David Quintana · Yago Saez · Asuncion Mochon · Pedro Isasi

Published online: 6 June 2007

Abstract Bankruptcy prediction has long time been an active research field in finance. One of the main approaches to this issue is dealing with it as a classification problem. Among the range of instruments available, we focus our attention on the Evolutionary Nearest Neighbor Classifier (ENPC). In this work we assess the performance of the ENPC comparing it to six alternatives. The results suggest that this algorithm might be considered a good choice.

Keywords Evolutionary computation · Bankruptcy · Evolutionary nearest neighbor classifier · Finance · Artificial intelligence

1 Introduction

Bankruptcy prediction has long been an active research field in finance [1–5]. Nevertheless, the number of parties that are affected by corporate failure is considerable. Investors, auditors, creditors or employees among many others, have a lot to gain from accurate forecasts regarding the solvency of companies.

D. Quintana (✉) · Y. Saez · P. Isasi
Department of Computer Science, Universidad Carlos III de Madrid, Avda. Universidad, 30, Leganes 28911, Spain
e-mail: david.quintana@uc3m.es

Y. Saez
e-mail: yago.saez@uc3m.es

P. Isasi
e-mail: pedro.isasi@uc3m.es

A. Mochon
Department of Applied Economics, UNED, Pº Senda del Rey, 11, Madrid 28040, Spain
e-mail: amochon@cee.uned.es

There are two major ways to deal with bankruptcy prediction. The first one, the structural approach, involves detailed projections of financial statements and a thorough understanding of the economics of the firm to be studied. The second one, to which this research belongs, is a statistical approach. This category would cover those efforts that understand this as a classification problem, which could be tackled looking for patterns relating to a number of relevant parameters.

This domain has been the subject of prediction efforts by means of very different techniques such as k-nearest neighbor, multiple discriminant analysis, logit models, artificial neural networks or classification trees among many others. We must note that since the pioneering works such as the one carried out by Odon and Sharda [6], the attempts to predict corporate distress with artificial neural networks are especially abundant. Anyone interested might find of interest the paper by Atiya [7] surveying these efforts. In terms of predictive accuracy, there is mixed evidence, but so far it seems that multi layer perceptrons tend to fare a bit better. This can be observed in references that compare different alternatives in the same sample [8–10].

The approach that we suggest is based on the Evolutionary Nearest Neighbor Classifier (ENPC) developed by Fernández and Isasi [10]. Among its features we can highlight the fact that this classifier is fully integrated. Both the size of the classifier and the learning algorithm cannot be split into specialized blocks to be taken care of by different techniques. Moreover, one of the main advantages offered by this algorithm is the lack of necessity for initial parameters to be defined by the user to guide the process.

ENPC has been proved to offer competitive results in synthetic domains such as squared spiral, uniform distributed data and popular benchmarks such as the iris data set

and the Pima Indians diabetes database from the UCI repository. This research constitutes the first attempt made to test its performance in the early bankruptcy prediction domain.

The remainder of the paper is structured in the following way. In Sect. 2, we introduce the ENPC. Section 3 presents the explanatory variables and describes the data. Section 4 deals with the results of the empirical analysis and, finally, Sect. 5 covers the summary and conclusions.

2 The evolutionary nearest neighbor classifier (ENPC)

The algorithm that we present as a competitive alternative in the domain is the ENPC [10]. This classifier can be included within the supervised nearest prototype class, as it assigns patterns to labeled prototypes depending on a distance measure. This family labels those patterns that are closer to a prototype with the same class as the prototype. This general description fits a good number of alternatives. However, there can be important differences among them according to issues such as the way of defining the number of prototypes to be used or the initial set to start from. ENPC dynamically finds the answer to both questions and, unlike most of the alternatives, does not rely on any parameter to be defined by the user.

This particular system is based on a set of prototypes that control a region defined by the patterns that are closer to them, in terms of the sum of squared errors. Each of these prototypes has a quality measure that considers the number of patterns in its region and whether those patterns belong to the same class as the prototype or not. The driving force behind the algorithm is the effort of the prototypes to enhance their quality relying on several operators (mutation, reproduction, fight and die) which allow them to do so.

