Expert Systems with Applications 39 (2012) 7512–7523

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

A hybrid device for the solution of sampling bias problems in the forecasting of firms' bankruptcy

Fernando Sánchez-Lasheras [a], Javier de Andrés [b,*], Pedro Lorca [b], Francisco Javier de Cos Juez [c]

[a] University of Oviedo, Department of Construction and Manufacturing Engineering, Campus de Gijón, Edificio 5, 33204 Gijón, Spain
[b] University of Oviedo, Faculty of Economy and Business, Department of Accounting, Avda. del Cristo s/n, 33006 Oviedo, Spain
[c] University of Oviedo, Department of Exploitation and Exploration of Mines, c/ Independencia No. 13, 33004 Oviedo, Spain

ARTICLE INFO

Keywords:
Bankruptcy
Self Organized Maps (SOM)
Multivariate Adaptive Regression Splines (MARS)
Construction
Sampling bias

ABSTRACT

This paper proposes a new approach to the forecasting of firms' bankruptcy. Our proposal is a hybrid method in which sound companies are divided in clusters using Self Organized Maps (SOM) and then each cluster is replaced by a director vector which summarizes all of them. Once the companies in clusters have been replaced by director vectors, we estimate a classification model through Multivariate Adaptive Regression Splines (MARS). For the test of the model we considered a real setting of Spanish enterprises from the construction sector. With this procedure we intend to overcome the sampling-bias problems that matched-pairs models often suffer. We estimated two benchmark models: a back propagation neural network and a simple MARS model. Our results show that the proposed hybrid approach is much more accurate than the benchmark techniques for the identification of the bankrupt companies.

1. Introduction

Correct resource allocation decisions are critical to guarantee the survival of banks and other lenders. So, bankruptcy forecasting models are key tools to help bank managers/officers in their investment/lending decisions. From the late 1960 onwards many models have been developed and tested. During the last years the importance of such systems is even higher due to the current financial crisis, which demands an even more careful management of financial resources. Furthermore, under Basel II Accord recommendations (Bank for International Settlements (BIS), 2006), banks which choose to develop their own empirical model to quantify required capital for credit risk (internal rating-based approach) are required to maintain less capital than those using the standardized approach. So, an accurate device to estimate loan default probabilities lets a financial entity to minimize the resources held as reserves and therefore to reach a higher level of profitability.

According to Sueyoshi and Goto (2009a), research on bankruptcy-based performance assessment can be classified into three broad categories. First, those studies centered on a particular model, which test how such model performs in comparison with others. Second, research focused on the selection of an appropriate set of variables to implement a particular model. The third category comprises papers which investigate the bankruptcy process.

Among these categories, the first is the one which has received most attention by researchers. The tested models are mainly statistical methodologies (for a review of the most outstanding studies see Balcaen & Ooghe, 2006; Keasey & Watson, 1991, among others) and artificial intelligence techniques (for a review see, e.g., Aziz & Dar, 2006; Ravi Kumar & Ravi, 2007).

Ravi Kumar & Ravi (2007) discuss the models which have been most frequently used in studies focused in insolvency prediction via intelligent systems. These models are Fuzzy Logic (FL), Neural Networks (NN), Genetic Algorithms (GA), Case-Based Reasoning Systems (CBR), Rough Sets (RS), Support Vector Machines (SVM), Decision trees (DT), Data Envelopment Analysis (DEA) and Hybrid Systems (HS).

Among these, HS are the most promising. These combine two or more intelligent techniques in several forms to derive the advantages of all of them. HS have received considerable attention from researchers as they amplify the advantages of the intelligent techniques while simultaneously nullifying their disadvantages. Most HS require a considerable amount of data to reach to accurate estimations. This is not a problem nowadays, as there exist publicly available databases containing financial information of listed and unlisted firms.

However, studies using HS for bankruptcy prediction suffer from a drawback which is that the majority of them estimate the model upon the basis of a sample in which non-failed companies are underrepresented. In most cases a matched-pairs design is used. The selection of non-failed firms is arbitrary, which makes the model to achieve a high in-sample percentage of correct classi-

* Corresponding author. Tel.: +34 985103287; fax: +34 985109599.
E-mail addresses: sanchezfernando@uniovi.es (F. Sánchez-Lasheras), jdandres@uniovi.es (J. de Andrés), plorca@uniovi.es (P. Lorca), fjcos@uniovi.es (F.J. de Cos Juez).

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2012.01.135

fications but it is likely to be inaccurate for failure prediction in new cases drawn from a real population.

Another strategy is to consider a "real" population as the sample. That is, to consider all the companies for which we have financial information available. However, as only a very small percentage of firms enter into financial distress in a normal economic situation, such samples are very unbalanced. This causes coefficient instability and leads to poor performance ability of the models.

