Corporate bankruptcy prediction using data mining techniques

M. F. Santos[1], P. Cortez[1], J. Pereira[2] & H. Quintela[3]
[1]*Department of Information Systems, University of Minho, Portugal*
[2]*School of Management of the Polytechnic Institute of Cávado and Ave, Portugal*
[3]*School of Technology and Management of the Polytechnic Institute of Viana do Castelo, Portugal*

Abstract

The interest in the prediction of corporate bankruptcy is increasing due to the implications associated with this phenomenon (e.g. economic, and social) for investors, creditors, competitors, government, although this is a classical problem in the financial literature.

Two kinds of models are generally adopted for bankruptcy prediction: (i) accounting ratios based models and (ii) market based models. In the former, classical statistical techniques such as discriminant analysis or logistic regression models have been used, while in the latter the Moody's KMV model was adopted.

This paper follows the first approach (i), and it is based on the analysis of the evolution of several financial indicators during a three-year period. A framework was developed, encompassing a total of 16 models. These differ in the data mining algorithm (e.g. Artificial Neural Networks or Decision Trees), the data used (all three years or just the last one) and the input attributes adopted (e.g. all accounting ratios or just the most significant ones). The experiments were conducted using the new Business Intelligence Development Studio of the Microsoft SQL Server. Very good results were achieved, with performances between 86% and 99% for all 16 models.

Keywords: data mining, knowledge discovery from databases, decision support, corporate bankruptcy, artificial neural networks, decision trees.

WIT Transactions on Information and Communication Technologies, Vol 37, © 2006 WIT Press
www.witpress.com, ISSN 1743-3517 (on-line)
doi:10.2495/DATA060351

1 Introduction

The corporate bankruptcy prediction is a classical problem in the financial literature. Indeed, bankruptcy is one of the four generic terms that are generally found in literature for corporate distress and could be defined as the condition in which a business cannot meet its debt obligations and petitions a court for either reorganization of its debts or liquidation of its assets. The causes of business failure and bankruptcy can be pointed into: economic, financial, neglect, fraud, disaster and others. Economic factors include industry weakness and poor location, while financial factors include excessive debt and insufficient capital. For example, investors want to minimize credit risk and prevent non-profitable investments. Therefore, several authors have researched this subject in the past. Bankruptcy predictions' impact is high for financial markets. Beaver introduced the Naïve Bayes approach in 1966 using a single variable [1] and Altman in 1968 [2] proposed the use of Linear Discriminant Analysis (LDA). Since then several contributions have been made to improve the Altman's results, using different parametric, semi parametric and non-parametric models. The use of data mining techniques such as Artificial Neural Networks (ANN), decision trees, and Support Vector Machine (SVM) for bankruptcy prediction started in the late 1980s [3–7]. Aziz and Dar [8] carried a work of critical analysis of methodologies of corporate bankruptcies prediction models concluding that almost all models are capable of doing well their job, but the advantage of developing models based in Data Mining techniques is the future integration in Intelligent Systems.

This paper presents a framework that evaluates a total of 16 distinct models, by comparing different algorithms (e.g. Artificial Neural Networks and Decision Trees), training strategies (e.g. balanced training sets), and feature selection (e.g. the use of one or all three years data). The aim is to diagnose bankruptcy prediction based in real data collected in the North of Portugal.

The paper is organized as follows: first, the basic concepts are introduced and the data is presented and described; then the data mining techniques applied are presented; next, the experiments performed are described, being the results analysed in terms of several criteria; finally, closing conclusions are drawn.

2 Materials and methods

2.1 Data

In general, the datasets of bankrupt prediction studies use financial ratios (e.g. liquidity, solvency, leverage, profitability, asset composition, firm size, growth), cash flow information, and other variables of interest that include information on macroeconomic, industry specific, location or spatial and firm specific variables.

The data collected used financial indicators from a total of 2468 companies located at the North of Portugal during a consecutive 3-year period. The time window for this data collection was from 1999 until 2003. The database includes enterprises from the industry and retail fields. Each row of the original dataset

contains the following attributes: a key internal code; an identification of the business area of activity (e.g. textile); 58 ratios of activity, collected during three years in the form: *Ratio* x_n, where x is the name of the ratio and n the ordinal year (1, 2 or 3), and the situation of the company at the end of the 3 year period observation (0 for success and 1 for bankruptcy).

