academicJournals

Vol. 8(44), pp. 2181-2186, 25 November, 2013
DOI: 10.5897/SRE09.174
ISSN 1992-2248 © 2013 Academic Journals
http://www.academicjournals.org/SRE

Scientific Research and Essays

Full Length Research Paper

An examination of the effect of discretization on a naïve Bayes model's performance

Arezoo Aghaei Chadegani[1]* and Davood Poursina[2]

[1]Department of Accounting, Mobarakeh Branch, Islamic Azad University, Isfahan, Iran.
[2]Department of Statistics, Isfahan University, Isfahan, Iran.

Accepted 3 June, 2013

A Bayesian network (or a belief network) is a probabilistic graphical model that represents a set of variables and their probabilistic independencies. Some researches often involve continuous random variables. In order to apply these continuous variables to BN models, these variables should convert into discrete variables with limited states, often two. During the discretization process, one problem that researchers faced is to decide the number of states for discretization. Does the number of states chosen for discretization impact models' power? In this study, this issue is examined empirically. The study examines this issue in the financial distress prediction field. The sample consists of 144 firms listed in Tehran stock exchange from 1997 to 2007. In order to develop Naïve Bayes models, two methods for choosing variables were used. The first method is based upon conditional correlation between variables and the second method is based upon conditional likelihood. The accuracy in predicting financial distress 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%. Collectively, the results showed that the performance of the second naïve Bayes model that based upon conditional likelihood is better than the first one. Further analyses also showed that the number of states chosen for discretization has effect on models' performance. In comparing the model's performance when continuous variables are discretized into two, three, four and five states, the results showed that the naïve Bayes model's performance increases when the number of states for discretization increases from two to three, and from three to four but when the number of states increases from four to five the model's performance decreased.

Key words: Bayesian networks, naïve Bayes, selection of predictors, discretization, continuous variables, financial distress predictors, firms, Tehran Stock Exchange (TSE).

INTRODUCTION

Reasoning with incomplete and unreliable information is a central characteristic of decision making in some industry such as medicine and finance. Bayesian networks provide a theoretical framework for dealing with this uncertainty using an underlying graphical structure and the probability calculus. Bayesian networks have been successfully implemented in areas as diverse as medical diagnosis and finance (Holmes and Jain, 2008). In this study, Bayesian networks are used for developing two naïve Bayes models for predicting financial distress and the effect of discretization on naïve Bayes model's performance were investigated.

A Bayesian network (or a belief network) is a probabilistic graphical model that represents a set of variables and their probabilistic independencies. Naïve Bayes models work only with discrete variables and to

*Corresponding author. E-mail: arezooaghaie2001@yahoo.com. Tel: ++989131885269.

Abbreviations: EP-T, Extended Pearson-turkey; **BN,** Bayesian network; **TSE,** Tehran stock exchange; **DAG,** direct acyclic graph.

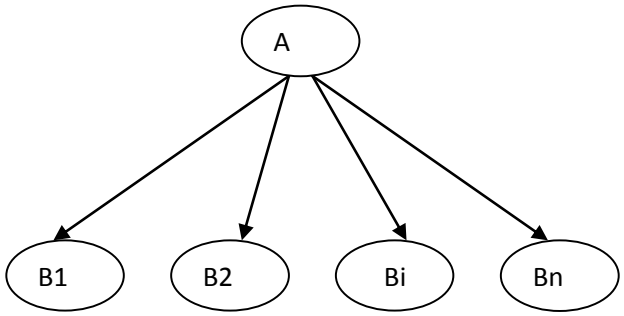

Figure 1. A naïve Bayes BN model.

apply these continuous variables to BN models, these variables should convert into discrete variables with limited states.

Bayesian networks are powerful tools both for graphical representation of the relationships among a set of variables and for dealing with uncertainties in expert systems (Pearl, 1988). BNs have become an active research in the past decade (Heckerman, 1999). Bayesian networks have been successfully applied to create consistent probabilistic representations of uncertain knowledge in diverse fields such as medical diagnosis (Spiegelhalter, 1989), image recognition (Booker and Hota, 1986), language understanding (Charniak and Goldman, 1989), search algorithms (Hansson and Mayer, 1989), and many others.

