



## Partial Least Square Discriminant Analysis for bankruptcy prediction

Carlos Serrano-Cinca<sup>\*</sup>, Begoña Gutiérrez-Nieto

Department of Accounting and Finance, University of Zaragoza, Spain

### ARTICLE INFO

#### Article history:

Received 29 June 2011

Received in revised form 13 September 2012

Accepted 25 November 2012

Available online 3 December 2012

#### Keywords:

Bankruptcy  
Financial ratios  
Banking crisis  
Solvency  
Data mining  
PLS-DA

### ABSTRACT

This paper uses Partial Least Square Discriminant Analysis (PLS-DA) for the prediction of the 2008 USA banking crisis. PLS regression transforms a set of correlated explanatory variables into a new set of uncorrelated variables, which is appropriate in the presence of multicollinearity. PLS-DA performs a PLS regression with a dichotomous dependent variable. The performance of this technique is compared to the performance of 8 algorithms widely used in bankruptcy prediction. In terms of accuracy, precision, *F*-score, Type I error and Type II error, results are similar; no algorithm outperforms the others. Behind performance, each algorithm assigns a score to each bank and classifies it as solvent or failed. These results have been analyzed by means of contingency tables, correlations, cluster analysis and reduction dimensionality techniques. PLS-DA results are very close to those obtained by Linear Discriminant Analysis and Support Vector Machine.

© 2012 Elsevier B.V. All rights reserved.

### 1. Introduction

Bankruptcy prediction from financial ratios using mathematical models is a classical approach in data mining research. Since Beaver's [7] pioneer work, based on univariate ratio analysis, many different techniques have been employed in this context. Altman [3] used Linear Discriminant Analysis (LDA); Ohlson [42] used Logistic Regression (LR); Marais et al. [35] used Decision Trees such as Id3, C4.5 and Random Trees; Tam and Kiang [57] used Multilayer Perceptron (MLP), a neural network model and K-Nearest Neighbors (KNN); Serrano-Cinca [54] and du Jardin and Séverin [18] applied Self Organizing Feature Maps; Fan and Palaniswami [21] used Support Vector Machine (SVM) and Sarkar and Sriram [52] applied Naive Bayes (NB). Techniques of ensembles, such as Boosting or Bagging, have been applied by Foster and Stine [26], who combined C4.5 and Boosting; while Mukkamala et al. [39] combined Bagging and Random Tree (BRT). See Olson et al. [44] for a recent comparative analysis on data mining methods for bankruptcy prediction. This paper applies Partial Least Square Discriminant Analysis (PLS-DA) to the 2008 banking crisis in the USA. To the best of our knowledge, this technique has not previously been applied to bankruptcy prediction.

Partial Least Squares (PLS) regression combines features from Principal Component Analysis (PCA) and Multiple Linear Regression [60]. PLS is a mathematical estimation approach that builds a model by

sequentially adding data points so that model parameters are continuously updated. PLS models are popular in structural model building and in regression analysis. PLS-DA is based on the PLS regression model, being the dependent variable a categorical one. This approach is useful for classification tasks [6]. For example, PLS-DA is a standard tool in Chemometrics, the science that analyzes chemical data [62]. Its attraction resides in its ability to successfully address the problem of multicollinearity [6,59].

Multicollinearity is a major problem when building models based on financial data. Financial analysts request tools able to accurately predict distress from financial data. But they also want to model bankruptcy symptoms, by identifying the relevant variables. However, it is difficult to select an appropriate model when using collinear data, as there is no unique data reduction method, and different orderings of the hypothesis testing procedure result in different models, something that affects interpretability. Multicollinearity and model selection procedures in regression have long been debated in Econometrics; see for example Hendry and Mizon [28]. Only when regressors are orthogonal, any model selection procedure ends up by identifying the same model. This is certainly not the case with financial data. A possible solution is to do regression analysis on principal components, which are orthogonal by definition, but this always results in a loss of information in the data set since only a small number of components are employed in the distress prediction model. In this paper we consider if the PLS-DA methodology offers a possible way forward in this context. The paper poses two research questions: **RQ1**: How does PLS-DA perform in terms of model interpretability, when facing correlated data, compared to other techniques? The case of multicollinearity in financial information will be specially studied. **RQ2**: How does PLS-DA perform in terms of classification accuracy, compared to other techniques?

