Focus

Soft Computing 7 (2003) 297–303 © Springer-Verlag 2003
DOI 10.1007/s00500-002-0216-4

Learning Bayesian networks using various datasources and applications to financial analysis

J. Gemela

Abstract This paper illustrates opportunities of using Bayesian networks in fundamental financial analysis. In it, we will present an application based on construction of a Bayesian network from a database of financial reports collected for the years 1993–97. We will focus on one sector of the Czech economy – engineering – presenting an example that use the constructed Bayesian network in the sector financial analysis. In addition, we will deal with the rating analysis and show how to perform this kind of analysis by means of neural and Bayesian networks.

Keywords Bayesian networks, Learning, Fundamental financial analysis, Neural networks, Engineering, Application

1
Introduction

Bayesian Belief Networks offer a mathematical model for representing uncertain information. This model includes methods for constructing (learning) networks, storing probabilistic information, and an evidence propagation scheme. Bayesian networks use graphical representations of uncertain knowledge which can be easily interpreted. First, the structure of the graph forms qualitative relationships between domain variables. Second, quantitative aspects of knowledge are represented by a set of conditional probability tables. These tables are stored in the nodes of the influence graph. For more information about Bayesian networks we refer the reader to [1] and for information about learning [2–4].

Because financial analysis is a very wide field of interest, we have chosen only one part – fundamental analysis of enterprises – for the focus of this paper. The object of the fundamental analysis is to determine an enterprise's economic situation by means of data gathered from within the enterprise itself. Technical financial analysis, on the other hand, would analyze the enterprise's position using statistical data obtained at financial markets such as stock markets, bond markets, options markets etc. This difference in data utilization is the main difference between these two types of analysis.

As previously mentioned, the fundamental financial analysis uses the financial data obtained from the enterprise itself. These data summarize the types of the enterprise's assets and give an overview of the company's basic economic activities. The data are systematically summarized in accounting ledgers and regularly published in annual reports. On the basis of these reports, financial analysts compute various financial indices which give the basic characteristics of the enterprise's economic situation. It is not surprising to note that the computed financial characteristics usually vary with the respective economic sector. For example metallurgy, energy-supply industry and chemistry are characterized by high measure of investments. On the other hand, trade and research require low investments but higher labor expenditures. Therefore, if we want to analyze a particular enterprise, we have to compare its financial characteristics with the given sector or subsector characteristics rather than with the economic averages. Comparing financial characteristics of the enterprise, one can discover potential threats and opportunities relative to its current economic position. Details can be found in [5] or [6].

2
Objectives of the application

In this paper we shall concentrate on the analysis of the financial indices of one economic sector: engineering. Our choice was motivated by its relative stability in the Czech Republic (compared to other sectors) and by the fact that we have at our disposal a sufficient number of annual reports. We have collected about 4400 reports of engineering enterprises for the years 1993–1997, from which we selected 10 basic financial indices which are typically used by bank analysts to judge an enterprise's application for a loan. Therefore, the given indices focus on that part of enterprise's finances which are supposed to be used for repayment of the loan and interest.

We can consider the selected financial indices to be continuous random variables. In general, unfortunately, these random variables do not follow standard probability distributions and there is no statistical model to represent the joint probability distribution. In the light of this we opted to categorize these random variables and transform the continuous variables to discrete ones. Observing basic dependencies among discrete random variables, one can construct a Bayesian network which can be used as a tool for analyzing a given economic sector.

J. Gemela
Laboratory for Intelligent Systems,
University of Economics, Ekonomicka 957,
CZ-148 00 Prague 4, Czech Republic
e-mail: gemela@mefisto.cz

This work was supported by the grant VS96008 of the Ministry of Education of the Czech Republic.

3
Financial indices

In this section we give a list and brief description of financial indices used in our application. We will use ratio indices, which are probably the most widely used financial analysis technique. Ratio analysis investigates relationships between two or more line items on the financial statements. It is important to emphasize that the ratio indices evaluate the enterprise position statically in the context of one year. In order to measure the development of the enterprise from the historical point of view, we also use trend indices. These indices indicate changes of selected ratio indices from year to year. Finally, since financial experts have observed some dependencies between certain ratio indices and the Owner's Equity value (which will be explained later), we also take this index into account.

