;
Panel B: Selection of second-order variables		
	B.1: Selection of second-order variables for CH	
14	$\mathrm{Corr}(CH, OF\|CR) = -0.04$,	$\mathrm{Corr}(CH, CR\|OF) = 0.34$;
15	$\mathrm{Corr}(CH, OF\|CS) = -0.22$,	$\mathrm{Corr}(CH, CS\|OF)$=0.12;
16	$\mathrm{Corr}(CH, OF\|TA) = -0.18$,	$\mathrm{Corr}(CH, TA\|OF) = -0.25$;
17	$\mathrm{Corr}(CH, CR\|CS) = 0.42$,	$\mathrm{Corr}(CH, CS\|CR) = 0.08$;
18	$\mathrm{Corr}(CH, CR\|TA) = 0.38$,	$\mathrm{Corr}(CH, TA\|CR) = -0.23$;
19	$\mathrm{Corr}(CH, CR\|CA) = 0.34$,	$\mathrm{Corr}(CH, CA\|CR) = 0.52$;
20	$\mathrm{Corr}(CH, S\|CR) = -0.16$,	$\mathrm{Corr}(CH, CR\|S) = 0.38$;
21	$\mathrm{Corr}(CH, TA\|CA) = -0.11$,	$\mathrm{Corr}(CH, CA\|TA) = 0.52$;
22	$\mathrm{Corr}(CH, TA\|S) = -0.31$,	$\mathrm{Corr}(CH, S\|TA) = -0.21$;
23	$\mathrm{Corr}(CH, CA\|S) = 0.63$,	$\mathrm{Corr}(CH, S\|CA) = -0.41$;
	B.2: Selection of second-order variables for LM	
24	$\mathrm{Corr}(LM, CR\|S) = 0.31$,	$\mathrm{Corr}(LM, S\|CR) = -0.03$;
	B.3: Selection of second-order variables for IT	
25	$\mathrm{Corr}(IT, OF\|TA) = -0.04$,	$\mathrm{Corr}(IT, TA\|OF) = 0.06$;
	B.4: Selection of second-order variables for M	
26	$\mathrm{Corr}(M, OF\|TA) = 0.02$,	$\mathrm{Corr}(M, TA\|OF) = 0.72$;
27	$\mathrm{Corr}(M, OF\|CA) = 0.14$,	$\mathrm{Corr}(M, CA\|OF) = -0.15$;
28	$\mathrm{Corr}(M, TA\|CA) = 0.72$,	$\mathrm{Corr}(M, CA\|TA) = 0.16$;
29	$\mathrm{Corr}(M, TA\|E) = 0.69$,	$\mathrm{Corr}(M, E\|TA) = 0.11$;
30	$\mathrm{Corr}(M, TA\|NT) = 0.69$,	$\mathrm{Corr}(M, NT\|TA) = 0.12$;
31	$\mathrm{Corr}(M, TA\|RE) = 0.68$,	$\mathrm{Corr}(M, RE\|TA) = 0.03$;
32	$\mathrm{Corr}(M, E\|NT) = 0.15$,	$\mathrm{Corr}(M, NT\|E) = 0.02$;
33	$\mathrm{Corr}(M, E\|RE) = 0.16$,	$\mathrm{Corr}(M, RE\|E) = 0.10$;
34	$\mathrm{Corr}(M, RE\|NT) = 0.14$,	$\mathrm{Corr}\ (M, NT\|RE) = 0.10$.

tal contribution in predicting B given the other variable in the pair because all partial correlations are significant (correlations $\geqslant 0.10$). Therefore, no variable is eliminated. The structure of the naïve Bayes with the eight selected variables, LM, CHN, IT, M, AU, R, IFR, and CH is shown in Fig. 3. The model consists of financial-accounting factors, market variables, auditors' opinions, and industry failure rate.

The naïve Bayes model is typically used with discrete-valued data. Prior research (e.g., Sarkar and Sriram, 2001) has used bracket median method for discretization, which divides the continuous cumulative probability distribution into n equally probable intervals. For the demonstration of the proposed heuristic method, we adapt the extended Pearson–Tukey (EP–T) method (Keefer and Bodily, 1983), a method of three-point approximations, to convert continuous variables into discrete. Under the EP–T method, a continuous distribution is approximated by a discrete distribution with probabilities 0.185, 0.63, and 0.185. Compared to bracket median method, the EP–T is able to better capture the tails of continuous variables. This feature is very suitable for the context of bankruptcy prediction since soon-to-be bankrupt firms tend to have values at the tails of the distributions (e.g. unusually high profit (McKee and Lensberg, 2002), unusually high leverage, unusually low cash flow, etc.). Besides, according to Keefer (1994), the EP–T method is one of those three-point discrete-distribution approximations that accurately represent certainty equivalents for continuous random variables.