Following Fernández and Isasi [10], we provide a brief description of the algorithm. The training process consists of eight stages represented in Fig. 1.

1. Initialization. The process starts with a single prototype, whose initial location is irrelevant. There are no learning parameters to be defined as the algorithm will self-adjust automatically.
2. Information gathering. At the beginning of each iteration the algorithm gathers the information regarding prototypes, classes and pattern sets required for the following five stages.
3. Mutation. The algorithm allows for the changes in the class associated to each prototype, depending on the distribution of the data classes of the patterns that are closer. The aim of the mutation operator is to perform this function.
4. Reproduction. This stage introduces new prototypes in the classifier. Each prototype has the chance of introducing additional prototypes if it considers that it would enhance the probability of controlling patterns that belong to the same class.

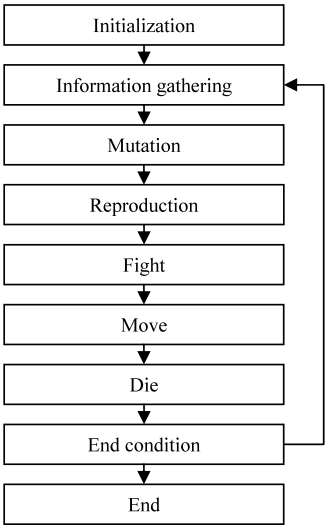

Fig. 1 ENPC algorithm flow

5. Fight. This operator allows a prototype to gain control of patterns owned by neighboring prototypes. The process is implemented by means of a roulette system where probabilities are granted depending on the homogeneity of the regions controlled by each prototype.
6. Move. The previous stages are likely to result in changes in the distribution of classes and prototypes, hence the need to re-allocate the prototypes in order to increase the performance of the system. This is achieved by relying on the second step or Lloyd iteration, to make a local optimization.
7. Die. This step of the process prunes the prototype set discarding the prototypes with a probability inverse to their quality.
8. End condition. The population evolves until the user decides to stop. The stop criteria might be target accuracy, a number of iterations, a mix of the previous two or a convergence to any of the first two.

3 Variables and data

3.1 Variables

In 1968, Altman published an article [1] in which he tried to determine a set of explanatory variables that could be useful to discriminate, out of a sample of manufacturers, which companies would eventually file for bankruptcy and which would keep operating. The result was the identification of five variables that he used to develop one of the most influential models in bankruptcy prediction, the Z-score. These variables have been widely used to test different bankruptcy prediction models differently [11–13].

Table 1 Predictive variables

Name	Definition
WC/TA	Working Capital / Total Assets
RE/TA	Retained Earnings / Total Assets
EBIT/TA	Earnings Before Interest and Tax / Total Assets
MVE/TA	Market Value of Equity / Total Assets
S/TA	Sales / Total Assets
CR	Current Assets / Current Liabilities

Table 2 Correlation matrix of predictive variables

	WC/TA	RE/TA	EBIT/TA	MVE/TA	S/TA	CR
WC/TA	1.00	0.31	0.27	0.32	−0.05	0.56
RE/TA	0.31	1.00	0.61	−0.03	0.09	−0.02
EBIT/TA	0.27	0.61	1.00	−0.06	0.28	0.02
MVE/TA	0.32	−0.03	−0.06	1.00	−0.22	0.68
S/TA	−0.05	0.09	0.28	−0.22	1.00	−0.21
CR	0.56	−0.02	0.02	0.68	−0.21	1.00

The set of explanatory variables that we will use in our analysis is provided in Table 1. The list basically mirrors the financial items suggested by Altman plus the current ratio (Current Assets / Current Liabilities). This additional variable is supposed to be a good indicator of short-term solvency and has also been used in the past [13, 14].