3.1
Ratio indices

The used ratios (including the proposed discretization of their values) are listed in Table 1.

Now, we give a brief description of the indices:

- *Debt ratio* measures the risk of the possibility that the company will become insolvent before all creditors' claims are met. Desired levels of this ratio usually vary with the stability of company income. Generally speaking, the more stable the historical income, the greater the likelihood that investors and creditors will tolerate increased debt.
- *Liquidity* compares current assets to current liabilities. It is an indicator of a company's ability to meet its short-term obligations with current assets.
- *Return on sales* is one of profitability ratios. This ratio indicates the return a company receives for each dollar of sales.
- *Inventory turnover*. Turnover is the relationship between the amount of an asset and some measure of its use. Inventory turnover is the number of days the inventory is turning in the production process.
- *Accounts receivable turnover* provides an indication of how quickly the average amount of receivables are being collected.
- *Accounts payable turnover*, apart from accounts receivable, is the number of days in which the average amount of payables are settled.

3.2
Trend indices and owner's equity

Trend indices measure the change of a particular ratio index in two subsequent years. The indices indicate whether the financial situation is improving or declining. The computational formula is simple:

$$\text{Index trend}(year) = \text{Index}(year) - \text{Index}(year - 1) \ .$$

The owner's equity index roughly indicates the size of the company. In fact, it is the share of the business that the owner owns outright. Owner's equity consists of the owner's investment in the business plus profits made from

Table 1. List of ratio indices

Debt ratio (*DR*)	Debt/Assets	0.00–0.25 0.25–0.50 0.50–0.75 0.75–1.00
Liquidity (*L*)	Quick assets/Current liabilities	0.0–0.5 0.5–1.0 1.0–1.5 1.5–2.0 Over 2.0
Return on sales (*RS*)	Net income/net sales	Under −0.05 −0.05–0.00 0.00–0.05 0.05–1.00 Over 1.00
Inventory turnover (*IT*)	Average inventory/Cost of goods sold/360	0–30 days 30–90 90–180 180–360 Over 360
Accounts receivable turnover (*RT*)	Average net acc. receivable/Cost of goods sold/360	0–30 days 30–90 90–180 180–360 Over 360
Accounts payable turnover (*PT*)	Average net acc. payable/Cost of goods sold/360	0–30 days 30–90 90–180 180–360 Over 360

the business that the owner has not withdrawn from the business. Table 2 contains a list of used trend indices and the owner's equity index, which will be indicated with E, and the corresponding proposed discretization.

4 Constructing Bayesian network

As we have already stated, we construct a Bayesian network structure using a probabilistic learning algorithm. We will not go into details of the learning algorithm (we refer the reader to [7]). Instead, we will only present its basic properties, ideas and implementation parameters.

4.1 Basic principles

The algorithm uses the heuristic approach and systematically goes through the search set of all Bayesian network structures with respect to the ten discrete random variables. The algorithm seeks the networks with the best mixture of the following measures:

- *Accuracy*. The primary goal of the learning algorithm is to find a network which approximates the empirical probability distribution given by the data. As a measure of accuracy, the Kullback-Leibler divergence (see [8] for the definition and basic properties) is used:

$$\mathrm{Div}(\hat{P}, P_M) = \sum_x \hat{P}(x) \log \frac{\hat{P}(x)}{P_M(x)} \quad ,$$

  where $\hat{P}$ is the empirical probability distribution and P_M is the probability distribution represented by the model.
- *Complexity*. Generally, the more complex the network is, the higher the accuracy that can be achieved. Therefore we are faced with a problem of finding a tradeoff between accuracy and complexity of the network. As a measure of complexity, MDL (minimum description length) principle (see [4]) considers the space necessary to store the network in the computer memory (i.e. the number of bits needed to encode the structure of the graph and store all conditional probability tables).
- *Meeting expert's requirements*. In our application we intended to combine statistical data with an expert's knowledge. However, we have had some difficulty in explaining the notion of conditional independence to financial experts. In the end, we were given only a number of causal relationships which are supposed to hold true. These relationships were of the form "the change of the variable A influences the variable B." Thus, according to the heuristic learning approach, the algorithm prefers those networks which obey this expert knowledge.