Figure 1: Distribution of companies in original dataset and after pre-processing phase.

In the pre-processing phase of the Knowledge Discovery in Databases (KDD) process, it was identified that 180 rows contained missing attributes. To solve this issue, the strategy followed was the deletion of the rows with missing attributes. This cleansing procedure was carried due to the difficulty to fill the missing attributes with their real values. Nevertheless, it should be noted that the remaining dataset still has a high number of cases (2 288 companies) and distribution of the dependent variable in the post-processed data set was kept (Figure 1).

2.2 Data mining

2.2.1 KDD pre-processing phase

All experiments reported in this section were conducted using the Microsoft SQL Server 2005. In the pre-processing phase of the KDD process, it was used the SQL Server Integrations Service (SSIS). Formerly known as Data Transformation Services (DTS), SSI is the Microsoft's extraction, transformation and loading tool [9]. The modelling phase was carried out using the SQL Server Business Intelligence Development Studio.

Since the dataset is biased, with the majority of the examples belonging to the no bankruptcy class, two modelling approaches were considered. The first approach uses the original distribution (86% of successful companies and 14% of bankrupt companies) in both test and training samples. For the second approach, a balancing procedure was adopted, where a random number of false cases was deleted from the training set, in order to obtain a 50-50% distribution. However, the original 86-14% ratio is kept in the test sample (Figure 2). In general, the balancing procedure has the advantage of reducing the number of false positives.

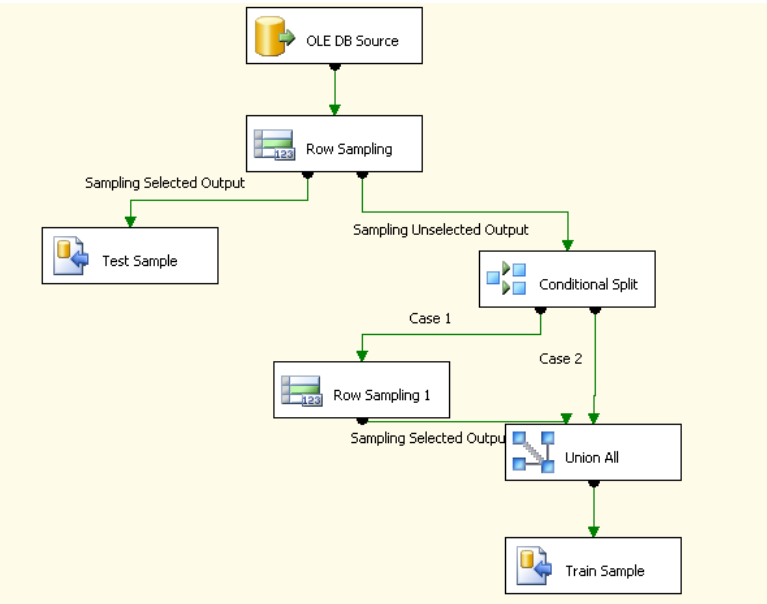

Figure 2: Data flow for generation of test and training samples.

Both approaches use 1/3 of the whole dataset for the test set (model evaluation). The first approach uses of the rest 2/3 of the data set for training (model estimation). The balanced training set contained less data, with a total of 416 rows (213 of bankrupt companies and 213 of successful ones).

2.2.2 KDD data mining phase

The experiments for developing the predictive model for bankruptcy prediction were done using the SQL Server Analysis Services, in the SQL Server Business Intelligence Development Studio, using the following DM techniques: Decision Trees and Neural Networks. These are denoted by the terms Microsoft Decision Trees and Microsoft Neural Networks in the Business Intelligence Suite.

Decision Tree is one of the most efficient and popular DM classification technique. It adopts a branching structure of nodes and leaves, where the knowledge is hierarchically organized. Each node tests the value of some feature, while each leave assigns a class label. The most popular decision tree algorithms for classification are ID3, C4.5 and C5.0 proposed by Ross Quinlan. The CART Classification algorithm proposed by Briemann is also widely adopted. The Decision Trees used in this work (Microsoft Decision Trees), are built has a hybrid of these algorithms (C4.5 and CART).