The list does not mean to be exhaustive by any means. The literature on bankruptcy is abundant and so are the potentially interesting ratios. The selected ones have a bearing on the going-concern and, as we already mentioned, have been proven popular and useful. However, this does not mean that there is not ample room for inconsistencies in the financial ratios and misclassified output. The identification of the ultimate explanatory variable set is an open issue that still puzzles researchers in the financial side. This adds complexity to the task since it is not likely that a perfect classification based on these ratios is at all possible.

3.2 Data

The sample consists of 552 US companies, 138 of which went bankrupt between the years 1995 and 2004. The group that showed solvency issues was identified using COMPUSTAT, and it includes all the companies for which all the required information was available. However, it excludes utilities, financial services and transportation companies since they are structurally different and have bankruptcy environments that are different from the rest [14]. The date of bankruptcy filings was obtained through inspection of documents filed in the Securities and Exchange Commission and accessible through the Electronic Data Gathering, Analysis, and Retrieval System (EDGAR). For each of the troubled companies, we include three non-bankrupt comparables. These are companies which operate in the same industry and had a comparable size in terms of total assets a year prior to bankruptcy. There does not seem to be consensus in the literature regarding the appropriate sample break. On one hand, there are some authors that suggest that a bankrupt/non-bankrupt company split that is close to the one observed in the real world is likely to result in better classification systems [9]. On the other, there is a vast majority which tends to include from one to four sound companies for every bankrupt example. The main advantage of including more than one is the increase in sample size. This, however, comes at the price of potentially biasing the classification methods to predict soundness. Given the fact that some of the techniques that we will be using are quite data intensive, the alternative would have been increasing the number of bankrupt companies by extending the considered time period. The risk of facing a structural change due to changes in the general economy that might affect the results has led us to proceed the way we have already mentioned.

The financial information required to perform the analysis was also obtained from COMPUSTAT. For each distressed company we consider the financial information reported two years before. Whenever the company went into bankruptcy during the last quarter, we considered the precedent year. The information regarding the matching comparables is reported at the same point in time. Table 2 shows the cross-correlations.

4 Analysis

The aim of this paper is to show the relative performance of ENPC in the domain of bankruptcy prediction. To do so, we have measured and compared the forecasting ability of algorithm to the one achieved by alternative approaches.

We will start breaking the sample into two subsamples. The first one, made up of 331 companies, will be used as a training set to fit the models that will be subsequently used to predict the rest of the original set of data. As we mentioned before, the composition of the global is divided into 25% bankrupt and 75% non-bankrupt companies. The subsamples are balanced. The training set is split 24.77%/75.23% and the test set 25.34%/74.66%. We first run the program on the train set and then we test the set of identified prototypes in the rest of the sample.

The algorithm has been compared to six classifiers: Naive Bayes [15, 16], logistic regression with a ridge estimator [17], C4.5 [18], PART [19], support vector machine trained with sequential minimal optimization [20] and multilayer perceptron (MLP). The implementation selected for all of them was the one provided in WEKA [21]. In every

Table 3 Summary results on test sample

	NB	LR	C4.5	PART	SVM	MLP	ENPC
Accuracy	73.30%	76.47%	71.95%	74.66%	74.66%	79.19%	80.09%
T. I Error	76.79%	85.71%	96.43%	99.99%	99.99%	76.79%	71.43%
T. II Error	9.70%	2.42%	4.85%	0.00%	0.00%	1.81%	2.42%

case but the MLP, they were run using the default parameters unless a random seed was required. Whenever this was the case, five different seeds were tested. For the MLP, the number of training iterations was increased to 1500 cycles and different combinations of learning rates, random seeds, and the number of nodes in the hidden layer was tested. In this case, we report the results obtained by the best performing network configuration.

The results are reported in Table 3. For each classifier considered: naive Bayes (NB), logistic regression (LR), C4.5, PART, support vector machine (SVM), multilayer perceptron (MLP) and ENPC, we show the global prediction error. In addition to that, we provide additional information regarding the kind of error committed by each system. On one hand, we define Type I error as misclassifying a failed firm as healthy. On the other hand, Type II errors would occur whenever non-distressed companies are being classified as bankrupt.