Nigam et al. (2000) and Ng and Jordan (2002) have shown that classification decision rules that are based on naïve Bayesian networks (that is, conditional independence assumption) work well in practice. The conditional independence is useful because it makes computation much more tractable (Nigam et al., 2000; Winkler, 1988),

Kyprianidou (2002) used Bayesian networks for analyzing basic genetics. 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. They used bracket median method for discretization. Sun and Shenoy (2007) developed Bayesian networks models for predicting financial distress. They adapt the Extended Pearson-Tukey (EP-T) method to convert continuous variables into discrete.

The aim of this study is to examine the effect of discretization on a Naïve Bayes Model's Performance.

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). A Bayesian Network consists of the following:

1) A set of variables and a set of directed edges between variables.
2) Each variable has a finite set of mutually exclusive states.
3) The variable together with the directed edges from a directed acyclic graph (Kyprianidou, 2002).

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 (Kyprianidou, 2002). Bayes rule can be expressed as follows:

Inference in Bayesian networks

Because a Bayesian network is a complete model for the variables and their relationships, it can be used to answer probabilistic queries about them. For example, the network can be used to find out updated knowledge of the state of a subset of variables when other variables (the evidence variables) are observed. This process of computing the posterior distribution of variables given evidence is called probabilistic inference. Indeed inference in a Bayesian networks means computing the conditional probability for some variables, given information (evidence) on other variables (Jensen, 1990). The complexity of Bayesian network inference depends on the network structure. There are several well-known methods of exact inference in Bayesian networks: variable elimination and clique tree propagation being particularly popular. The methods of approximation used mostly are: stochastic MCMC simulation and bucket elimination.

NAÏVE BAYESIAN NETWORKS 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 (Ceruti, 2002). Figure 1 presents a

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 flow/ total liabilities
X5	Current assets/ total liabilities
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 in –t)/(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

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 financial distress prediction context, in Figure 1, (A) represents the financial distress variable. B1, B2 …..Bn represent n financial distress predictor variables. The naïve Bayes model assumes the following conditional independence: the assumption says that predictors, B1, B2 …, Bn are conditionally mutually independent given the state of financial distress (Sun and Shenoy, 2007).

Discretization

The naïve Bayes model is typically used with discrete-valued data for which the research's data are continuous and they should be first discretisize. 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 on the number of states for discretization.

There are many different methods for discretization and prior research (Sarkar and Sriram, 2001) has used bracket median method for discretization, which divides the continuous cumulative probability distribution into n equal probable intervals. Sun and Shenoy (2007) adopted the extended Pearson-Tukey (EP-T) method (Keefer and Bodily, 1983), a method of three-point approximations, to convert continuous variables into discrete.

Hypotheses

In this research, the investigation carried out stated as follows: does the number of states chosen for discretization impact models' power? Then, the hypotheses of this research are:

H0: The number of states chosen for discretization has effect on the naïve Bayesian models' performance.

H1: The number of states chosen for discretization does not have effect on the naïve Bayesian models' performance.

Sample and data

Sample firms used in this study are companies that were listed in Tehran Stocks Exchange (TSE) across various industries during the period 1997 to 2007 for developing naïve Bayes models for financial distress prediction. Through analysis and reviewing of past research, 20 variables are identified as potential financial distress predictors. These variables were included in the financial ratios used in measuring firm's liquidity, leverage, turnover, profitability and firm's size and other factors like auditors' opinions. Most of the data are continuous and for developing naïve Bayes models, they should be converted first into discrete variables. The variables in this study are shown in Table 1.

RESEARCH PROCESS AND RESULTS

First method for Variable Selection in Naïve Bayes Models

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. Subsequently, a small number of predictive variables are preferred over a very large number of variables including irrelevant and redundant ones. Two different methods for selecting the variables were compared. The first method is depending upon correlations and conditional correlations among variables.

