Forecasting Corporate Bankruptcy with an Ensemble of Classifiers

Despina Deligianni[1] and Sotiris Kotsiantis[2]

[1] Hellenic Open University, Greece
devi287@hotmail.com
[2] Department of Mathematics, University of Patras, Greece
sotos@math.upatras.gr

Abstract. Prediction of corporate bankruptcy is a phenomenon of growing interest to investors, creditors, borrowing firms, and governments alike. Timely identification of firms' impending failure is really wanted. The aim of this research is to use supervised machine learning techniques in such an environment. A number of experiments have been conducted using representative machine learning algorithms, which were trained using a data set of 150 failed and solvent Greek firms. It was found that an ensemble of classifiers could enable users to predict bankruptcies with satisfying precision long before the final bankruptcy.

1 Introduction

The problem of Bankruptcy prediction is a classical one in the financial literature (see e.g. [3] for a review). The main impact of Bankruptcy prediction is in bank lending. Banks need to forecast the possibility of default of a potential counterparty before they expand a loan. This can lead to sounder lending decisions, and consequently result in important savings.

There are two main approaches to loan default/bankruptcy prediction. The first approach, the structural approach, is based on modeling the underlying dynamics of interest rates and firm attributes and deriving the default probability based on these dynamics. The second approach is the statistical approach. Instead of modeling the relationship of default with the attributes of a firm, this relationship is discovered from the data. The focus of this article is on the empirical approach, particularly the use of supervised machine learning in bankruptcy prediction. The automated system uses financial ratios as predictors of performance, and assesses posterior probabilities of financial health (on the other hand, financial distress).

Balcaen and Ooghe [4] provide an overview of the standard statistical methodologies applied on business failure. Kumar and Ravi [15] present a survey of bankruptcy prediction via statistical and intelligent techniques. These techniques of corporate bankruptcy prediction have their own strengths and weaknesses and, hence, choosing a particular model may not be easy. Searching for best distress prediction models is still in progress [26]. This study provides a critical analysis of most frequently used

I. Maglogiannis, V. Plagianakos, and I. Vlahavas (Eds.): SETN 2012, LNAI 7297, pp. 65–72, 2012.

corporate bankruptcy learning models. Accordingly, we use a representative algorithm for each one of the most widespread machine learning techniques so as to investigate the efficiency of ML techniques in such an environment. Finally, it was found that an ensemble of classifiers could enable users to predict bankruptcies with satisfying precision long before the final bankruptcy.

The following section provides a short literature review in the domain of corporate bankruptcy learning models. Section 3 describes the data set of our study and the variable selection process. Section 4 presents the experimental results for the compared algorithms. Finally, section 5 discusses the conclusions and some future research directions.

2 Literature Review

Many studies have been conducted for bankruptcy prediction using models such as neural networks [16], [7], instance based learners [1], Bayesian models [21], rule learners [23], decision trees algorithms [8] and Support Vector Machines [22], [29]. Olson et al [18] applies a variety of data mining tools to bankruptcy data, with the purpose of comparing accuracy and number of rules. For that data, decision trees were found to be relatively more accurate compared to neural networks and support vector machines, but there were more rule nodes than desired.

Verikas et al [27] present a comprehensive review of hybrid and ensemble-based soft computing techniques applied to bankruptcy prediction. Tsai and Hsu [24] presented a meta-learning framework, which is composed of two-level classifiers for bankruptcy prediction. The first-level multiple classifiers perform the data reduction task by filtering out unrepresentative training data. Then, the outputs of the first-level classifiers are used to create the second-level single (meta) classifier. Hung and Chen [14] propose a selective ensemble of three classifiers, i.e. the decision tree, the back propagation neural network and the support vector machine, based on the expected probabilities of both bankruptcy and non-bankruptcy.

In the Greek context, logit analysis, probit analysis, and the linear probability model are the most commonly used techniques applied [17]. The performance of alternative non-parametric approaches has been explored in the Greek context to overcome the aforementioned shortcomings of the statistical and econometric techniques such as rough sets [9] and multicriteria discrimination method [12]. Tsakonas et al [25] used neural logic networks for bankruptcy prediction.

