ELSEVIER

Knowledge-Based Systems 14 (2001) 189–195

Knowledge-Based SYSTEMS

www.elsevier.com/locate/knosys

A data mining approach to the prediction of corporate failure

Feng Yu Lin*, Sally McClean[1]

Faculty of Informatics, University of Ulster, Coleraine BT52 1SA, UK

Accepted 2 February 2001

Abstract

This paper uses a data mining approach to the prediction of corporate failure. Initially, we use four single classifiers — discriminant analysis, logistic regression, neural networks and C5.0 — each based on two feature selection methods for predicting corporate failure. Of the two feature selection methods — human judgement based on financial theory and ANOVA statistical method — we found the ANOVA method performs better than the human judgement method in all classifiers except discriminant analysis. Among the individual classifiers, decision trees and neural networks were found to provide better results. Finally, a hybrid method that combines the best features of several classification models is developed to increase the prediction performance. The empirical tests show that such a hybrid method produces higher prediction accuracy than individual classifiers.

Keywords: Corporate failure; Data mining; Hybrid method

1. Introduction

Due to recent changes in the world economy and as more firms, large or small, seem to fail now more than ever corporate failure prediction is of increasing importance. Corporate failure prediction is not only an interesting but also a challenging problem that has led to several studies over the past four decades. Numerous statistical classifiers have been constructed for the prediction of corporate failure. The main techniques used include discriminant analysis (DA) [1,28], logistic regression (LG) [19,23,32], probit analysis [9,27], mathematical programming [7], expert systems [15], artificial neural networks (NN) [2,30], decision tree method [14], rough sets [11] and multi-criteria decision aid [18]. In this paper, we will use the data mining approach, which is a systematic approach to find hidden patterns, trends and relationships in data and sometimes we refer to it as knowledge discovery.

Altman [1], Argenti [4] and Lincoln [22] argue that corporate failure is not an instantaneous occurrence but that it is a process which evolves over a considerable period of time. Companies do not fail overnight. Since corporate failure evolves over a considerable period of time, this gives us the foundation for predicting corporate failure. According to UK 1985 Act (Section 228): 'every balance sheet and profit and loss account shall give a true and fair view', so the financial statement data are a set of facts which should include extensive warning signals pointing toward failures. Using these facts, we can extract the discovery of patterns or rules describing a firm's financial status; such patterns or rules potentially lead to decision support using methods such as classification and forecasting.

2. Data mining

Data mining is a set of techniques carried out in a logical order. There are a number of basic tasks that are frequently carried out as part of data mining, many of which may be tackled using a number of different approaches. In this study, we follow a procedure for data mining: the SAS SEMMA methodology [16].

Fig. 1 shows the SEMMA process. This acronym SEMMA is based on the five stages of the data mining process — sampling, exploration, manipulation, modelling and assessment. Not all steps are critical to the whole data mining process, however, a data mining project should at least consider all of them.

2.1. Sampling

The data sample consists of company financial data from the UK. The companies are divided into two groups: one is the failed companies group and the other is the nonfailed

* Corresponding author. Tel.: +44-28-70324702; fax: +44-28-70324916.

E-mail addresses: fy.lin@ulst.ac.uk (F.Y. Lin), si.mcclean@ulst.ac.uk (S. McClean).

[1] Tel.: +44-28-70324602; fax: +44-28-70324916.

0950-7051/01/$ - see front matter
PII: S0950-7051(01)00096-X

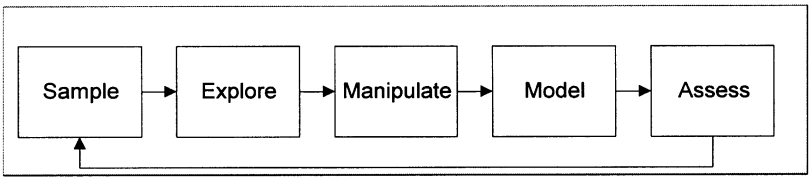

Fig. 1. The SEMMA data mining process.

companies group. The nonfailed companies are those listed on the London Stock Exchange, while the failed firms in this study are of those 1 year prior to failure. They are all UK listed companies of negligible value, or companies in receivership and/or liquidation.

The data collected relates to 1133 companies over the span of the last 20 years, 1980–1999. No attempt is made to match the failed and nonfailed companies. Ohlson [24] states 'the appropriate criteria to be used for matching purposes are not obvious'. Frydman et al. [14], Taffler [28], Watson [29] and Lacher et al. [20] have previously undertaken this kind of unmatched sample study.

The financial data were accessed from Datastream/ICV, which is part of Primark global information services (NYSE/PSE: PMK). The attributes used are account items from both the balance sheet and P/L (profit and loss) account (i.e. income statement). In this study, the initial collected variables covering the important areas of overall business performance are listed as in Table 1.

This training sample consists of 690 nonfailed companies and 106 failed companies. The data for the failed group relates to 1 year before failure. All the companies' data in the training sample are between 1980 and 1990. Estimated models based on this training sample are used to predict the unseen test dataset which consists of companies' data between 1991 and 1999. Details of the training sample and test sample are shown in Table 2.

Table 1
Initial collected variable sets

Concept	Variables	Meaning
Profitability (rates of return)	X701	Return on S'holders equity%
	X703	Return on S'holders capital%
	X705	Return on long-term capital%
	X707	Return on capital employed%
	X709	Return on net fixed assets%
	X711	Trading profit margin%
Profit margins	X713	Operating profit margin%
	X716	Pre-tax profit margin%
	X717	Net profit margin%
	X718	Earnings margin%
	X719	Cash flow margin%
Efficiency (turnover ratios)	X721	Turnover/assets employed
	X722	Turnover/fixed assets
	X723	Turnover/N. current assets
	X724	Stock turnover
	X725	Stock ratio (days)
	X726	Debtors turnover
	X727	Debtors ratio (days)
	X728	Creditors turnover
	X729	Creditors ratio (days)
Gearing ratios	X731	Capital gearing%
	X732	Income gearing%
	X733	Borrowing ratio
	X734	Income gearing% (inc ass)
	X735	GCF/total liabilities
	X736	Pref and loan/equ. and reserves
	X737	Loan cap./equity and reserves
Liquidity ratios	X741	Working capital ratio
	X742	Quick assets ratio
	X743	Cash/curr. liabilities
Productivity ratios	X761	Tax ratio
	X762	Sales per employee
	X763	Operating profit per employee
	X764	Capital employed/employee
	X765	Stock and W.I.P. per employee
	X766	Average salary per employee
Per-share items and yields	X792	Cash earnings per share

2.2. Data exploration

Exploration or preprocessing of data is very important and is sometimes the most time-consuming part of the data mining process. Firstly, the data collected is not often presented in a ready to use state, so we need to transform it into a table or database. Secondly, usually the sample data collected contains missing data, this is especially true of data from the financial statement. Missing data is normally expressed as a blank or N/A (not available). Some software will treat a blank or N/A as zero, and some will not accept blank or N/A in the numerical field. Hence, we need to transform the blank or N/A into a reserved number, e.g. −999.99, which is a surrogate for the missing value. If processed in SPSS, we can put a statement 'MISSING VALUE ALL (−999.99)' to specify that the value −999.99 represents a missing value. Thirdly, we must filter out the redundant records such as duplicated data. To filter out such records, we can make use of a unique index, for

Table 2
Training set and test set

Data sample	No. of nonfailed firms	No. of failed firms	Time span
Training set	690	106	1980–1990
Test set	289	48	1991–1999