Accurate Prediction of Financial Distress with Machine Learning Algorithms

A. S. Vieira*, João Duarte*, B. Ribeiro♣ and J. C. Neves+
*ISEP, Rua de S. Tomé, 4200 Porto, Portugal, asv@isep.ipp.pt
♣Department of Informatics Engineering, University of Coimbra, P-3030-290 Coimbra, Portugal
+ISEG - School of Economics, Rua Miguel Lupi 20, 1249-078 Lisboa, Portugal

Abstract

Prediction of financial distress of companies is analyzed with several machine learning approaches. We used DIANE, a large database containing financial records of small and medium size French companies from the year of 2002 up to 2007. It is shown that inclusion of historical data, up to 3 years priori to the analysis, increase the prediction accuracy. In particular, fluctuations of some financial ratios are found to be crucial. Due to the inclusion of a large amount of inputs particular attention is given to feature selection. An accuracy of up to 94% is achieved with the best models.

1. Introduction

One of the most important threats for business is the credit risk associated with counterparts. The rate of bankruptcies have increased in recent years and its becoming harder to estimate as companies become more complex and develop sophisticated schemes to hide their real situation. Due to the recent financial crisis and regulatory concerns, credit risk assessment is a very active area both from academic and business community. The ability to discriminate between faithful customers from potential bad ones is thus crucial for commercial banks and retailers.

The problem of bankruptcy prediction can be addressed as follows: given a set of financial ratios describing the situation of a company over a given period, predict the probability that this company may become bankrupted in a near future, normally during the following year.

Prediction of financial distress of companies with financial ratios has been addressed by several models. Despite all its limitations, Linear Discriminant Analysis is still largely used as a standard tool for bankruptcy prediction [1,2]. In particular, versions of the Logistic model [3] are widely used by credit ranking agencies.

In previous works we show that some recent machine learning approaches, like Genetic Algorithms and Support Vector Machines, are able to achieve superior accuracy in early detection of bankruptcy [4, 5]. For a review on the application of machine learning algorithms to financial distress prediction of companies see [6-8].

For this study we use a large database of French companies. This new database is very detailed containing information on a wide set of financial ratios spanning over a period of several years. It contains up to three thousands distressed companies and about sixty thousand healthy ones.

Using this dataset we compare the efficiency of the Logistic model with other machine learning algorithms, namely: Support Vector Machines, Neural Networks and AddaboostM1.

In order to make predictions more accurate we tested the models with data from three previous years priori to failure. It is shown that inclusion of these historical records can boost precision and robustness of the classifiers, particularly in early detection.

This paper is organized as follows: Section 2 describes the dataset, Section 3 presents the models used, Section 4 contains the results obtained and in Section 5 the conclusions are presented.

2. The dataset

We used a sample obtained from Diane, a database containing financial statements of French companies. The initial sample consisted of financial ratios of about 60 000 industrial French companies, for the years of 2002 to 2006, with at least 10 employees. From these companies, about 3000 were declared bankrupted in 2007 or presented a restructuring plan (“Plan de redressement”) to the court for approval by the creditors. We decided not to distinguish these two categories as both signals companies in financial distress.

The dataset includes information about 30 financial ratios defined by COFACE of the companies covering a wide range of industrial sectors.

2.1. Preprocessing

Our database contains many cases with missing values, especially for defaults companies. For this reason we sorted the default cases by the number of missing values and selected the examples with 10 missing values at most. A final set of 600 default examples was obtained. In order to obtain a balanced dataset we selected randomly 600 non-default examples resulting in a set of 1200 examples.

The remaining missing data was treated as follows. For the ratios of the years 2003 and 2006 each missing value was replaced by the value of the closest available year; for 2004 and 2005, if values of the next and previous years were available, each missing value was replaced by their mean, otherwise it was replaced by the remaining value. In some cases there was no data available for a ratio in any of the years. In this very few cases the missing data was replaced by the median value of the ratio in each year. Finally, all ratios were logarithmized and then standardized to zero mean and unity variance.

2.2. Feature selection

The 30 financial ratios produced by COFACE are described in Table 1. These ratios allow a very comprehensive financial analysis of the firms including the financial strength, liquidity, solvability, productivity of labor and capital, margins, net profitability and return on investment. Although, in the context of linear models, some of these variables have small discriminatory capabilities for default prediction, the non-linear approaches here used may extract relevant information contained in these ratios to improve the classification accuracy without compromising generalization.