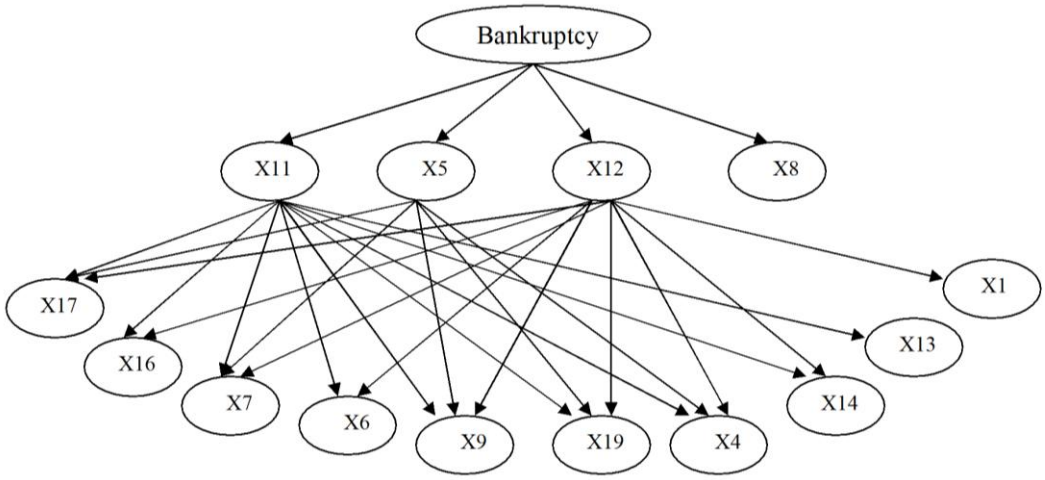

Figure 2. The structure of the first naïve Bayes model.

Table 2. The result of the discretization in first naïve Bayes model.

Firms	Discretization							
	Two states		Three states		Four states		Five states	
	1	0	1	0	1	0	1	0
Number of firms 1	60	12	59	13	65	7	58	14
Number of firms 0	19	53	11	61	8	64	7	65
1%	83	17	84	16	90	10	80	20
0%	26	74	16	84	11	89	10	90

Firms with financial distress=1; firms without financial distress=0

Following Sun and Shenoy research, first, the correlations among all variables were obtained, including 20 potential predictors and the variable of interest, and firm's financial distress status. Variables that have significant correlations were assumed to be dependent and therefore connected. At first stage, four predictors (x5, x8, x11, and x12) are connected with financial distress, since they have dependency with financial distress. These variables are first – order variable. The second-order variables were identified, that is, the variables that have effect on first – order variables. Conceptually, second-order variables are those that have significant correlations with first-order variables. To select a given first-order variable's and second-order variables, a similar method used to select first-order variables is followed. The major difference is that now a first-order variable is considered differently instead of financial distress as a root variable. For example the conditional correlation between x5 and x17, x7, x9, x19, x4 were significant so they were 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 models that were obtained.

The naïve Bayes model is typically used with discrete-valued data. Uniform Widths method to convert continuous variables into discrete were used. During the discretization process, one problem that researchers face is to decide on the number of states for discretization. The research study started from two states to five states and tested the performance of the model with all samples in the research. When continuous variables are discretized into 2 states, the model's accuracy is 83% while the number of discretization states increases to 3, the model's accuracy is 84%. Moreover, the number of states increases to 4, the model's accuracy is 90% and when the numbers of states for discretization were further increased, the model's performance continues to drop (Table 2).

Second method for variable selection in naïve Bayes models

In the second method, variables are selected based upon

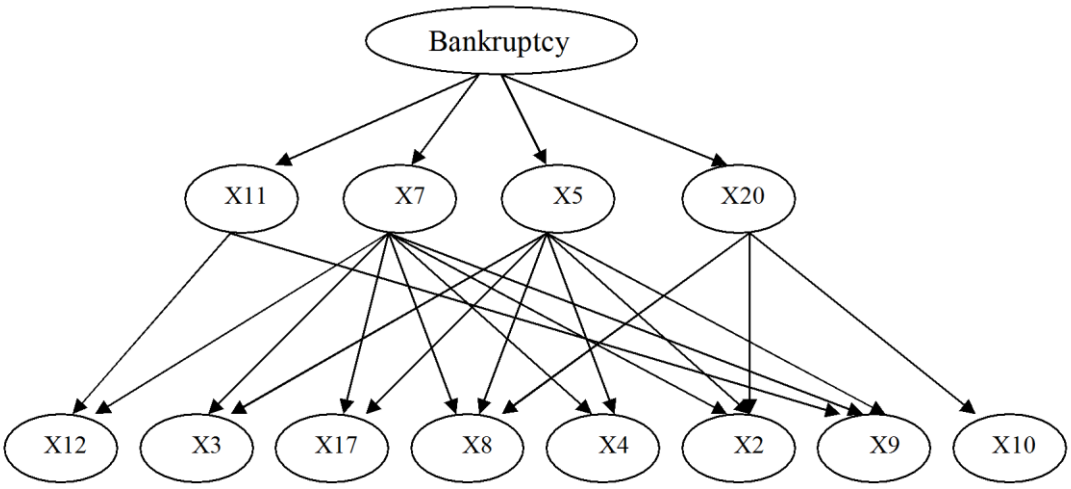

Figure 3.The structure of the second naïve Bayes model.