To stay in the sample for training and testing the naïve BN model, a firm needs to have at least one variable available among the eight selected children nodes. Thus, the maximum sample size for this stage of the study is 7822, including 6932 non-bankruptcies and 890 bankruptcies. Ideally, only data in the training sample should be used to identify the cutoff points. However, this ideal procedure requires a lot of repetitive work given the 10 training samples

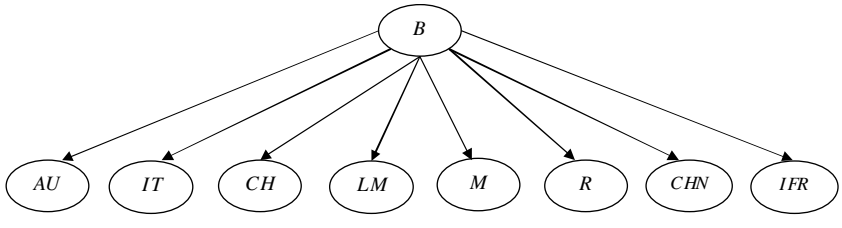

Fig. 3. The structure of the naïve Bayes model.

used under the 10-fold analysis. Therefore, for each child node of the naïve Bayes Model in Fig. 3, we use the entire (including both training and test) sample of 7822 firms to identify two points, x_1, x_2, which are respectively at 18.5 percentile, and 81.5 (18.5 + 63) percentile. These two points, x_1 and x_2, are used as cutoffs to determine to which status, 'High', 'Medium', or 'Low', a certain variable value should belong. A firm is assigned a status of 'Low' for the variable if it has a value smaller than x_1; a status of 'Medium' if it has a value between x_1 and x_2, and a status of 'High' if it has a value larger than x_2. Since we estimate two conditional distributions for each predictor variable, one conditioned on bankruptcy and one conditioned on non-bankruptcy, there is no bias introduced by the fact that the sample proportion of bankruptcies (11.4%) in this study is larger than the population proportion of bankruptcies.

To make the test results more robust, a 10-fold analysis is employed. This means that the entire sample (including bankrupt and non-bankrupt sample) is divided randomly into 10 equal sized subsets. Each time, nine subsets are randomly selected to form the training sample to learn the probabilities parameters; the remaining subset is used as the test sample to test the model's performance. On an average, each training sample consists of 801 bankruptcies and 6239 non-bankruptcies; each test sample consists of 89 bankruptcies and 693 non-bankruptcies.

Models with probabilities parameters learned from training samples are used to predict the status of bankruptcy for firms in the test sample. When testing the model, we ignore the prior since the sample proportion of bankruptcies is larger than the population proportion (e.g., McKee and Greenstein, 2000). If a firm's posterior likelihood of bankruptcy given values of observed predictors is larger than 1, it is predicted as bankrupt, otherwise non-bankrupt. The predicted results are checked for accuracy with actual statuses of bankruptcy. Table 5 reports models' prediction ability in 10 test samples.

On an average, the naïve BN model with eight selected variables accurately predicts 81.12% of bankruptcies, and 81.85% of non-bankruptcies. For comparison, we also obtain the prediction ability of the naïve model with all 20 potential variables (without any selection), reported in the right hand part of Table 5. The naïve model with all variables, on an average, correctly predicts 81.57% of bankruptcies, and 81.78% of non-bankruptcies. To conduct statistical tests of significance in models' performance differences, we assume the prediction rates (of the 10-fold results) to be normally distributed with the same variance. Thus, test of significance in the average rate between models after 10-fold analysis is equivalent to testing for difference of means of normal distribution. Untabulated *T*-test results suggest that there is no significant difference in two models' performance. This indicates that our proposed heuristic of variable selection has successfully eliminated redundant and less relevant variables, and achieved an equivalent level of performance with much fewer variables.

Table 5
Prediction ability in the test sample

Set no.	The naïve Bayesian in Fig. 4 with eight selected variables		The naïve Bayesian with all 20 potential variables	
	% bpt correct	% nbpt correct	% bpt correct	% nbpt correct
1	79.78	82.25	74.16	80.38
2	86.52	81.24	84.27	81.82
3	78.65	82.56	80.90	83.29
4	87.64	83.14	85.39	83.86
5	74.16	83.69	73.03	83.98
6	79.78	81.53	87.64	82.40
7	85.39	77.06	89.89	78.35
8	79.78	82.25	80.90	81.24
9	79.78	82.40	78.65	81.10
10	79.78	82.40	80.90	81.39
Average	81.12	81.85	81.57	81.78

Appendix A presents the tables of conditional probabilities[8] underlying the naïve BN model in Fig. 3. These conditional probabilities are informative in regard to the relationships between *B* (bankruptcy status) and its predictors. For instance, the probability of having a low *M* (market capitalization) given *B* is 44%, which is much higher than that (15%) given *NB* (non-bankruptcy).