3 Data Description

Bankruptcy filings in the years 2003 and 2004 were supplied directly from the National Bank of Greece directories and the business database of the financial information services company called ICAP, in Greece. Financial statement data for the fiscal years prior to bankruptcy were supplied from ICAP financial directories. The financial statements of these firms were gathered for a period of three years. The critical year of failure denoted as year 0, three years before as year –3 and year –1 is the final

year prior to bankruptcy filing. As the control sample, each selected bankrupt firm was matched with two non-bankrupt (healthy) firms of exactly the same industry, by care-fully comparing the year of the reported data (year –1) assets size and the number of employees. The selected non-bankrupt corporations were within 20% of the selection criteria. Following the prior literature, we examine the probability of a firm's initial filing for bankruptcy and eliminate any observations for a firm after it has filed for bankruptcy during our sample period. Our final bankruptcy sample consists of 50 initial bankruptcies in the year period 2003-2004 and is similar in size but more complete compared to previous studies. The final pooled set of failed and healthy firms is composed of 150 individual firms with financial data for a three-year period, which attributes 450 firm-year observations. Through extensive literature review on bankruptcy prediction about 50 financial ratios were traced. The final set of the calculated input features is 21 because of missing financial data and financial ratio duplication. Table 1 provides a brief description of the financial variables. in order to show how much each attribute influences the induction, we rank the influence of each one according to different statistical measures e.g. Information Gain, Gain Ration and Relief Score [28]. The attributes that mainly influence the induction are: *WC/TA*, EQ/CE and GRNI (see ReliefF Score in Table 1). It seems that the attributes: CA/CL, NIMAR, ROCE, GRNS, ROE, QA/CL, S/TA and OPIMAR do not influence the induction in any way.

Table 1. Research Variables description and Average ReliefF score of each variable

Independent variable	Variable Description	Average Score
WC/TA	Working capital divided by total assets	0.035
EQ/CE	Shareholder's equity to capital employed	0.011
GRNI	Growth rate of net income	0.012
SIZE	Size of firm is the ln(Total Assets/GDP price index)	0.006
GRTA	Growth rate of total assets (TAt – TA t-1)/(ABS(TAt)+ ABS(TA t-1)	0.004
TD/EQ	Total debt to shareholder's equity capital	0.003
S/CE	Sales divided by capital employed	0.003
COLPER	Average collection period for receivables	0.002
S/EQ	Sales divided by Shareholder's equity capital	0.002
CE/NFA	Capital employed to net fixed assets	0.002
PAYPER	Average payment period to creditors	0.001
INVTURN	Average turnover period for inventories	0.001
GIMAR	Gross income divided by sales	0.001
CA/CL	Current assets to current liabilities	0
NIMAR	Net income divided by sales	0
ROCE	Net income pre tax divided by capital employed	0
GRNS	Growth rate of net sales	0
ROE	Net income pre tax divided by Shareholder's equity capital	0
QA/CL	Quick assets to current liabilities	0
S/TA	Sales divided by Total Assets	0
OPIMAR	Operating income divided by net sales	0

4 Experimental Results and Proposed Technique

Supervised machine learning is the investigation for algorithms that reason from externally supplied examples to produce general hypotheses, which will make predictions about future examples. For the purpose of this study, a representative algorithm for each learning technique was used. The most commonly used C4.5 algorithm [20] was the representative of the decision trees in our study. RBF algorithm [28] - was the representative of the ANNs. The RIPPER algorithm [6] was the representative of the rule-learners in our study. The Naïve Bayes algorithm [11] was the representative of the Bayesian networks in our study. The 1-NN algorithm was also used as a representative of lazy learners [28]. Finally, the Sequential Minimal Optimization (or SMO) algorithm was the representative of the SVMs [19].

The algorithms discussed here aim at achieving high classification accuracy that is lower error rate in the prediction of unseen instances. However, these algorithms do not differentiate the types of errors. That is for these algorithms classifying a bankrupt case as a non-bankrupt has the same error as classifying a non-bankrupt as a bankrupt. However, in real life, these costs is not the same for the decision maker. For example, the cost of predicting a case as non-bankrupt that is actually bankrupt is higher than vice versa. For our experiments, we used in the cost matrix 2 times more cost in the case that a non-bankrupt instance is actually bankrupt. We made this choice because in our data the set of non-bankrupt firms is two times the set of bankrupt firms, too. Cost-sensitive meta-learning converts existing cost insensitive base learning algorithms into cost-sensitive ones without modifying them. Therefore, it can be regarded as a middleware component that pre-processes the training set. All accuracy approximations were obtained by averaging the results from stratified 10-fold cross-validation in our data. It must be mentioned that we used the free available source code for our experiments by the book [28]. The results are presented in Table 2.