Table 2. Trend indices and owner's equity

Debt ratio trend (*DT*) and return on sales trend (*ST*)	Under −0.10 −0.10–−0.05 −0.05–0.00 0.00–0.05 0.05–0.10 Over 0.10
Liquidity trend (*LT*)	Under −0.50 −0.50–−0.25 −0.25–0.00 0.00–0.25 0.25–0.50 Over 0.50
Equity (*E*)	0–25 000 CZK 25 000–100 000 100 000–500 000 Over 500 000

Precisely speaking, the fitness of a model M is of the following form:

$$\mu(M) = \frac{\mathrm{Div}(\hat{P}, P_M)}{\mathrm{Div}(\hat{P}, P_{S_{\min}})} + \frac{c(M) - c(S_{\min})}{c(S_{\max})} + \frac{U - u(M)}{U} \quad , \tag{1}$$

where $S_{\min}$, $S_{\max}$ denote the model without arrows and complete graph respectively, $c(.)$ is the complexity of a model, $u(.)$ is the number of the expert's arrows presented in a model and U is the total number of arrows given by the expert. Thus, the fitness is computed as a sum of the three normalized measures. Of course, one can add weights to these measures according to personal preferences. In our application we used the following weights: 0.5, 0.4, 0.1.

The heuristic algorithm is built upon an iterative approach. In each iteration it examines network structures with the same number of arrows. The algorithm uses a working list, OPEN, of models whose size is limited by a parameter value (in this application we used the value 30). In the beginning, OPEN contains only the Bayesian network with no arrows (if some arrows are required, i.e. there are some arrows representing causal relationships given by the expert, then OPEN contains the network with the required arrows only). In each iteration, the algorithm performs the following operations:

- *Generation*. Generate several networks from every network in the list OPEN, by adding just one arrow. New arrows are added between nodes with the highest value of mutual information (see [4]). The algorithm takes both directions into account. The total number of newly-generated networks is limited by a parameter of the algorithm (100).
- *Elimination*. Search for generated equivalent network structures (equivalent in terms of representing the same set of conditional independencies). If equivalent structures are detected, those with lower μ are eliminated.[1]
- *Iteration step*. Replace the networks in OPEN with the newly generated ones. Because the size of OPEN is limited, choose the best ones within those limits.

During the learning process the algorithm seeks for a Bayesian network with the lowest fitness value. This network is then reported as the result of the learning process. The algorithm stops if the maximum number of arrows is reached (for 10 variables it is in 45 iterations) or, in the case of the extended elimination process, OPEN is empty.

[1] If one does not consider the third measure (user's requirements) we can also eliminate those models for which $\mathrm{Div}(\hat{P}, P_M) = 0$.

The computing parameters (30, 100) were estimated as optimal parameters after several runs of the algorithm.

4.2 Implementation

As we have already mentioned above, we used a database of 4400 accountants' annual reports. The data come from the years 1993–97. There were no missing values in the database. Because the Czech economy has been transforming from the central planning system into the free market since 1990, the economic conditions have been dramatically changing. In order to cope with these uncertain economic conditions, we decided to assign various weights to database cases having their origin in different years. These weights can be read from Table 3.

The next step of the implementation process was to obtain some expert knowledge and exploit this knowledge in the learning process. When we asked financial experts for a list of various dependencies among the financial indices, they found difficulties in producing such a list. It is likely they had not thought of the dependencies among financial factors before, as this is probably not of essential importance for the financial theory like it is in other ones (for example in medicine, where experts systematically explore dependencies among various factors). Eventually, we managed to obtain the following two lists of causal dependencies:

- *Sure dependencies* – well observed or based on theory: $E \rightarrow DR$, $E \rightarrow RT$, $E \rightarrow IT$.
- *Uncertain dependencies* – based on an expert's experience, expected, and sometimes observed causal relationships (without any theoretical explanation): $DR \rightarrow L$, $PT \rightarrow L$, $RT \rightarrow PT$, $IT \rightarrow RS$, $DT \rightarrow ST$.

The learning algorithm considers only those networks which include "obligatory" arrows and prefers those with the "expected" ones.

Figure 1 introduces the network structure found by the probabilistic learning algorithm. Note that all "expected" dependencies are included except $PT \rightarrow L$, however, the opposite $L \rightarrow PT$ has been discovered.