In terms of global predictive power, the results are consistent with the results obtained in similar studies. Even though there is not a big difference among the approaches, MPL seems to perform better than the rest and C4.5 offers substandard results. Naive Bayes is not useful either since both of them show a degree for accuracy below 74.66%. The absolute coincidence of PART and SVM is due to the fact that both systems reach the same solution, which is identifying the most abundant class in the training sample and using it to predict every pattern in the test set. ENPC offers the best global results although they are not significantly different from the ones achieved by the artificial neural network (MLP). The results were exactly the same regardless of whether the input data was normalized to a [0, 1] range or not.

The solution found by ENPC consisted of four prototypes. Three of them were labeled as "sound" and the fourth one as "bankrupt". To achieve this, the system was run with a double stop condition, either a perfect classification or running out of generations. The latter, set to 200, was the first one to occur.

In this domain, there is a significant difference between committing one kind of error or the other. From a practical point of view, the penalty associated with misclassifying a company as sound is more important than in the opposite case. For this reason, it is worthwhile going beyond the global results and looking into the confusion matrices reported in Appendix. If we analyze the kind of error committed by each system we observe that, consistent with the evidence from previous studies, the identification of bankruptcy is more challenging. The number of bankrupt companies classified as non-bankrupt as a percentage of total firms in distress is very high. It is likely that additional variables that have not been identified yet, or a substantial increase in the number of bankrupt companies in the sample, might enhance the results. Despite the high probability of error, ENPC outperforms the other approaches. This system beats the second best performers; MLP, also beats NB by more than five percent, and it is well ahead of LR and C4.5. As we mentioned before, SVM and PART do not classify any company as bankrupt and, therefore, show the worst possible results by this standard. Obviously, this very same weakness pays off when you consider Type II error and that is exactly what is evident when we inspect the last row of the table. Comparatively, the probability of misclassifying sound companies is very low. The composition of the sample biases is the classifiers which label any company as sound, since the likelihood of making a mistake is much lower. The performance of ENPC in this respect is average. Naive Bayes is significantly worse when it comes to correctly label non-distressed companies. The results of the MLP are slightly better than ENPC and worse than the naive predictions by PART and SVM.

In this particular case, the global error is the suitable magnitude for comparing the different classifiers. Focusing the evaluation on Type I error, which would clearly favor ENPC against the rest, would not be fair since classifiers were meant to minimize that percentage, not the probability of making any specific kind of error. Adding asymmetric cost functions, which is seldom done perhaps due to the difficulty in choosing the right values, might offer different results. Another point worth considering is that this exercise is meant to be used as a benchmark. The sample is very difficult to predict, but the degree of difficulty is the same for all the systems and, therefore, the results should be analyzed in relative terms.

5 Conclusions

In this work, we have assessed the performance of Evolutionary Nearest Neighbor Classifier (ENPC) in the early bankruptcy detection domain. In order to do so, we have compared results offered by ENPC in a matched sample of 552 companies to those provided by six alternatives:

Naive Bayes, logistic regression with a ridge estimator, C4.5, PART, support vector machine trained with sequential minimal optimization and multilayer perceptron (MLP).

This supervised nearest prototype classifier is based on the effort of a set of prototypes to enhance a quality index by means of several operators (mutation, reproduction, fight and die). This index is specific for each prototype and depends on the number of patterns in its region and whether those patterns belong to the same class as the prototype or not.

In this experiment its performance is in line with, if not better than, the best classic alternative, multilayer perceptron, with the advantage of offering a lower probability of misclassifying companies likely to go bankrupt without sacrificing much accuracy in the classification of sound companies. ENPC offers a big advantage over MPS in terms of ease of use. Finding the right network structure and training parameters is a difficult task. The search for the right combination of parameters might be costly in time and computational resources. Unlike MLP, ENPC finds the required parameters autonomously on execution time without the need for any other argument but a stopping criterion.

All the above mentioned suggests that ENPC is a competitive algorithm worth including in the toolkit of those who deal with bankruptcy prediction from a statistical point of view.