To facilitate the presentation and discussion of the results, each year prior to financial distress is denoted as year –1, year –2, year –3, Year –1 refers to the first year prior to financial distress (e.g., for the firms that faced financial distress in 2004, year –1 refers to 2003); year –2 refers to the second year prior to financial distress (e.g., for the firms that faced financial distress in 2004, year –2 refers to 2002), etc.

Table 2. Accuracy of the algorithms in each testing step

		Naive Bayes	***1-NN***	***RIPPER***	***C4.5***	***SMO***	***RBF***
Year(-3)	Bankrupt	26.5	49.0	63.3	69.4	30.6	40.8
	Non Bankrupt	85.4	71.9	49.0	35.4	81.3	67.7
Year(-2)	Bankrupt	26.5	40.8	59.2	55.1	38.8	38.8
	Non Bankrupt	92.7	82.3	51.0	57.3	78.1	81.3
Year(-1)	Bankrupt	26.5	40.8	65.3	59.2	65.3	46.9
	Non Bankrupt	94.8	82.3	75.0	75.0	70.8	77.1

In a comparative assessment of the models' performance we can conclude that RIPPER predicts more right the true positive bankrupt cases and NB the true positive non-bankrupt cases. For this reason, we implemented an algorithm that is based in RIPPER and NB decisions.

The concept of combining classifiers is proposed as a direction for the improvement of the performance of individual learners [5]. The goal of combining classification algorithms is to generate more certain, precise and accurate system results.

The most direct method for dealing with skewed class distributions with unequal misclassification costs is to use cost sensitive learning [13]. For our implementation, we used in the cost matrix 2 times more cost in the case that a non-bankrupt instance is actually bankrupt. As we have already mentioned, we made this choice because in our dataset the sample of non-bankrupt firms is two times the sample of bankrupt firms. The proposed ensemble algorithm is illustrated in Fig. 1.

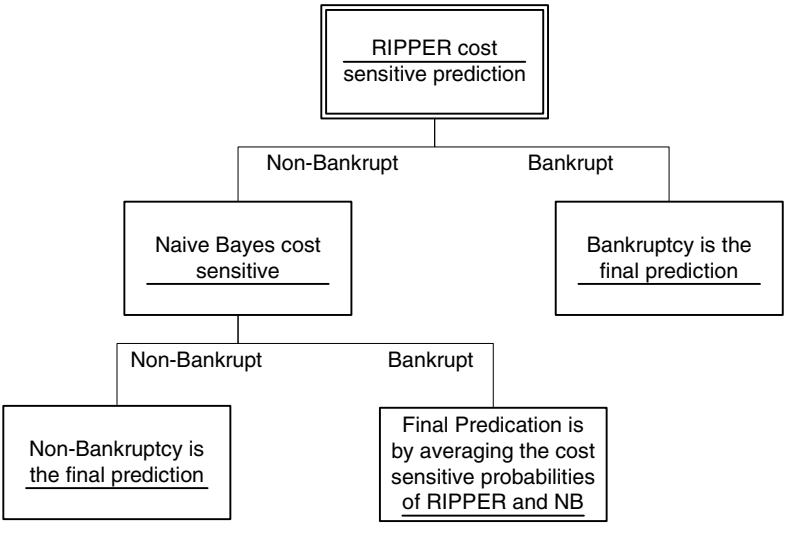

Fig. 1. The presented method

Since RIPPER predicts more right the true positive bankrupt cases, in the presented ensemble we choose RIPPER algorithm to start the classification process. The cost of predicting a case as non-bankrupt that is actually bankrupt is higher than vice versa. NB predicts more right the true positive non-bankrupt cases and for this reason, in the presented ensemble we choose to trust the decision of NB classifier for non-bankrupt cases in the second step. In the remaining cases, the presented model gives the final prediction by averaging the cost sensitive probabilities of the two chosen classifiers.