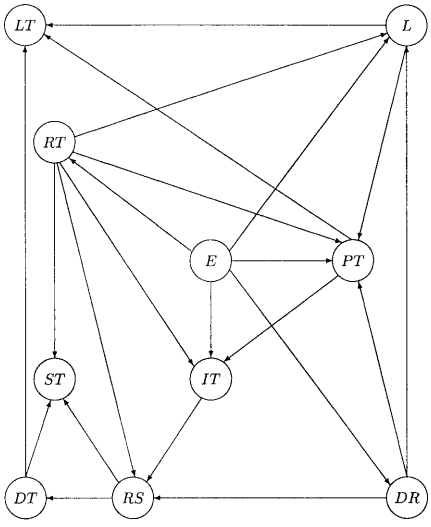

Fig. 1. Bayesian network found by the learning algorithm

5 Example

A financial expert, when evaluating the financial position of a particular company, should compare the financial characteristics of this company with the characteristics of the respective economic sector, i.e., averages received for a group of similar companies. If, for example, the expert evaluates an engineering company, he/she should compare its financial characteristics with the engineering sector averages. An example of these sector characteristics for liquidity, inventory turnover, accounts receivable turnover and accounts payable turnover is content of Table 4, which summarizes unconditional probability distributions of these four financial indices.

What can we read from the values in this table?

- *Liquidity.* If the liquidity index of a company falls within the interval 0.0–0.5 it should be regarded as too small and alternatively, values between 0.5 and 1.5 and over 2 will be considered "normal."
- *Inventory turnover* and *accounts receivable turnover.* The values of these indices will have the maximum probability if they are in the interval 30–90 days. The probability decreases if values divert from this "normal" interval.
- *Accounts Payable turnover.* Analogously, we can find out that values of accounts payable turnover are "good" in the interval 0–180 days and "bad" over 180 days.

Now, let us focus our attention on the Liquidity index, which is, in a way, most interesting. We can observe that the index is not homogeneous in the sense that the "normal" values are those in interval 0.5–1.5 and over 2.0. As the probability of the interval 1.5–2.0 is 0.10, we do not consider these values to be "normal." The existence of this "gap in normality" could be explained by a hypothesis that there are different subsectors in the engineering sector whose values of the liquidity index differ.[2] For a given company having liquidity in the interval 1.5–2.0, an expert can conclude that this value is not desirable but he/she cannot say whether the value is too high or small. Thus, it seems necessary to perform a more detailed analysis of that particular subsector of the engineering sector. Financial experts did it and observed that companies which have similar production

Table 3. Weights of years

Year	Weight
1993	1
1994	1.5
1995	2
1996	3
1997	5

[2] One may argue that the gap in normality may only be a result of inappropriate categorization. However, detailed exploration gives the probability 0.18 for the interval 2.0–2.5.

Table 4. Sector characteristics

Index	Value	Probability
Liquidity (*L*)	0.0–0.5	0.13
	0.5–1.0	0.29
	1.0–1.5	0.22
	1.5–2.0	0.10
	Over 2.0	0.26
Inventory turnover (*IT*)	0–30 days	0.18
	30–90	0.40
	90–180	0.30
	180–360	0.10
	Over 360	0.02
Accounts receivable turnover (*RT*)	0–30 days	0.13
	30–90	0.60
	90–180	0.20
	180–360	0.04
	Over 360	0.03
Accounts payable turnover (*PT*)	0–30 days	0.16
	30–90	0.48
	90–180	0.24
	180–360	0.09
	Over 360	0.03

Table 6. Subsector characteristics

Index	Value	Subs. A	Subs. B	Subs. C
Liquidity (*L*)	0.0–0.5	0.03	0.06	0.16
	0.5–1.0	0.05	0.26	0.40
	1.0–1.5	0.13	0.33	0.28
	1.5–2.0	0.13	0.19	0.08
	Over 2.0	0.66	0.16	0.08
Inventory turnover (*IT*)	0–30 days	0.14	0.17	0.20
	30–90	0.39	0.39	0.39
	90–180	0.38	0.36	0.33
	180–360	0.09	0.08	0.08
	Over 360	0.00	0.00	0.00
Accounts receivable turnover (*RT*)	0–30 days	0.11	0.11	0.12
	30–90	0.62	0.63	0.64
	90–180	0.22	0.21	0.19
	180–360	0.03	0.03	0.03
	Over 360	0.02	0.02	0.02
Accounts payable turnover (*PT*)	0–30 days	0.41	0.11	0.11
	30–90	0.50	0.58	0.47
	90–180	0.06	0.26	0.32
	180–360	0.02	0.04	0.08
	Over 360	0.01	0.01	0.02