5.2. Missing information and second-order variables

Some sample firms have missing values on one or multiple children nodes used in Fig. 3. Specifically, among the entire sample of 7822 firms, 1678 firms

[8] The conditional probabilities are learned based upon each set of training sample for each fold of analysis. Conditional probabilities presented in Appendix A are learned from one set. Conditional probabilities learned from each of the other nine sets are substantially similar to those in Appendix A.

have missing value on child node *IT*; 2419 firms have missing values on *M*; 1537 firms missing on *AU*; 2331 firms missing on *R*; 964 firms missing on *IFR*; 5345 firms missing on *LM*; 1086 firms missing on *CH*, 1679 missing on *CHN*. In the following discussion, we call the eight children nodes in Fig. 3 first-order variables. Next we discuss how to identify 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. Next we explain how each first-order variable's second-order variables are selected.

To select second-order variables for *CH*, we identify those non-first-order variables that are connected to *CH* in Fig. 2. These variables have significant correlations with *CH*. Such variables include *OF*, *CR*, *CS*, *TA*, *CA*, *S*. Among these variables, there are 10 pairs of significant relationships. Next we examine the partial correlation between one variable with *CH* after controlling for the other variable in the pair (Pairs 14–23, Panel B.1 of Table 4). For the pair of *OF* and *CR* (Pair 14), *CR* has a significant ($\geqslant 0.10$) partial correlation with *CH*, given *OF*, but *OF* does not have a significant partial correlation with *CH*, given *CR*. This indicates that *OF* does not have significant incremental contribution in predicting *CH* beyond *CR*; therefore, *OF* is deleted. Similarly, for the pair of *CR* and *CS* (Pair 17), *CS* is eliminated. Partial correlations among all other pairs are significant. Therefore, no other variable is deleted. To summarize, *CH* has four second-order variables, which are, *CR*, *TA*, *CA*, *S*.

Non-first-order variables that have significant correlations with *LM* include *CR*, and *S*. Since there is a significant correlation between *CR* and *S*, we obtain the partial correlation between *LM* and *CR* (*S*) given *S* (*CR*) (Pair 24, Panel B.3 of Table 4). *CR* has a significant partial correlation with *LM*, given *S*, but *S* does not have a significant partial correlation with *LM* after controlling for *CR*. Therefore, *S* is deleted. *LM*'s second-order variable is *CR*. Non-first-order variables that have significant correlations with *IT* include *TA*, and *OF*. Since there is a significant correlation between *TA* and *OF*, next we examine the relevant partial correlations (Pair 25, Panel B.3 of Table 4). Neither of the partial correlations is significant. In such case, the one with the higher partial correlation is selected while the one with the lower partial correlation is deleted. Therefore, *OF* is deleted. *IT*'s second-order variable is *TA*.

Non-first-order variables that have significant correlations with *M* include *OF*, *TA*, *CA*, *E*, *NT*, and *RE*. There exist significant correlations between *TA* and *OF*, *OF* and *CA*, *TA* and *CA*, *TA* and *E*, *TA* and *NT*, *TA* and *RE*, *E* and *NT*, *RE* and *E*, *RE* and *NT*. To avoid double counting information, next we examine the partial correlation between one variable with *M* after controlling for the other variable in the pair. The partial correlations are presented in Panel B.4 (Pairs 26–34). For the pair of *OF* and *TA* (Pair 26), *TA* has a significant ($\geqslant 0.10$) partial correlation with *M*, given *OF*, but *OF* does not have a significant partial correlation with *M*, given *TA*. Therefore *TA* is selected, while *OF* is deleted. Similarly, for the pair of *RE* and *TA* (Pair 31), *RE* is deleted; for the pair of *NT* and *E* (Pair 32), *NT* is eliminated. Partial correlations among other pairs are all significant, which does not suggest elimination of any other variable. To summarize, *M*'s second-order variables include *TA*, *CA*, *E*.

Only one non-first-order variable, *E*, has a significant correlation with *R*. Therefore, *E* is the second-order variable for *R*. Similarly, *S* is the second-order variable for *IFR*; *CA* is the second-order variable for *AU*. There are no non-first-order variables that have significant correlations with *CHN*. Therefore, *CHN* has no second-order variables. By incorporating selected second-order variables into the naïve Bayes model in Fig. 3, we form the following cascaded naïve Bayes model shown in Fig. 4.

Using the 10-fold analysis, we obtain the average prediction performance of the cascaded naïve Bayes model as presented in Panel A of Table 6. The cascaded BN model accurately predicts 81.12% of bankruptcies and 80.08% of non-bankruptcies. *T*-test results suggest that, compared to the naïve model with only eight first-order variables, the cascaded model has indifferent prediction ability in predicting bankruptcy, but has a significantly ($p < 0.05$) worse performance in predicting non-bankruptcy.