Table 3. The result of the discretization in second naïve Bayes model.

Firms	Discretization							
	Two states		Three states		Four states		Five states	
	1	0	1	0	1	0	1	0
Number of firms 1	63	9	66	6	68	4	62	10
Number of firms 0	14	58	9	63	6	66	14	58
1%	87	13	92	8	94	6	86	14
0%	20	80	13	87	8	92	20	80

Firms with financial distress=1; firms without financial distress=0.

conditional likelihood. First, the correlations among all variables were obtained, including 20 potential predictors and the variable of interest, firm's financial distress status. Variables that have significant correlations were assumed to be dependent and therefore connected. At the first stage, four predictors (x5, x7, x11, and x20) were connected with financial distress, since they have dependency with financial distress. These variables are first – order variables. Then, second-order variables were identified. 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 was obtained. When continuous variables were discretized into 2 states, the model's accuracy is 87%. On the other hand, when the number of discretization states increases to 3, the model's accuracy increase to 92% and when the number of states increases to 4, the model's accuracy become 94%. In addition, when the number of states for discretization increases, the model's performance continues to drop. The rates showed that this model's performance is better than the first one (Table 3).

SUMMARY AND CONCLUSIONS

In this study, naive Bayesian networks for developing two different models for predicting financial distress and examining the effect of discretization on naive Bayes models performance were introduced. First, two different methods that guide the selection of predictor variables from a pool of potential variables were provided. Under the first method, only variables that have significant correlations with the variable of interest, the status of financial distress were selected. As a result, 4 variables were selected from a pool of 20 potential predictors. Afterwards, second order variables were selected. The first naïve BN consisting of these selected variables have an average prediction accuracy of 90%. Hence, a conditional likelihood method for selecting variables and run the second naïve Bayes model were used. This model has an average prediction accuracy of 93%. Secondly, the impacts on a naïve Bayes model's performance of the number of states into which continuous variables are discretized were also investigated. It was found 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.

REFERENCES

Booke LB, Hota N (1986). Probabilistic Reasoning about Ship Images. Paper presented at the Second Annual Conference on Uncertainty in Artificial Intelligence. University of Pennsylvania. Philadelphia. PA.

Ceruti MG (2002). Establishing a Data-mining Environment for Warning Event Prediction with an Object-oriented Command and Control Database. Data Acquisition Exploitation. pp. 92-99.

Charniak E, Goldman RP (1989). Plan Recognition in Stories and in Life. Paper presented at the Fifth Workshop on Uncertainty in Artificial Intelligence. Mountain View. California.

Hansson O, Mayer A (1989). Heuristic Search as Evidential Reasoning. Paper presented at the Fifth Workshop on Uncertainty in Artificial Intelligence. Windsor. Ontario. Canada.

Heckerman D, Mamdani A, Wellman MP (1999). Real-World Applications of Bayesian Networks. Communications of the ACM. pp. 24-68.

Jensen A, Andersen K (1990). Approximations in Bayesian belief universes for knowledge-based systems. In Proceedings of the Sixth Conference on Uncertainty in Artificial Intelligence. pp. 162-169.

Kyprianidou C (2002). Analyzing Basic Genetics Using Bayesian Networks. (Ph.D) dissertation. MSc Actuarial Science Cases Business School City University.

Nigam K, McCallum AK, Thrun S, Mitchell T (2000). Text Classification from Labeled and Unlabelled Documents using EM. Mach. Learn. 39:103-134.

Pearl J (1988). Probabilistic Reasoning in Intelligent Systems. Network of Plausible Inference.

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

Spiegelhalter DJ, Franklin R, Bull K (1989). Assessment, Criticism, and Improvement of Imprecise Probabilities for a Medical Expert System. In Proceedings of the Fifth Conference on Uncertainty in Artificial Intelligence pp. 285-294.

Sun L, Shenoy P (2007). Using Bayesian Networks for Bankruptcy Prediction. Eur. J. Oper. Res. 180(2):738-753.

Winkler WE (1988). Using the EM Algorithm for Weight Computation in the Fellegi-Sunter Model of Record Linkage. Proceedings of the Section on Survey Research Methods. American Statistical Association pp. 667-671.