programs have similar debt ratio and return on sales index. While debt ratio is strongly influenced by production technology used by the company, the return on sales index is given by the market position of the given company. Therefore, fixing both these indices we can identify some subsectors of the engineering sector which have similar production programs (e.g. car production, production of agricultural machines etc.).

On the other hand, the remaining three indices seem to be quite homogeneous. We can see the intervals where the indices achieve the maximum probability and the probability decreases according to the distance from the interval of "normality."

Let us consider three subsectors A, B and C, identified by various instances of the debt ratio and return on sales indices as suggested in Table 5.

Table 6 contains conditional probability distributions of the four indices within these subsectors (i.e., given the respective values of debt ratio and return on Sales). This table gives us deeper insight into the engineering sector:

- *Liquidity*. For the subsector A "normal" Liquidity values are those over 2.0. However, for companies in subsector B the values with maximum probability are in the interval 0.5–1.5 and for companies in subsector C are in the interval 0.5–1.0.
- *Inventory turnover* and *accounts receivable turnover*. We can consider these two indices as to be quite stable throughout the engineering sector. The nature of these indices does not change within particular engineering subsectors.
- *Accounts payable turnover*. The subsector analysis presented above discovered that the engineering sector is not homogeneous in terms of the accounts payable turnover index as it seemed from the previous sector analysis (cf. Table 4). Conditional probabilities of this index significantly vary within particular subsectors as can be seen from Table 6.

Table 5. Subsectors identification

Subsector	Debt ratio	Ret. on sales
A	0.00–0.25	0–0.05
B	0.25–0.50	0–0.05
C	0.50–0.75	0–0.05

The presented sector and subsector analysis offers the following conclusion: An expert evaluating the current financial position of a particular company should take into account the insufficient sensitivity of sector financial characteristics in comparison with the subsector ones. The probabilistic belief network applied in this paper to the financial analysis seems to be useful for the decision-maker since it enables quantitative modeling of various economic subsectors and thus, more accurate and confident evaluation.

6
Rating analysis

The presented Bayesian network models various relationships among the financial indices. Unfortunately, the network does not give any analytical tool for evaluating a given economic position of an enterprise. The network does not say if the financial position is good or bad. In order to overcome this problem we decided to introduce a new variable – Rating score – which will indicate the quality of the current financial position. Table 7 contains proposed values of this variable.

How can we incorporate the Rating score variable into the Bayesian network. To answer this question we should solve the following two problems:

- Computing the rating score
- Re-learning of the Bayesian network

Table 7. Rating score

Rating score (R)	1 – very good 2 – good 3 – average 4 – bad 5 – very bad

6.1 Computing the rating score

There is no general mathematical formula for computing the rating score. Instead, various banks and experts use a number of methods (mathematical and statistical tools, expert systems, etc.) for evaluating the score. Very often the experts estimate the financial position on the basis of their knowledge and experience without using any mathematical tool. In the light of this we decided to construct a neural network trained on a set of evaluated examples. In this section we describe the whole process.

The expert determined eight financial indices which he uses for the rating analysis. These are: *DR*, *L*, *RS*, *IT*, *DT*, *LT*, *ST*, *E*. The following task was to prepare a set of typical examples of engineering enterprises expressed in terms of the chosen variables. There are two possible alternatives: (a) sampling on the basis of the uniform probability distribution over the eight indices and (b) sampling using the empirical probability distribution represented by the database. In our application we used a combination of these two approaches. For this purpose we used the method of Kohonen maps (see [9] for details): Consider the eight-dimensional cube with edges of the length 1. Next, consider a regular grid with 96 vertices inside the cube. Then, take the first case from the database and normalize the values of considered indices. This case represents a point inside the cube. Find a vertex within the grid, closest to the database case. Move this vertex towards the case (the movement step is controlled by a parameter). After examining all the database cases the 96 vertices will be reordered according to the empirical probability distribution represented by the database. These 96 vertices represent the training set for the constructed neural network. For testing the neural network we prepared additional 32 examples using the same methodology. Afterwards we asked the expert to assign the Rating score value to these 128 examples. The expert did not use any mathematical technique. He classified the examples on the basis of his experience and intuition.