It is possible that for the full sample, the above comparison result is affected by those instances for which missing information on first-order variables is not that much. Take an extremely case in which the sample has complete information on all eight

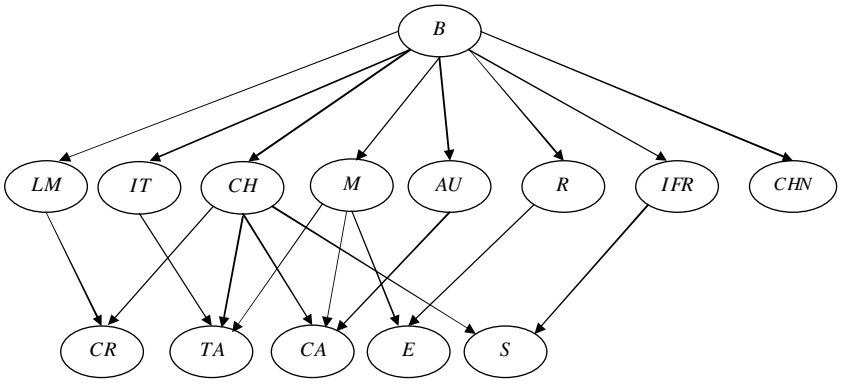

Fig. 4. The structure for the cascaded naïve Bayes model with first and second-order variables.

Table 6
Prediction ability in the test sample for the cascaded naïve Bayes model in Fig. 4

	% bpt correct	% nbpt correct
Panel A: Average performance in 10-fold analysis using the full sample		
Cascaded	81.12%	80.08%
Naïve	81.12%	81.85%
T-test	0.000	2.101*
Panel B: Average performance in 5-fold analysis using the sample with two or more missing values on first-order variables		
Cascaded	77.74%	81.09%
Naïve	77.36%	83.80%
T-test	0.081	1.860*

* Significant at $p < 0.05$.

first-order variables, the performance of the cascaded model would be identical to the naïve model because adding second-order variables does not make any difference given the Markov properties of BNs. To better examine whether the cascaded model is able to compensate for missing values on first-order variables, next, we identify firms with missing values on at least two first-order variables[9] and redo the comparison between the cascaded model and the naïve model. In our sample, 265 bankruptcies and 3501 non-bankruptcies qualify for such a selection. Given the reduced sample size, we perform 5-fold analysis instead of 10-fold analysis here. *T*-test results (Panel B of Table 6) suggest that, compared to the naïve model, the cascaded model performs the same in predicting bankruptcy, while still performs significantly ($p < 0.05$) worse in predicting non-bankruptcy. Overall speaking, we do not observe significant improvement on the model's performance after adding the second-order variables. From this perspective, the naïve Bayes model presented in Fig. 3 becomes more appealing with fewer variables and equivalent performance. Nevertheless, our results do not deny the possible superiority of the cascaded model over the naïve model in situations where missing information on first-order variables are even more substantial.

5.3. Number of states for discretization

Bankruptcy prediction often involves continuous random variables. To apply these continuous variables to BN models, past research usually employs a discretization approach (Sarkar and Sriram, 2001). This approach converts continuous variables into discrete variables with limited states, often two. During the discretization process, one problem that researchers face is to decide the number of states for discretization. Does the number of states chosen for discretization impact models' prediction power? In this study, we empirically examine this issue. The advantage of increasing the number of states is to reduce the information loss during the discretization process. However, more states require more parameters to define models. Unless one has either data or knowledge to estimate these parameters, one can easily succumb to over-fitting resulting in degradation in performance.

We use the naïve Bayes model in Fig. 3 to test the effect of discretization states. In the naïve Bayes

[9] The verification would be more appropriate if we select firms with missing values on more than two first-order variables, for instance 3, or more. However, this is not doable in our sample because the number of bankrupt firms in our sample which have missing values on at least three first-order variables is very few.

Table 7
The effect of number of states for discretizing continuous variables

# States for discretization	Average performance in the 10-fold analysis	
	% bpt correct	% non-bpt correct
2	82.58	77.55
3	83.37	77.44
4	83.82	74.94
5	83.37	75.45
6	83.15	73.83
7	82.25	75.13
8	82.36	72.36
9	81.57	71.44
10	80.67	69.46

model, six continuous variables, *M*, *R*, *IFR*, *CH*, *LM*, *CHN*, are discretized into various states, from $2,3,4,\ldots,10$. Since bankrupt firms tend to have extreme values that reside in the tails of distributions, we use the following $n-1$ points: $\frac{1}{n+1}$, $\frac{2}{n+1},\frac{3}{n+1},\ldots,\frac{n-1}{n+1}$ or $\frac{2}{n+1},\frac{3}{n+1},\frac{4}{n+1},\ldots,\frac{n}{n+1}$[10] to discretize continuous variables into *n* states. The model's performance with continuous variables discretized into various states is tested using the 10-fold analysis. Table 7 presents the model's average performance in the 10 test samples.