On the other hand, Neural Networks are other important DM techniques, denoting a set of connectionist models inspired in the behaviour of the human brain. The Multilayer Perceptron (MLP) is a popular architecture, where neurons are grouped in layers and only forward connections exist, capable of nonlinear mappings. In particular, the Microsoft Neural Network is a feedforward network,

which uses de *weighted sum* approach as combination function and *tanh* as the activation function in the hidden nodes. For output nodes it uses the *sigmoid* function. The number of hidden nodes in the hidden layer is automatically set by the algorithm.

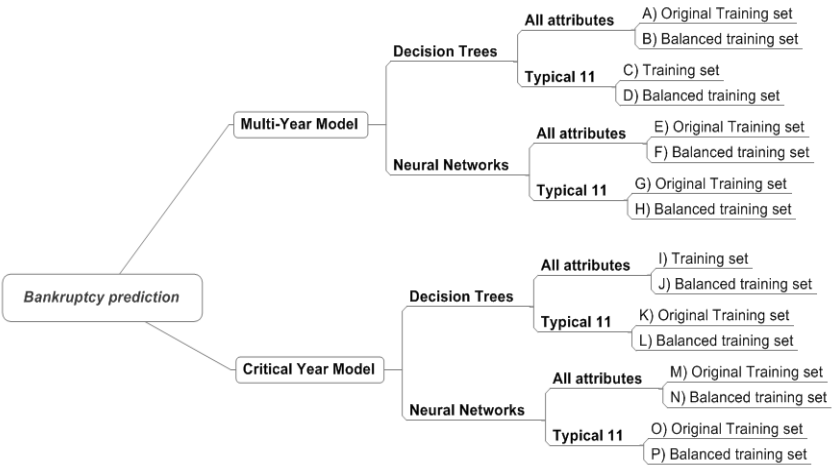

Figure 3: The 16 models of the bankruptcy prediction framework.

3 Results

3.1 Framework

For bankruptcy prediction, it was followed the approach presented in Figure 3. This framework includes a total of 16 distinct models, which differ in the Data Mining algorithm, training strategy and feature selection approach. Regarding the latter, two classes of models were considered. One uses data from just the last year, while the other uses all the data collected from three consecutive years. Furthermore, the input attributes used consisted in all the collected indicators (58 ratios), and in a second experiment, guided by a finance expert, just the eleven presented in Table 1. It should be noted that these are considered in literature the core ratios for predicting bankruptcy due to their inherent high level of discrimination power. On the other hand, the training strategy was divided into balanced and non-balanced training sets. Finally, two Data Mining algorithms were tested Decision Trees and Artificial Neural Networks.

The accuracy was estimated using the Holdout Method. In each simulation, the available data was randomly divided into two mutually exclusive partitions: the training set, with 2/3 of the available data and used during the modelling phase; and the test set, with the remaining 1/3 examples, being used after training, in order to compute the accuracy values.

For the classification analysis, it was used the confusion matrix which is a matrix of size L x L, where L denotes the number of possible classes. This matrix is created by matching the predicted (given by the Data Mining model) and actual (desired result).

Table 1: Core ratios.

Attribute	Description
Ratio 5_n	Current assets/Total assets
Ratio 6_n	Current assets/Current liabilities
Ratio 10_n	Equity/Total assets
Ratio 12_n	Equity/Liabilities
Ratio 13_n	Cash-flow/Current liabilities
Ratio 14_n	Cash-flow/Liabilities
Ratio 24_n	Working capital/Total assets
Ratio 43_n	Retained earnings/Total assets
Ratio 46_n	Net profit/Total assets
Attribute	**Description**
Ratio 48_n	Net profit/Liabilities
Ratio 55_n	Sales/Total assets

3.2 Experimental results

Table 2 and Figure 4 present the Bankruptcy, Nonbankruptcy and Global Accuracy rates for all the tested models. The analysis of the results shows that there are no significant differences for the feature selection, algorithms and approach followed for data partition (train set and test set) used in this experiments.