<sup>\*</sup> Corresponding author at: Department of Accounting and Finance, Fac. Economía y Empresa, Univ. Zaragoza, Gran Vía 2, Zaragoza (50.005), Spain. Tel.: +34 876 554643; fax: +34 976 761769.

E-mail address: [serrano@unizar.es](mailto:serrano@unizar.es) (C. Serrano-Cinca).

URL: <http://ciberconta.unizar.es/charles.htm> (C. Serrano-Cinca).

The paper compares PLS-DA with 8 algorithms previously used in distress prediction (LDA, LR, MLP, KNN, NB, SVM, C4.5, and BRT). The data used relates to the 2008–2011 USA banking crisis. The paper joins a long literature on bankruptcy prediction. Many studies analyze bankruptcy, and compare the results obtained by the different techniques, see Ravi Kumar and Ravi [50] and Demyanyk and Hasan [16] for a revision. It can be concluded that no technique is clearly better than others, because it depends on the problem analyzed and the performance measure chosen, Caruana and Niculescu-Mizil [9]. LDA is optimum if data satisfied some statistical properties. Unfortunately, financial ratios are not normally distributed, as shown by Ezzamel and Mar-Molinero [20], and this calls for the use of alternative techniques. LR needs fewer requisites, and being regression based, results can be interpreted in a standard way. When relationships are not linear, MLP performs well, because neural networks are universal approximators, as Hornik et al. [29] have shown. But results by MLP are difficult to interpret. By contrast, decision tree algorithms, such as C4.5, have the advantage of resulting in general rules that can be incorporated in the design of expert systems. Boosting and Bagging techniques improve the performance of decision trees. The strength of PLS-DA stems from its capability to deal with multicollinearity, because it transforms original variables into orthogonal components. It is also robust to missing data and skew distributions, see Cassel et al. [10]. It can also deal with the ‘too few cases/too many variables’ problem. These problems are common in financial information; a priori, PLS-DA seems to be a promising technique.

In order to compare PLS-DA with the rest of the techniques, a benchmark on which to analyze performance has to be established. But performance can be measured in many ways. Ferri et al. [24] analyze the behavior of 18 different performance metrics; being accuracy, precision and *F*-score the most popular. One technique can obtain better results than other by using a given performance measure; but it can obtain poorer results by using another performance measure, Caruana and Niculescu-Mizil [9]. But it can also be the case that two techniques achieve the same performance, but one of them correctly classifies some cases and the other one correctly classifies different cases. In this paper we analyze – by means of contingency tables and Phi correlations – coincidences and divergences in classifications produced by the 9 techniques. Each technique generates a score for each bank, which can be interpreted as a solvency measure. This resulted in a two-way table with banks in rows and the 9 scores in the columns. This table can be analyzed with multivariate analysis, by using Cluster Analysis (CA) and Categorical Principal Components Analysis (CATPCA). This approach results in a taxonomy of techniques, it being specially interesting to know the techniques closest to PLS-DA.

The paper analyzes the 2008 banking crisis. Financial data are taken from the Federal Deposit Insurance Corporation (FDIC), an independent agency created by the US Congress to maintain stability and public confidence in the banking system. The crisis has caused the failure of an important number of USA banks. 140 banks failed in 2009, 157 in 2010, and there are still banks in difficulties in 2011, since the crisis is still not over. The FDIC [22] affirms that there are a good number of banks, 884, on their “problem list.” The procedure followed by this paper, based on pattern recognition, will allow the identification of these banks.

The following section presents the conceptual background, with an emphasis on presenting the PLS-DA technique. The third section presents the empirical study, focusing on performance comparison. Finally, conclusions and bibliography are presented.