The results of the presented ensemble are compared with the well known cost-sensitive ensemble MetaCost [10] and the well known voting technique. MetaCost's procedure begins to learn an internal cost-sensitive model by applying a cost-sensitive procedure, which utilizes a base learning algorithm. Then, MetaCost procedure

approximates class probabilities using bagging and then re-labels the training instances with their minimum expected cost classes, and as a final point relearns a model using the modified training set.

Table 3. Accuracy of ensembles in our dataset

		Metacost Naive Bayes	*Metacost RIPPER*	*Voting RIPPER & Naive Bayes*	*Presented Method*
Year(-3)	Bankrupt	26.5	46.9	36.7	63.3
	Non Bankrupt	85.4	70.8	79.2	69.0
Year(-2)	Bankrupt	28.6	44.9	30.6	64.2
	Non Bankrupt	89.6	70.8	86.5	71.0
Year(-1)	Bankrupt	28.6	49.9	42.9	71.3
	Non Bankrupt	89.6	85.4	84.4	78.1

According to Table 3, our approach performs better than other examined ensemble methods as far as the average value of the true positive precision in both classes in the examined dataset. Both the training and classification time cost of the presented model is comparable with that of simple voting and less than the cost of Metacost.

5 Conclusion

With the help of supervised machine learning techniques, the experts are in the position to know which of the firms will bankrupt or not with satisfactory accuracy. For this reason, a prototype version of a software support tool has been constructed implementing the presented ensemble of classifiers (see Figure 2). Tracking progress is a time-consuming job that could be handled automatically by such a tool. While the experts will still have the crucial role in monitoring and evaluating progress, the tool could use the data required for reasonable and efficient monitoring. The prediction model developed from the present study proposes the importance of liquidity defined by the ratio working capital to total assets, capital structure defined as equity to capital employed and profitability growth defined as net income growth.

Nevertheless, there were a number of limitations in this study that must be noted. First, the sample size was fairly small. Thus, the generalization of the research results is somewhat limited. The second limitation was that only financial ratio attributes were included in this study. There may be other essential quantitative attributes (i.e. market value, stock data, age) as well as qualitative variables (leadership, type of ownership, reputation, etc.) and there is enough literature in organization theory that reports the value of these attributes. These limitations open up a open opportunity for future research.