When continuous variables are discretized into two states, the model's accuracy in predicting bankruptcy is 82.58%, and its accuracy in predicting non-bankruptcy is 77.55%. When the number of discretization states increases to 3, the model's accuracy in predicting bankruptcy is 83.37% and its accuracy in predicting non-bankruptcy is 77.44%. Untabulated *T*-test results suggest that there is no significant difference in the model's performance between two and three states. When the number of states increases to 4, the model's accuracy in predicting bankruptcy is 83.82%, which is statistically indifferent to the model's performance with two or three states. However, the model's accuracy in predicting non-bankruptcy is decreased to 74.94%, which is significantly ($p < 0.001$) worse than that with two or three states. When we increase the number of states for discretization further, the model's performance continues to drop. With the 10 states of discretization, the model's accuracy in predicting bankruptcy is decreased to 80.67% (insignificantly different from that with two or three states), and its accuracy in predicting non-bankruptcy is decreased to only 69.46% (significantly ($p < 0.001$) worse than that with two or three states). To summarize, using a large training sample (on average 801 bankruptcies and 6239 non-bankruptcies) and a naïve Bayes model in which six out of eight predictor variables are continuous, we find that discretizing continuous variables into two or three states leads to the best performance. One possible interpretation of this finding is that bankrupt firms tend to have extreme values at one end of the distributions, while non-bankrupt firms tend to have extreme values at the opposite end. Two or three states are sufficient enough to capture the distinction. With four or more states, the model's performance significantly deteriorates, probably due to over-fitting.

5.4. Modeling continuous variables with probability density functions

The discretization of continuous variables has been criticized by researchers (Poland and Shachter, 1993). For instance, Miller and Rice (1983) and Keefer (1992) note that representing continuous distributions accurately with a few points is tricky if the tails of the distributions are significant. Next instead of discretizing continuous variables (*M*, *R*, *LM*, *CH*, *CHN*), we fit them using the normal distribution to see whether the prediction ability of the naïve Bayes model in Fig. 3 can be improved. Note that we choose to discretize *IFR* here because the goodness-of-fit of the normal distribution for this variable is too low. One possible reason for the low goodness-of-fit is that, different from other variables that are firm specific, industry failure rate, *IFR*, is an industry level factor. Again, the 10-fold analysis is used here. For each fold, we use the training sample to estimate the parameters (mean and standard deviation) of the normal distributions modeling continuous variables. The probability density function for each continuous variable given bankruptcy (*B*) and that given non-bankruptcy (*NB*) are then used to calculate the likelihood of bankruptcy given values of variables. Assuming that the prior of bankruptcy is unknown $\left(\frac{P(B)}{P(NB)}=1\right)$, the posterior likelihood of bankruptcy is calculated as

[10] If bankrupt firms tend to have extreme values at the left tails of variables, $\frac{1}{n+1},\frac{2}{n+1},\frac{3}{n+1},\ldots,\frac{n-1}{n+1}$ are used to discretize them (M,R,CH,LM,CHN), while if bankrupt firms tend to have extreme values at the right tails of the variable (*IFR*), $\frac{2}{n+1},\frac{3}{n+1},\frac{4}{n+1},\ldots,\frac{n}{n+1}$ are used to discretize such variable.

$$Likelihood\_Ratio\left(\frac{B|AU,IT,IFR,M,R,LM,CH,CHN}{NB|AU,IT,IFR,M,R,LM,CH,CHN}\right)$$

$$=\frac{P(AU|B)}{P(AU|NB)}\times\frac{P(IT|B)}{P(IT|NB)}\times\frac{P(IFR|B)}{P(IFR|NB)}$$

$$\times\frac{f(M|B)}{f(M|NB)}\times\frac{f(R|B)}{f(R|NB)}\times\frac{f(LM|B)}{f(LM|NB)}$$

$$\times\frac{f(CH|B)}{f(CH|NB)}\times\frac{f(CHN|B)}{f(CHN|NB)}.$$

The right column of Table 8 presents the 10-fold analysis result when modeling five continuous variables (*M*, *R*, *LM*, *CH*, *CHN*) using normal distributions. For comparison purposes, the left column of Table 8 shows the model's performance when continuous variables are discretized into three discrete states under the EP–T method. Untabulated *T*-test results suggest that compared to discretizing continuous variables into three states, modeling them with normal distributions leads to a statistically indifferent performance in predicting bankruptcy (83.60% vs. 81.13%), but a statistically significantly ($p < 0.001$) worse performance in predicting non-bankruptcy (77.51% vs. 81.85%). One possible explanation for this finding is that the normal distribution does not represent the underlying distributions of empirical data very well because financial ratios tend to be skewed (e.g., Karels and Prakash, 1987). We also experiment to identify and use the best-fit distributions for continuous variables using Crystal Ball software.[11] The results are substantially similar to those using the normal distribution. Again, it is possible that even the best-fit distributions do not represent the underlying distribution of the real world data very well. This finding provides some justification for discretizing continuous variables in the context of bankruptcy prediction.