Due to the large number of attributes available, we used several ranking algorithms to select the most relevant. We used the following methods: SVM Attribute evaluation, Chisquared, Consistency Subset, GainRatio.

In SVMA [13] attributes are ranked by the square of the weight assigned by the SVM. Attribute selection is handled by ranking attributes for each class separately using a one-vs-all method and then "dealing" from the top of each pile to give a final ranking.

In Chisquared algorithm the worth of an attribute in calculated by computing the chi-squared statistic with respect to the class.

Consistency Subset [9] evaluates the worth of a subset of attributes by the level of consistency in the class values when the training instances are projected onto the subset of attributes.

GainRatio evaluates the worth of an attribute *A* by measuring the information gain ratio with respect to the class *C*:

$$Gain(C,A) = \frac{H(C) - H(C,A)}{H(A)},$$

where *H*(*X*) is a measure of the entropy.

2.3. Historical data

A company is a dynamic entity, subjected to fluctuation of the market, economy cycles and unavoidable contingencies related to its business activity. Therefore, yearly variations of important financial ratios reflecting the balance sheet, sometimes quite relevant, are common particularly for small companies. Yearly variations of over 50% in some ratios are not atypical.

In order to accommodate these fluctuations, we decided to use an extended record from years preceding the default. However care must be taken in choosing the relevant information. If many years are used, we increase the complexity of the problem and may obscure the present situation of the company by averaging over a remote past. On the other hand if few years are used we may not properly characterize the company background. In this study we considered data from 3 years priori to the bankruptcy event.

This adds complexity to the analysis as the number of inputs is increased threefold - from 30 to 90 ratios. Furthermore, we found that more relevant than the ratios themselves, are the variations that occurs over the period range of the analysis.

We consider the following parameters. Ratios of the current year, R_0^i, ratios from previous year, R_1^i, and fluctuations of the ratios over the period considered, R_2^i. They are defined as:

$$R_0^i = R^i$$
$$R_1^i = R^{i-1}$$
$$R_2^i = \frac{\max(R^i) - \min(R^i)}{\bar{R}}$$

where $\bar{R}$ is the ratio average over the period considered.

The variables selected by the feature ranking algorithms are presented in Table 1. Note that many of selected attributes are the ratios variations over the period, in this case three years, R_2^i. Features selected by different algorithms differ considerably meaning that important correlations can exist in the ratios.

3. Models used

We analyze the data with four machine learning algorithms: Logistic, Neural Networks with a Multilayer Perceptron (MLP), Support Vector Machine (SVM) [10,11] and AdaBoost M1.

Table 1. Ratios used with historical data set. Selection procedure: [SVMA - Top 20 (I), Top 15 (II), Top 10 (III), Top 5 (IV)], [Cfs Subset Evaluation - Greedy (V), Genetic (VI)], [Chisquared: Top 20 (VII)], [Consistency Subset: Greedy (VIII), Genetic (IX)] and [GainRatio: Top 20 (X)]. The labels means: 0 - current year, 1 - previous year, 2 - variation over three previous years.

#	**Designation**	**I**	**II**	**III**	**IV**	**V**	**VI**	**VII**	**VIII**	**IX**	**X**
1	Number of employees	0	0	0	0	0	0		0	0	
2	Financial Debt / Capital Employed %										
3	Capital Employed / Fixed Assets	2	2	2							
4	Depreciation of Tangible Assets (%)	0	0		0	0					
5	Working capital / current assets						0		2	1,2	
6	Current ratio	0					2		0		
7	Liquidity ratio	1				0,2	0	0	0,2	0,2	
8	Stock Turnover days						2			0	
9	Collection period	0	0	0							
10	Credit Period						2				
11	Turnover per Employee	0,2	0,2	0,2	0,2	0,2	0,2	2	2		2
12	Interest / Turnover	0	0							0	
13	Debt Period (days)	2	2	2						2	
14	Financial Debt / Equity (%)					0	0	0			0
15	Financial Debt / Cashflow	0	0	0		0,2	0,2	0,2			0,2
16	Cashflow / Turnover (%)	0,1	0,1	0	0	0,2	0,2	0,2	2		0,2
17	Working Capital / Turnover (days)	0,1	0,1	1			2			0	
18	Net Current Assets/Turnover (days)	0	0								
19	Working Capital Needs / Turnover (%)	2									
20	Export (%)						2			1	
21	Value added per employee					2	0,2		1	0	
22	Total Assets / Turnover					2	2		2		
23	Operating Profit Margin (%)						0,2	0,2		0,2	0,2
24	Net Profit Margin (%)						0,2	0,2			0,2
25	Added Value Margin (%)								0		
26	Part of Employees (%)					0	0,1	0		1	0
27	Return on Capital Employed (%)	2				0	0,2	0,2		1,2	0,1,2
28	Return on Total Assets (%)	0	0	0	0	0,2	0	0,2	0		0,2
29	EBIT Margin (%)							0,2		2	0,2
30	EBITDA Margin (%)	0					0	0,2			0,2