After evaluating the examples we constructed a three layered feed-forward neural network. The network has 8 neurons in the input layer, 5 neurons in the hidden layer and 5 neurons in the output layer. The number of the input neurons is given by the number of the input variables, the number of the output neurons is given by the number of classes to which the input vector should be classified. The neurons in the hidden layer use the nonlinear transforming function *tanh* and the output neurons use the function SOFTMAX, which gives the output in the interval $(0, 1)$. The input vector is then classified into the class represented by the output neuron with the maximal output value. Table 8 summarizes the quality of the constructed neural network.

The first row of the table states that 89 (93%) training and 28 (88%) testing examples were classified by the expert as well as the neural network into the same class. The second row reports those examples where the network's rating and the expert's rating differ by 1 (i.e. they were classified not into the same but neighbouring classes). As states the third row, there is only one example in the testing set which was evaluated as "4 – bad" by the expert and "2 – good" by the network.

6.2 Re-learning of the Bayesian network

Our goal was to incorporate the Rating score variable into the Bayesian network. In order to use the probabilistic learning algorithm we need, first, incorporate this variable into the database. Of course, we cannot ask the expert to evaluate all the database cases. Instead, we used the neural network to asses the database cases and extend the original database. After this process we achieved the new database with 11 variables and used this database for repeated learning of the Bayesian network. During this re-learning process we kept the structure of the previously learned Bayesian network. We added only arrows connecting the rating score node with the remaining ones. The following new arrows were introduced: $R \to LT$, $R \to T$, $E \to R$, $DR \to R$ and $R \to RS$.

Rating analyses using the constructed Bayesian network bear new dimensions to the rating process. Using the Bayesian network one can obtain the Rating score in the form of a probability distribution instead of one number. Moreover, using the inference tool, we can fix a rating score we want to achieve and recompute conditional probabilities of the remaining variables to discover which financial outcomes are required to conform the expected rating score.

7 Summary

We have proposed an application of using probabilistic Bayesian networks in fundamental financial analysis. The Bayesian network has been constructed by means of the probabilistic learning algorithm from a database of annual reports of Czech engineering enterprises. As the random variables of our interest, we have taken categorized values of ten widely used financial indices. Next, we have presented the neural network which was trained on the set of examples evaluated by the expert. The neural network enables us to estimate a rating score of a particular enterprise. Using the neural network we have introduced the

Table 8. Quality of NN

Rating score difference	Train	Train (%)	Test	Test (%)
0	89	93	28	88
1	7	7	3	9
2	0	0	1	3
3	0	0	0	0
4	0	0	0	0
Sum	96	100	32	100

new rating variable in the database. The extended Bayesian network was constructed by means of the the extended database.

References

1. **Jensen FV** (1996) Introduction to Bayesian Network, UCL Press, London
2. **Heckerman D, Geiger D, Chickering DM** (1994) Learning Bayesian Networks: The Combination of Knowledge and Statistical Data. Technical report MSR-TR-94-09. Microsoft Research, Advanced Technology Division, Microsoft Corporation, Redmond, WA 98052
3. **Gemela J** (1997) A Bayesian Approach to Learning Bayesian Networks, Technical report LISp-97-01. Laboratory for Intelligent Systems, Prague University of Economics, Prague
4. **Lam W, Bacchus F** (1994) Learning Bayesian Belief Networks: An Approach Based on the MDL Principle. Comput. Intelligence **10**(3)
5. **Danos P, Imhoff EA** (1992) Introduction to Financial Accounting, Irwin
6. **Hermanson RH, Edwards J, Maher MW** (1992) Accounting Principles, Irwin
7. **Gemela J** (1997) Financial Analysis using Bayesian Networks, Diploma thesis. Prague University of Economics, Prague
8. **Hajek P, Havranek T, Jirousek R** (1992) Uncertain Information Processing in Expert Systems, CRC Press, Inc. Boca Raton
9. **Sima J, Neruda R** (1996) Theoretical Questions of Neural Networks, matfyzpress