Acknowledgements This article has been financed by the Spanish founded research MCyT project OPLINK, Ref: TIN2006-08818-C04-02, and CAM-UC3M project Computación con Inspiración Biológica para la Minería de Datos, Ref: UC3M-DEC-05-029.

Appendix. Confusion matrices. Test sample

Confusion matrices for naive Bayes (NB), logistic regression (LR), support vector machine (SVM), multilayer perceptron (MLP), C4.5, PART and ENPC. Real class in rows (B: bankrupt / NB: non-bankrupt), output from the models in columns.

	NB	
	B	NB
B	13	43
NB	16	149

	LR	
	B	NB
B	8	48
NB	4	161

	SVM	
	B	NB
B	0	56
NB	0	165

	MLP	
	B	NB
B	13	43
NB	3	162

	C4.5	
	B	NB
B	2	54
NB	8	157

	PART	
	B	NB
B	0	56
NB	0	165

ENPC	
B	NB
16	40
4	161

References

1. Altman EL (1968) Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. J Financ 23(3):589–609
2. Altman EL (1982) Accounting implications of failure prediction models. J Account Audit Financ 6:4–19
3. Beaver W (1996) Financial ratios and predictors of failure. In: Empirical research in accounting: selected studies. Suppl J Account Res 4:71–111
4. Johnson C (1970) Ratio analysis and the prediction of firm failure. J Financ 25:1166–1168
5. Zavgren C (1983) The prediction of corporate failure: the state of the art. J Account Lit 2:1–38
6. Odom M, Sharda R (1990) Neural network model for bankruptcy prediction. In: Proceedings of the IEEE International Conference on Neural Networks II, San Diego, USA, pp 163–168
7. Atiya AF (2001) Bankruptcy prediction for credit risk using neural networks: a survey and new results. IEEE Trans Neural Netw 12(4):929–935
8. Tam K, Kiang M (1992) Managerial applications of the neural networks: the case of bank failure predictions. Manag Sci 38(7):416–430
9. McKee TE, Greenstein M (2000) Predicting bankruptcy using recursive partitioning and a realistically proportioned data set. J Forecast 19:219–230
10. Fernández F, Isasi P (2004) Evolutionary design of nearest prototype classifiers. J Heuristics 10(4):431–454
11. Coats PK, Fant LF (1993) Recognizing financial distress patterns using a neural network tool. Financ Manag 22(3):142–155
12. Wilson RL, Sharda R (1994) Bankruptcy prediction using neural networks. Decis Support Syst 11:545–557
13. Zhang G, Hu MY, Patuwo BE, Indro DC (1999) Artificial neural networks in bankruptcy prediction: general framework and cross-validation analysis. Eur J Oper Res 116:16–32
14. Raghupathi W, Schkade LL, Raju BS (1991) A neural network approach to bankruptcy prediction. In: Proceedings of the IEEE 24th annual international conference on systems sciences, Hawaii, USA, vol 4, pp 147–155
15. Duda RO, Hart PE (1973) Pattern classification and scene analysis. Wiley, New York
16. John GH, Langley P (1995) Estimating continuous distributions in Bayesian classifiers. In: Proceedings of the eleventh conference on uncertainty in artificial intelligence, Montreal, Canada, pp 338–345
17. le Cessie S, van Houwelingen JC (1992) Ridge estimators in logistic regression. Appl Stat 41(1):191–201
18. Quinlan JR (1993) C4.5: programs for machine learning. Kaufmann, Los Altos
19. Frank E, Witten IH (1998) Generating accurate rule sets without global optimization. In: Proceedings of the fifteenth international conference on machine learning, Madison, USA, pp 144–151
20. Platt J (1999) Fast training of support vector machines using sequential minimal optimization. In: Schoelkopf B, Burges C, Smola A (eds) Advances in kernel methods—support vector learning. MIT, Cambridge, pp 185–208
21. Witten IH, Frank E (2005) Data mining: practical machine learning tools and techniques, 2nd edn. Morgan, San Francisco