For MLP, we used a neural network with 1 hidden layer with a number of neurons defined by: $(number\_ratios + 1) / 2$. The learning rate was set to 0.3 and the momentum to 0.2.

For the C-SVM algorithm we used the LibSVM [12] library with a radial basis function as kernel with the cost parameter $C = 1$ and used the shrinking heuristic.

For AdaBoost M1 algorithm we used a Decision Stump as weaklearner and set the number of iterations to 100. No resampling was used.

4. Results

4.1 Prediction using data from a single year

First we compare the efficiency of the classifiers using data from a single year. Table 2 presents the results obtained when all ratios are used. We used 10-fold cross validation in all classifiers.

Support Vector Machines achieved the highest accuracy, 92.42% and the lowest error types. For 2005, two years before bankruptcy, the Adaboost retrieved the best results. It is remarkable such a high accuracy taking into account the fluctuations on ratios occurring from one year to the other. The highest error is type II, as expected. Neural Networks (MLP) were the worst classifier due to the large dimensionality of the training data exposing it to the corresponding risk of overfiting.

Table 2. Accuracy and error types for different models in 2007 using data from the previous year (2006) and from 2005. All 30 variables were used.

	Classifier	**Accuracy**	**Type I**	**Type II**
2006	Logistic	91.25	6.33	11.17
	MLP	91.17	6.33	11.33
	C-SVM	**92.42**	**5.16**	**10.00**
	AdaboostM1	89.75	8.16	12.33
2005	**Classifier**	**Accuracy**	**Type I**	**Type II**
	Logistic	79.92	**19.50**	20.67
	MLP	75.83	24.50	23.83
	C-SVM	**80.00**	21.17	**18.83**
	AdaboostM1	78.17	20.50	23.17

4.2. Historical data

Most default prediction models use a small set of financial ratios, between 5 and 10, usually from a single year, to quantifying the profitability, cashflow and liabilities of the company. Since we have a large pool of data, we tested the models with several sets of attributes. The first, containing the top 5 attributes, selected by SVMA, the second the top 10, the third the top 15, the fourth the top 20 and finally the fifth with all the 90 attributes.

The results are presented in Table 3. The best accuracy (94%) was obtained again with SVM using 20 variables, which means a reduction of about 30% in type I error and 20% in type II error. This improvement is justified by the fact that more data is used. With the 5 top ratios we achieved a performance similar to the previous dataset with all ratios included.

The Adaboost algorithm is the least sensitive to overfiting and therefore is relatively immune to the curse of dimensionality when using the full set of attributes. In practice it is unwise to use the full set of attributes and the accuracy has to be sacrificed to simplicity. For the top 20 ratios, selected by SVMA, the Logistic model achieved again an accuracy very close to SVM.

The importance of using a large set of ratios is clearly evident on prediction two years before bankruptcy (Table 4). In this case, the accuracy of SVM increased substantially from 77.42% to 81.42%.

Table 3. Accuracy in predicting failures during 2007 using data from 3 previous years (2006, 2005 and 2004).

	# inputs to the model				
Classifier	**5**	**10**	**15**	**20**	**ALL**
Logistic	91.17	**93.33**	**93.42**	93.58	92.25
MLP	91.33	93.25	93.08	93.17	92.50
C-SVM	**91.67**	**93.33**	**93.42**	**94.00**	**93.17**
AdaboostM1	90.50	**93.33**	93.00	92.25	91.58

Table 4. Accuracy for different models using data from 3 previous years (2005, 2004 and 2003), two years before bankruptcy.