As an alternative to both strategies we propose a HS model where, upon the basis of a real population of firms, data are preprocessed to summarize the information of healthy firms. So, the initial unbalanced sample is transformed into a balanced one which retains the main features of the healthy firms. Self Organized Maps (SOM) is used in this stage. Then a classification device is developed upon the transformed sample, for which we use the Multivariate Adaptive Regression Splines (MARS) approach. The results are compared with benchmarks which are popular in bankruptcy prediction literature. As an important application of the combined approach, this paper applies it to the solvency assessment of Spanish construction firms.

The remainder of the paper is structured as follows. Section 2 revises prior studies on bankruptcy prediction using HS. Section 3 is devoted to build the database. Section 4 describes the algorithm and the analytical procedures we used. Section 5 comments on the main results, including the benchmark techniques applied. Finally, Section 6 is devoted to the summary and main conclusions, including also some further research avenues.

2. Prior bankruptcy research using hybrid systems

Basically, there are four types of HS which have been applied to financial distress prediction:

– Hybrid Algorithms (HA), where two or more intelligent algorithms are tightly integrated to form a new classification device (i.e., GA-trained NN, neuro-fuzzy systems).
– Ensemble Classifiers (EC), which consist of multiple single classifiers whose decision is combined to form that of the combined system, usually by applying a voting scheme.
– Feature Selectors (FS). In these systems, an algorithm is used for the selection of the predictors of failure among a list of feasible variables and another model is used to predict the bankruptcy status using the selected indicators.
– Clustering and Classificatory devices (CC). These HSs preprocess the financial information on the failed and non-failed firms and identify groups based on similarities. The grouping information is used in the subsequent estimation of a classification model.

Tables 1–4 contain a summary of a selection of the most outstanding studies on financial failure prediction using each type of HSs.

It must be pointed out that if the bankruptcy prediction models are eventually to be used in a predictive context, the estimation samples of failing and non-failing firms should be representative of the whole population of firms (Ooghe & Joos, 1990). Nevertheless, in the great majority of the hybrid prediction models revised in Tables 1–4, the samples are not representative of the whole population. Most studies oversample failing companies because of the low frequency rate of failing firms in the economy. A common strategy is the use of matched pairs samples (on the basis of size, sector, and/or age). This can lead to biased parameter estimates especially if the sample is made up of failed firms and very sound companies. In that case the model will achieve a high percentage of correct classifications but it is likely to be inaccurate for failure prediction in new cases drawn from a real population.

An alternate sampling strategy is to consider a real population. As Foglia, Iannotti, and Marullo-Reedtz (2001) point out, this procedure increases the variance of the estimates of coefficients due to the data imbalance between sound and unsound firms. An additional drawback is that, having into account that in a normal economy most companies are non-bankrupt, to classify all the firms as "not-bankrupt" would let the model reach a high percentage of correct classifications. To avoid this, the algorithm can be designed to consider the different misclassification costs (the costs of classifying as insolvent a company which is solvent are much lower than those of the opposite error). Such a model will pay more attention to accurately classifying the failing companies at the expense of more misclassifications of non-failing firms.

However, the estimation of the different misclassification costs is not straightforward as it depends on the financial decision to be taken. Furthermore, such estimation is a subjective task as it also depends on the risk profile of the agent who makes the decision.

As an alternative to both approaches, we propose a method which enables the formation of a sample which is representative of the main features of the population but retains the balanced design and the stability of the coefficients.

Our proposal is a hybrid method in which sound companies are divided in clusters according to their financial similarities and then each cluster is replaced by a director vector which summarizes all of them. The clustering process is made by means of a SOM procedure. The most relevant reasons for choosing SOM among the different methods for clustering are the following two: first, this technique was specifically designed for multidimensional datasets, and is able to take advantage of their complexity and second, unlike other methods for data-reduction and clustering, this family of algorithms is characterized by a learning process that is constantly updated as it takes more information from the input data, improving the output dynamically over the training stage and therefore producing more reliable results.

Prior to the calculation of clusters, sound companies are divided into two groups:

(a) Companies which are actually sound but whose financial features have a certain degree of similarity with those of failed ones. These are called "borderline" companies.
(b) Companies which are sound and whose financial features are clearly different from those of bankrupt companies.

The clustering process is carried out separately for each group of firms. Although the idea of considering a "grey zone" or group of doubtful firms has been previously introduced by other researchers (see, i.e., Alam, Booth, Lee, & Thordarson, 2000; Ooghe, Joos, & Devos, 1992; Tseng & Lin, 2005), we made the discrimination between sound and doubtful firms on a multivariate basis by using a non-euclidean distance measure (the Mahalanobis distance).

Once the companies in clusters have been replaced by director vectors, we estimate a classification model through MARS. The reason for choosing MARS as the second part of the hybrid system lies in the fact that this technique is a flexible procedure, which models relationships that are nearly additive or involve interactions with fewer variables (Hastie & Tibshirani, 1990). MARS builds flexible models by fitting piecewise linear regressions; that is, the nonlinearity of a model is approximated through the use of separate regression slopes in a limited number of intervals of the variable space. This is made by using a procedure which is inspired by the recursive partitioning technique governing Classification and Regression Trees (CART) algorithm (Breiman, Friedman, Olshen, & Stone, 1984). Such features make it especially suitable for the