Table 8
The effect of fitting continuous variables using normal distribution

Test set #	Discretizing continuous variables		Fitting continuous variables using normal distribution	
	% bpt correct	% nbpt correct	% bpt correct	% nbpt correct
1	79.78	82.25	79.78	78.64
2	86.52	81.24	84.27	75.47
3	78.65	82.56	86.52	78.53
4	87.64	83.14	89.89	77.81
5	74.16	83.69	83.15	80.09
6	79.78	81.53	80.90	76.62
7	85.39	77.06	86.52	71.28
8	79.78	82.25	82.02	78.79
9	79.78	82.40	82.02	78.21
10	79.78	82.40	80.90	79.65
Average	81.13	81.85	83.60	77.51

5.5. Naïve Bayes vs. logistic regression

In this section, we compare the performance of the naïve Bayes model in Fig. 3 with that of logistic regression, a widely used bankruptcy prediction tool. Since logistic regression is not applicable to observations with missing data unless proper techniques are used to estimate the missing values, this comparison[12] is restricted to firms with complete information on the eight predictors in Fig. 3. Thus, the study sample is reduced to 414 bankruptcies and 1435 non-bankruptcies. Given the small sample size, a 5-fold analysis is performed. Using the same eight variables presented in Fig. 3, logistic regression has an average prediction rate of 79.48% in bankruptcy sample, and 82.02% in non-bankruptcy sample. The naïve Bayes model in Fig. 3 has an average prediction rate of 80.43% in bankruptcy sample, and 80.00% in non-bankruptcy sample. Untabulated *T*-tests suggest that there is no significant difference (at the 5% level of significance) between two models'

[11] Crystal Ball software selects the following distributions for our experiment: Normal, Inverse Gaussian, Pareto, and Error Function, among a potential pool of 14 distributions, including Beta, Exponential, Extreme Value, Logistic, Log-Logistic, Log-normal, Pearson Type V, Triangular, Uniform, and Weibull.

[12] We also experiment the stepwise logistic regression at a selection criterion of $p = 0.05$ (Jones, 1987). In order to enter the stepwise logistic regression, a sample firm needs to have complete information on all 20 potential predictors used in this study. This restriction further reduces the sample to 304 bankruptcies and 1151 non-bankruptcies. The stepwise logistic regression is estimated as

$$y = -11.678 + 0.881AU + 0.661IT - 0.219LM - 1.452R + 0.205IFR - 0.728M + 1.377TL - 0.564E + 0.701TA,$$

where $y = \ln\frac{B}{1-B}$. Stepwise logistic regression selects nine predictors, six of which are the same as those used in the naïve Bayes model in Fig. 3. Based upon a 5-fold analysis, logistic regression has an average prediction rate of 84.20% in bankruptcy sample, and 84.10% in non-bankruptcy sample. For the same sample of firms, the naïve Bayes model has a prediction rate of 81.90% and 80.20%. Untabulated *T*-tests suggest that there is no significant difference (at the 5% level of significance) between two models' performance.

performance. The estimation[13] of logistic regression is as follows:

$$y = -4.755 + 0.933AU + 1.098IT - 3.165CH - 0.056LM - 2.222R - 0.391CHN + 0.356IFR - 0.156M,$$

where $y = \ln \frac{B}{1-B}$.

It is important to note that the naïve Bayes model is able to achieve an equivalent level of performance in a sub-sample of firms with missing data (see Panel B of Table 6), to which logistic regression is not applicable unless certain techniques of filling missing data is employed.

6. Summary and conclusions

In this study, we examine several important methodological issues related to the use of naïve Bayes Bayesian network (BN) models to predict bankruptcy. None of these issues have been studied by existing literature. First, we provide a heuristic method that guides the selection of predictor variables from a pool of potential variables. This method is very easy to implement and proves to be effective by the empirical results. Under this method, only variables that have significant correlations with the variable of interest, the status of bankruptcy, are selected. As a result, eight variables are selected from a pool of 20 potential predictors. Based on a 10-fold analysis, the naïve BN consisting of these eight selected variables have an average prediction accuracy of 81.12% for the bankruptcy sample and 81.85% for the non-bankruptcy sample. This prediction accuracy is appealing given the difficult nature of bankruptcy prediction and is comparable to results reported by some other studies (e.g. Ohlson, 1980; Hopwood et al., 1994; McKee and Greenstein, 2000; McKee and Lensberg, 2002) in this domain (see Table 9).

Bankruptcy prediction often involves incomplete information on some predictors. We further discuss how to select second-order variables that can compensate for missing information on selected predictors. Our empirical evidence does not show a significant improvement upon models' performance by incorporating second-order variables. Similar results are observed even after we restrict sample firms to those with at least two first-order variables missing. Nevertheless, our results do not deny the possible superiority of the cascaded model over the naïve model in situations where missing information on first-order variables are even more substantial.

Table 9
Bankruptcy prediction accuracy rates reported in some prior studies

Study	% bpt correct	% nbpt correct
Ohlson (1980)	87.6	82.6
Hopwood et al. (1994) [cost ratio of 50:1]	70.3	83.3
McKee and Greenstein (2000)	85	
McKee and Lensberg (2002)	80.3	

Second, we investigate the impact on a naïve Bayes model's performance of the number of states into which continuous variables are discretized. The naïve Bayes model consists of eight variables, six of which are continuous. Using an average training sample size of 801 bankruptcies and 6239 non-bankruptcies, we find that the model's performance is the best with the six continuous variables being discretized into two or three states. When the number of states is increased to four or more, the model's performance deteriorates, probably due to over-fitting.