Overall, the results are quite satisfactory, with performances ranging from 86% to 99%. For bankruptcy prediction the best models (Table 3) are **A** and **P** that were generated using two completely different strategies. Model **A** is based in a Decision Tree using all the attributes and a biased training set (the original frequency), while **model P** (Figure 5) is based in an Artificial Neural Network using only the eleven core ratios (Table 1) and a balanced training set.

Regarding the Model **A**, the most important attributes are: *Ratio* 40_3 and *Ratio* 14_3.

Turning to the Neural Network model, the most important attributes are the *Ratio* 46_3, *Ratio* 48_3, *Ratio* 14_3, *Ratio* 55_3. Figure 5 is an example of how the Neural Network model can be interpreted. For example, in first line, if the *Ratio* 46_3 is in domain [0,020,…,0,295] the probability of corporate no bankruptcy is of 99.66%. Instead, if the ratio is in domain [-0,727,…,-0.254] the probability of corporate bankruptcy is of 99.66%.

The results also suggest that for this problem, when using the Microsoft Decision Trees algorithm, there are no improvements when using balanced training sets.

Table 2: The predictive test performance (in %) for all 16 models of the proposed framework.

Model	Bankruptcy	Nonbankruptcy	Global Accuracy
A	**98%**	**99%**	**99%**
B	96%	90%	92%
C	96%	99%	98%
D	95%	95%	95%
E	95%	99%	99%
F	97%	97%	97%
G	94%	99%	98%
H	97%	98%	98%
I	86%	99%	97%
J	95%	95%	95%
L	95%	95%	95%
M	96%	99%	99%
N	97%	98%	97%
O	97%	99%	99%
P	**98%**	**98%**	**98%**

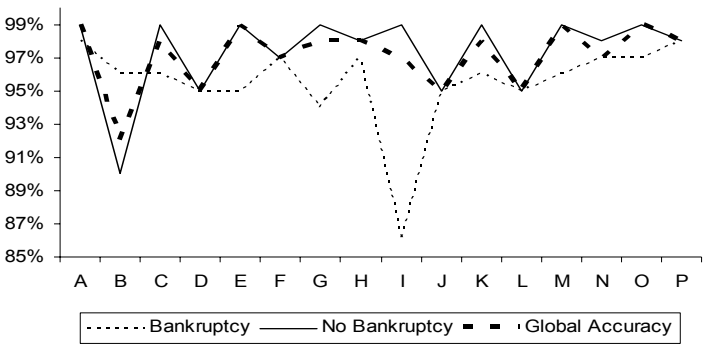

Figure 4: Classification accuracy.

4 Conclusions and further work

This paper presented a study of the Bankruptcy prediction based on Data Mining models (Artificial Neural Networks or Decision Trees). The models were induced making use of a framework containing a total of 16 models distributed

in two sets: the first one corresponding to a multi-year approach (tree consecutive years) and the second one based on one-year approach (the last year). The experiments were conducted using the new Business Intelligence Development Studio of the Microsoft SQL Server.

Table 3: The confusion matrix for models **A** and **P**.