	# variables				
Classifier	**5**	**10**	**15**	**20**	**ALL**
Logistic	75.83	77.58	79.25	79.33	79.00
MLP	75.83	75.00	76.17	75.33	77.42
C-SVM	76.42	76.58	78.83	78.41	**81.42**
AdaboostM1	**77.08**	**77.83**	**79.92**	**80.33**	81.33

In Fig. 1 the ROC curves for the four models are presented. AdaboostM1 has the greatest AUC (Area Under ROC curve), closely followed by the Logistic.

5. Conclusions

In this work it is shown that bankruptcy of small and medium size companies can be accurately predicted if a detailed training dataset is available. Of all the models tested Support Vector Machines achieved the best performance, but all approaches show comparable results.

We show that inclusion of information from previous years before default, especially fluctuations of relevant ratios, like debt to cash-flow, is crucial to achieve a good precision. Furthermore, we proved that the use of larger sets of inputs in the classifier can reduced both error types by up to 30%.

In future work we will consider inclusion of more years and the use of more efficient feature selection algorithms.

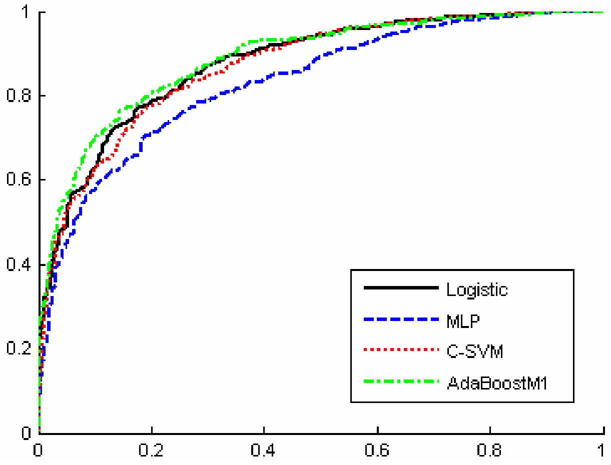

Figure 1. ROC curve using data from 3 previous years (2005, 2004 and 2003) and top 20 variables.

Acknowledgements

The authors like to thank “Fundação da Ciência e Tecnologia” for financial support under the project PTDC/GES/70168/2006.

References

[1] Altman, E. I. Financial Ratios, ”Discriminant Analysis and the Prediction of Corporate Bankruptcy”, *Journal of Finance*, 23 (1968) 589-609.

[2] R. A. Eisenbeis, “Pitfalls in the Application of Discriminant Analysis in Business, Finance and Economics”, Journal of Finance, 32 (3), June, (1977) 875-900.

[3] D. Martin, “Early Warning of Bank Failure: A Logit Regression Approach”, Journal of Banking and Finance, 1 (1977) 249-276.

[4] A. S. Vieira and J. C. Neves, “Improving Bankruptcy Prediction with Hidden Layer Learning Vector Quantization”, Eur. Account. Rev. **15** (2), (2006) 253-271.

[5] A. S. Vieira, B. Ribeiro, S. Mukkamala, J. C. Neves, A. H. Sung, “On the Performance of Learning Machines for Bankruptcy Detection”, IEEE International Conference on Computational Cybernetics, Vienna, Austria, August 30 – September 1, (2004) 223 – 227.

[6] F. Atiya, “Bankruptcy prediction for credit risk using neural networks: A survey and new results”, IEEE Trans. Neural Net., 4 (2001) 12-16.

[7] G. Udo Neural Network Performance on the Bankruptcy Classification Problem, Computers and Industrial Engineering, 25 (1993) 377-380.

[8] J. S. Grice and M. T. Dugan, The limitations of bankruptcy prediction models: Some cautions for the researcher, Rev. of Quant. Finance and Account., 17 no. 2 (2001) 151.

[9] H. Liu, R. Setiono, "A probabilistic approach to feature selection - A filter solution", 13th International Conference on Machine Learning (1996) 319-327.

[10] V. Vapnik, The Nature of Statistical Learning Theory. New York: Springer Verlag, 1995.

[11] C. Cortes and V. Vapnik, Support vector networks, Machine Learning 20 (1995) 273-297.

[12] C.-C. Chang and C.-J. Lin, LIBSVM: a library for support vector machines, Technical Report, Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan, 2000.

[13] I. Guyon, J. Weston, S. Barnhill, V. Vapnik (2002). Gene selection for cancer classification using support vector machines. Machine Learning. 46:389-422.