Finally, we compare the performance of the naïve Bayes model with continuous variables being discretized and the performance of the model with continuous variables being modeled with normal distributions. Our results show that replacing discretization with probability density functions does not increase the model's performance. On the contrary, modeling continuous variables with normal distributions leads to a significant decrease in predicting non-bankruptcy sample. We also experiment to identify and use the best-fit distributions for continuous variables. The results are substantially similar to those using the normal distribution. One potential explanation is that normal distributions (or even the best-fit distributions) do not represent variables' underlying distributions very well.

More importantly, the above reported results could also be applicable to contexts other than bankruptcy prediction. Of course, the study has its limitations, some of which imply the need for additional research. Based upon this study's results, we can conclude that our proposed heuristic for variable selection is simple to implement and performs well. However, this study does not examine the relative performance of the proposed heuristic

[13] Estimations of coefficients reported here are the averages of coefficient values in five regressions obtained in 5-fold analysis.

compared to other correlation-based algorithm (e.g. Hall, 1999). This is a limitation of our paper which desires some future research. This study adapts the extended Pearson–Tukey (EP–T) method (Keefer and Bodily, 1983), a method of three-point approximations, to convert continuous variables into discrete. According to Keefer (1994), the EP–T method is one of those three-point discrete-distribution approximations that accurately represent certainty equivalents for continuous random variables. However, we do not test the relative performance of the EP–T method compared to other discretization methods as proposed in machine learning literature (e.g., Fayyad and Irani, 1992). Future research is useful to do such a comparison. Various variable selection algorithms have been developed/utilized for other bankruptcy prediction techniques, such as genetic algorithms for neural networks (Back et al., 1996). It is interesting future research to explore how these algorithms can be applied into BN models.

In addition, the sample proportion of bankruptcies used in this study is larger than the realistic population proportion of bankruptcies, which leads to the ignorance of the prior during our study process. 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, such as noisy-OR (Vomlel, 2003), to bankruptcy prediction.

Acknowledgements

We are grateful for Roman Slowinski, the editor, and five anonymous referees for their constructive comments on earlier versions of this paper.

Appendix A. Conditional probabilities underlying the naïve Bayes model in Fig. 3

	M				*AU*	
	Low	Medium	High		1	0
B	44%	54%	2%	*B*	56%	44%
NB	15%	65%	20%	*NB*	25%	75%

	CH				*IT*	
	Low	Medium	High		1	0
B	29%	63%	8%	*B*	63%	37%
NB	19%	63%	18%	*NB*	24%	76%

	CHN				*IFR*		
	Low	Medium	High		Low	Medium	High
B	35%	53%	12%	*B*	9%	69%	22%
NB	16%	64%	20%	*NB*	19%	75%	6%

	LM				*R*		
	Low	Medium	High		Low	Medium	High
B	31%	64%	5%	*B*	67%	28%	5%
NB	15%	63%	23%	*NB*	12%	68%	21%

References

Altman, E., 1968. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. The Journal of Finance 23 (September), 589–609.

Altman, E., Haldeman, R., Narayanan, P., 1977. Zeta analysis: A new model to identify bankruptcy risk of corporations. Journal of Banking and Finance 10, 29–54.

Altman, E., Marco, G., Varetto, F., 1994. Corporate distress diagnosis: Comparisons using linear discriminant analysis and neural networks. Journal of Banking and Finance 18, 505–529.

Anderson, R.D., Mackoy, R.D., Thompson, V.B., Harrell, G., 2004. A Bayesian network estimation of the service-profit chain for transport service satisfaction. Decision Sciences 35 (4), 665–690.

Barth, M., Beaver, W., Landsman, W., 1998. Relative valuation roles of equity book value and net income as a function of financial health. Journal of Accounting and Economics 25, 1–34.

Back, B., Laitinen, T., Sere, K., 1996. Neural networks and genetic algorithms for bankruptcy prediction. Expert Systems with Applications, An International Journal 11 (4), 407–413.

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

Begley, J., Ming, J., Watts, S., 1996. Bankruptcy classification errors in the 1980s: An empirical analysis of Altman's and Ohlson's models. Review of Accounting Studies 1 (4), 267–284.

Clark, P., Niblett, T., 1989. The CN2 induction algorithm. Machine Learning 3, 261–283.

Emery, G.W., Cogger, K.O., 1982. The measurement of liquidity. Journal of Accounting Research 20 (Autumn), 290–303.

Fayyad, U.M., Irani, K.B., 1992. On the handling of continuous-valued attributes in decision tree generation. Machine Learning 8, 87–102.

Hall, M.A., 1999. Correlation-based Feature Selection for Machine Learning. PhD thesis, University of Waikato.