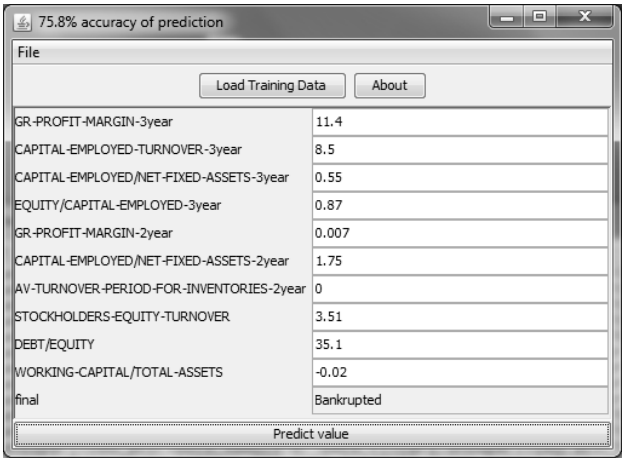

Fig. 2. A screenshot of the implemented decision support tool

References

1. Ahn, H., Kim, K.-J.: Bankruptcy prediction modeling with hybrid case-based reasoning and genetic algorithms approach. Applied Soft Computing Journal 9(2), 599–607 (2009)
2. Alfaro, E., García, N., Gámez, M., Elizondo, D.: Bankruptcy forecasting: an empirical comparison of AdaBoost and neural networks. Decision Support Systems 45(1), 110–122 (2008)
3. Altman, E.L.: Corporate Financial Distress and Bankruptcy. John Wiley and Sons (1993)
4. Balcaen, S., Ooghe, H.: 35 years of studies on business failure: An overview of the classic statistical methodologies and their related problems. The British Accounting Review 38, 63–93 (2006)
5. Berg, D.: Bankruptcy prediction by generalized additive models. Applied Stochastic Models in Business and Industry 23(2), 129–143 (2007)
6. Cohen, W.: Fast Effective Rule Induction. In: Proceeding of International Conference on Machine Learning, pp. 115–123 (1995)
7. Cho, S., Kim, J., Bae, J.K.: An integrative model with subject weight based on neural network learning for bankruptcy prediction. Expert Systems with Applications 36(1), 403–410 (2009)
8. Cho, S., Hong, H., Ha, B.-C.: A hybrid approach based on the combination of variable selection using decision trees and case-based reasoning using the Mahalanobis distance: for bankruptcy prediction. Expert Systems with Applications 37(4), 3482–3488 (2010)
9. Dimitras, A.I., Slowinski, R., Susmaga, R., Zopounidis, C.: Business failure prediction using rough sets. European Journal of Operational Research 114, 263–280 (1999)
10. Domingos, P.: MetaCost: A General Method for Making Classifiers Cost-Sensitive. In: Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining, pp. 155–164. ACM Press (1999)
11. Domingos, P., Pazzani, M.: On the optimality of the simple Bayesian classifier under zero-one loss. Machine Learning 29, 103–130 (1997)
12. Doumpos, M., Zopounidis, C.: A Multicriteria Discrimination Method for the Prediction of Financial Distress: The Case of Greece. Multinational Finance Journal 3(2), 71–101 (1999)

13. Elkan, C.: The foundations of cost-sensitive learning. In: Proceedings of the 17th International Joint Conference on Artificial Intelligence, pp. 973–978 (2001)
14. Hung, C., Chen, J.-H.: A selective ensemble based on expected probabilities for bankruptcy prediction. Expert Systems with Applications 36, 5297–5303 (2009)
15. Kumar, P.R., Ravi, V.: Bankruptcy prediction in banks and firms via statistical and intelligent techniques: a review. European Journal of Operational Research 180, 1–28 (2007)
16. Lee, K., Booth, D., Alam, P.: A comparison of supervised and unsupervised neural networks in predicting bankruptcy of Korean firms. Expert Systems with Applications 29, 1–16 (2005)
17. Negakis, C.: Robustness of Greek business failure prediction models. International Review of Economics and Business 42(3), 203–215 (1995)
18. Olson, D., Delen, D., Meng, Y.: Comparative analysis of data mining methods for bankruptcy prediction. Decision Support Systems 52, 464–473 (2012)
19. Platt, J.: Using sparseness and analytic QP to speed training of support vector machines. In: Kearns, M.S., Solla, S.A., Cohn, D.A. (eds.) Advances in Neural Information Processing Systems, vol. 11. MIT Press, MA (1999)
20. Quinlan, J.R.: C4.5: Programs for machine learning. Morgan Kaufmann, San Francisco (1993)
21. Sarkar, S., Sriram, R.S.: Bayesian Models for Early Warning of Bank Failures. Management Science 47(11), 1457–1475 (2001)
22. Shin, K., Lee, T., Kim, H.: An application of support vector machines in bankruptcy prediction model. Expert Systems with Applications 28, 127–135 (2005)
23. Thomaidis, N., Gounias, G., Zopounidis, C.: A fuzzy rule based learning method for corporate bankruptcy prediction. In: ACAI 1999, Chania, Greece (1999)
24. Tsai, C.-F., Hsu, Y.-F.: A Meta-learning Framework for Bankruptcy Prediction. Journal of Forecasting (2011), doi:10.1002/for.1264
25. Tsakonas, A., Dounias, G., Doumpos, M., Zopounidis, C.: Bankruptcy prediction with neural logic networks by means of grammar-guided genetic programming. Expert Systems with Applications 30(3), 449–461 (2006)
26. Tseng, F.-M., Hu, Y.-C.: Comparing four bankruptcy prediction models: logit, quadratic interval logit, neural and fuzzy neural networks. Expert Systems with Applications 37(3), 1846–1853 (2010)
27. Verikas, A., Kalsyte, Z., Bacauskiene, M., Gelzinis, A.: Hybrid and ensemble-based soft computing techniques in bankruptcy prediction: a survey. Soft. Comput. 14, 995–1010 (2010)
28. Witten, I., Frank, E., Hall, M.: Data Mining: Practical Machine Learning Tools and Techniques, 3rd edn. Morgan Kaufmann (2011) ISBN 978-0-12-374856-0
29. Zijiang, Y., Wenjie, Y., Guoli, J.: Using partial least squares and support vector machines for bankruptcy prediction. Expert Systems with Applications 38, 8336–8342 (2011)