Model A

Class	Test Set	
	0	**1**
0	112	9
1	2	632

Model P

Class	Test Set	
	0	**1**
0	109	15
1	2	629

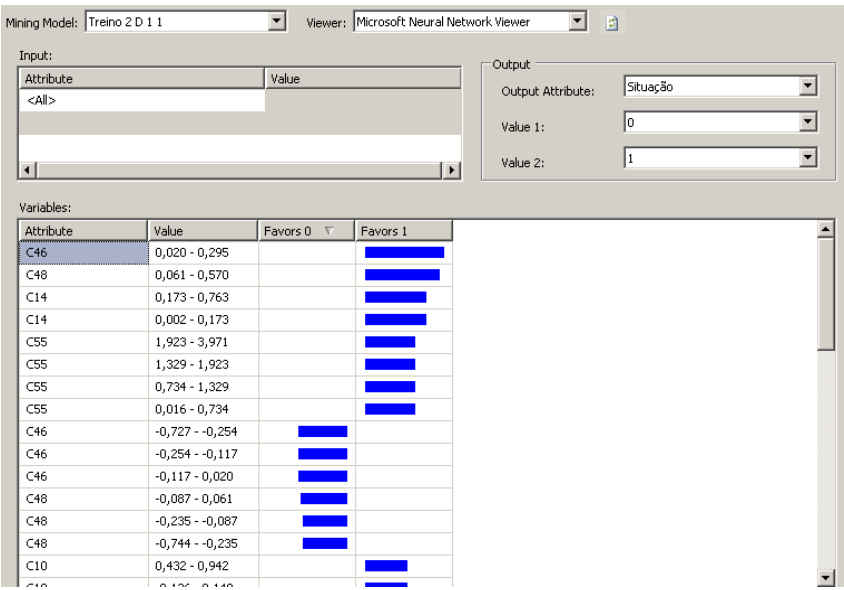

Figure 5: **Model P** viewer.

Accuracies between 86% and 99% were obtained, indicating that the followed approach enable the use of Data Mining models to predict the corporate bankruptcy. The most influent attributes considered in the generated models are the following: *Ratio* 46_3, *Ratio* 48_3, *Ratio* 14_3, *Ratio* 55_3 and *Ratio* 40_3.

It should be stressed that the results of this study are encouraging when compared with previous studies. For instance, the most relevant related works can be found in [10], where a Neural Network model obtained a maximum accuracy of 95%; and in [11] where a Decision Tree based model attained an

accuracy of 83%. Furthermore, this study considered a large dataset, with a total of 2 288 companies. In contrast, [10] used a reduced sample, with only 282 organizations, while [11] considered even a smaller number, with 72 companies.

Although quite good results were attained, there are still some limitations that should be considered in future research. For instance, macro-economic indicators (e.g. tight monetary police, the investor's expectations about economic conditions, the state of the economy) and qualitative variables (e.g. if there is a budgetary control system, if the skills on the board are unbalanced) could be used. It would be interesting also to consider a fine grained analysis; i.e. the results could be detailed by activity sector (e.g. textile or shoes).

References

[1] Beaver W.H., 1966, Financial ratios as predictors of failure, Journal of Accounting Research 4(3) 71-111.

[2] Altman, E.I., 1968, Financial ratios, discriminant analysis and the prediction of corporate bankruptcy, Journal of Finance 23(4), 589-609.

[3] Kotsiantis, S., Tzelepis, D., Koumanakos, E., Tampakas, V., 2005, 2nd International Conference on Enterprise Systems and Accounting (ICESAcc'05) 39-48, Greece.

[4] Pompe, P., Feelders, A., 1997, Using Machine Learning, Neural Networks, and Statistics to Predict Corporate Bankruptcy, Microcomputers in Civil Engineering 12, 267-276.

[5] Zang, G., Hu, M.Y., Patuwo, B., Indro, D.C., 1999, Artificial Neural Networks in Bankruptcy Prediction: General Framework and Cross-Validation Analysis, European Journal of Operational Research 116, 16-32.

[6] McKee, T., Greenstein, M., 2000, Predicting Bankruptcy Using Recursive Partitioning and a Realistic Proportioned Data Set, Journal of Forecasting, 19, 219-230.

[7] Shin, K., Lee, T., Kim, H., 2005, An Application of support vector machines in bankruptcy prediction model, Expert Systems with Applications, Volume 28, pp. 127-135.

[8] Aziz, M.A., Dar, H.A., Predicting Corporate Bankruptcy: Whither do We Stand.

[9] Rizzo, T., Machanic, A., Skinner, J., Davidson, L., Dewson R., Narkiewicz, J., Sack, J., Walters, R., Pro SQL Server 2005, Apress.

[10] Coats, P.K., Fant, L. F., 1993, Recognizing Financial Distress Patterns Using a Neural Network Tool, Financial Management, Vol. 22, n.º 3, Autumn, pp. 142-155.

[11] Sung, T. K., Chang, N., Lee, G., 1999, Dynamics of Modeling in Data Mining: Interpretive Approach to Bankruptcy, Journal of Management Information Systems, Vol. 16, nº 1, Summer, pp. 63- 85.