Hopwood, W.S., Mckeown, J.C., Mutchler, J.P., 1989. A test of the incremental explanatory power of opinions qualified for consistency and uncertainty. The Accounting Review 64 (January), 28–48.

Hopwood, W.S., Mckeown, J.C., Mutchler, J.P., 1994. A reexamination of auditor versus model accuracy within the context of the going concern opinion decision. Contemporary Accounting Research 10 (Spring), 409–431.

Jones, F., 1987. Current techniques in bankruptcy prediction. Journal of Accounting Literature 6, 131–164.

Karels, G.V., Prakash, A.J., 1987. Multivariate normality and forecasting of business bankruptcy. Journal of Business Finance & Accounting 14 (4), 573–593.

Keefer, D.L., 1992. Certainty equivalents for three-point discrete-distribution approximations. Working paper, Department of Decision and Information Systems, Arizona State University, Tempe, AZ.

Keefer, D.L., 1994. Certainty equivalents for three-point discrete-distribution approximations. Management Science 40 (6), 760–773.

Keefer, D.L., Bodily, S.E., 1983. 3-Point approximations for continuous random variables. Management Science 29, 595–609.

Koller, D., Sahami, M., 1996. Toward optimal feature selection. In: Proceedings of the Thirteenth International Conference in Machine Learning. Morgan Kaufmann Publishers, San Francisco, CA, pp. 284–292.

Kononenko, I., 1990. Comparison of inductive and naive Bayesian learning approaches to automatic knowledge acquisition. In: Wielinga, B. (Ed.), Current Trends in Knowledge Acquisition. IOS Press, Amsterdam, The Netherlands.

Kotsiantis, S., Tzelepis, D., Koumanakos, E., Tampakas, V., 2005. Efficiency of machine learning techniques in bankruptcy prediction. In: 2nd International Conference on Enterprise Systems and Accounting, Thessaloniki, Greece.

Langley, P., Iba, W., Thompson, K., 1992. An analysis of Bayesian classifiers. In: Proceedings of the Tenth National Conference on Artificial Intelligence. AAAI Press, San Jose, CA, pp. 223–228.

McKee, T.E., 1998. A mathematically derived rough set model for bankruptcy prediction. In: Brown, C.E. (Ed.), Collected Papers of the Seventh Annual Research Workshop on Artificial Intelligence and Emerging Technologies in Accounting, Auditing and Tax, Artificial Intelligence/Emerging Technologies Section of the American Accounting Association.

McKee, T.E., 2003. Rough sets bankruptcy prediction models versus auditor signaling rates. Journal of Forecasting 22, 569–586.

McKee, T.E., Greenstein, M., 2000. Predicting bankruptcy using recursive partitioning and a realistically proportioned data set. Journal of Forecasting 19, 219–230.

McKee, T.E., Lensberg, T., 2002. Genetic programming and rough sets: A hybrid approach to bankruptcy classification. European Journal of Operational Research 138, 436–451.

Miller, A.C., Rice, T.R., 1983. Discrete approximations of probability distributions. Management Science 29, 352–362.

Ohlson, J.A., 1980. Financial ratios and the probabilistic prediction of bankruptcy. Journal of Accounting Research 19, 109–131.

Pazzani, M., Muramatsu, J., Billsus, D., 1996. Syskill & Webert: Identifying interesting web sites. In: Proceedings of the Thirteenth National Conference on Artificial Intelligence. AAAI Press, Portland, OR, pp. 54–61.

Pearl, J., 1988. Probabilistic Reasoning in Intelligent Systems. Morgan Kaufmann.

Poland, W.B., Shachter, R.D., 1993. Mixtures of Gaussians and minimum relative entropy techniques. In: Uncertainty in Artificial Intelligence. Proceedings of the Ninth Conference. Margan Kaufmann, San Mateo, CA, pp. 183–190.

Sarkar, S., Sriram, R.S., 2001. Bayesian models for early warning of bank failures. Management Science 47 (11), 1457–1475.

Senetti, J.T., 1995. On the incoherent use of evidence: Why subjective Bayesian evidence is not held probative. Auditing: A Journal of Practice and Theory 13, 193–199.

Shumway, T., 2001. Forecasting bankruptcy more accurately: A simple hazard model. The Journal of Business 74, 101–124.

Tam, K.Y., Kiang, M.Y., 1992. Managerial applications of neural networks: The case of bank failure predictions. Management Science 38 (7), 926–947.

Titterington, D.M., Murray, G.D., Murray, L.S., Spiegelhalter, A.M., Skene, A.M., Habbema, J.D.F., Gelpke, G.J., 1981. Comparison of discrimination techniques applied to a complex data-set of head-injured patients (with discussion). Journal of the Royal Statistical Society, Series A 144, 145–175.

Vomlel, J., 2003. Noisy-OR classifier. In: Proceedings of the Sixth Workshop on Uncertainty Processing, pp. 291–302.

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