## 2. Conceptual background

Failure prediction from the analysis of financial ratios has been a fruitful research line, as shown by the state of the art revisions [16,50,64]. The use of financial ratios as a tool to analyze financial

distress has a long pedigree. The work by Beaver [7], who used univariate analysis, marks a starting point. But companies' health has a multivariate nature, which is not captured by Beaver's approach. Altman [3] introduced the use of multivariate techniques, particularly Linear Discriminant Analysis (LDA). LDA derives a linear combination of ratios which best discriminate between failed and non-failed firms. An overall score, known as Altman's *Z*-score, can be calculated from LDA, and used to estimate the financial health of a company. LDA depends on several restrictive assumptions, such as linearity, normality, or independence among predictors. But these assumptions are seldom satisfied [19]. Ohlson [42] took a step forward by applying Logistic Regression (LR). In common with LDA, LR weights financial ratios and obtains a score. LR and LDA share the same underlying linear model, but LR does not require multivariate normality, it only requires that variables are distributed in the exponential family [34]. This implies that it is not necessary for some of the restrictive assumptions of LDA to hold. Despite these differences, both LDA and LR obtain very similar results in practical applications with financial information. According to the empirical study by Lo [34], the null hypothesis that LDA and LR are equivalent may not be rejected.

A different approach is followed by Curram and Mingers [14] who use decision trees. Decision trees employ a recursive partitioning algorithm to induce rules on a given data set. Widespread algorithms in failure prediction are Id3 and C4.5 by Quinlan [47]. Decision trees have been successful in obtaining useful bankruptcy prediction rules. However, decision tree algorithms may suffer from overfitting: the algorithm reduces training set error at the cost of an increased test set error.

Tam and Kiang [57] employed Multilayer Perceptron (MLP), a neural network model to predict bankruptcy. Hornik et al. [29] proved that under certain weak conditions, multilayer feedforward networks perform as a class of universal approximators. This explains their powerful capability for classification. A meta-analysis performed by Adya and Collopy [1] reveals that neural networks outperformed alternative approaches in 19 out of the 22 analyzed studies; but they warn that the bias against publication of negative results may mean that successful applications are over-represented in the published literature. It is difficult to obtain an optimal combination of MLP parameters that produces the best prediction performance: there is a large number of controlling parameters, such as the number of hidden layers, the number of hidden nodes, the learning rate, the momentum term, epochs, and transfer functions [55].

*k*-Nearest Neighbors (KNN) is one of the most straightforward machine learning algorithms [25], which is very useful when there is no prior knowledge about the distribution of the data. Employed by Tam and Kiang [57] and Park and Han [45] in bankruptcy prediction, KNN is basis fundamental for many Case Base Reasoning (CBR) developments, widely used in bankruptcy prediction.

The aim of Support Vector Machine (SVM) is to find the best hyperplane that produces the largest separation between failed and non-failed firms. It has been applied to bankruptcy prediction by Fan and Palaniswami [21]. SVM are based on few restrictive assumptions, and they are obtaining very promising results in failure prediction [50].

The Naive Bayes (NB) algorithm is based on conditional probabilities. NB looks at the historical data and uses Bayes' Theorem to calculate the probability of an event occurring given the probability of another event that has already occurred. Sarkar and Sriram [52] used NB for predicting failure. Among the advantages of NB, Sun and Shenoy [56] highlight that it does not have any requirements on the underlying distributions of the variables; it does not require complete information for observations, and is easy to understand because the relationships among variables are explicitly represented. They empirically compare LR and NB, finding that there is no significant difference (at the 5% level) between both models' performance.

In recent years, techniques of ensembles combining multiple classifiers, like Boosting or Bagging (or bootstrap aggregating) have been developed. They have been applied to bankruptcy prediction by Foster and

متن کامل مقاله

دریافت فوری ←

**ISI**Articles

مرجع مقالات تخصصی ایران

- ✓ امکان دانلود نسخه تمام متن مقالات انگلیسی
- ✓ امکان دانلود نسخه ترجمه شده مقالات
- ✓ پذیرش سفارش ترجمه تخصصی
- ✓ امکان جستجو در آرشیو جامعی از صدها موضوع و هزاران مقاله
- ✓ امکان دانلود رایگان ۲ صفحه اول هر مقاله
- ✓ امکان پرداخت اینترنتی با کلیه کارت های عضو شتاب
- ✓ دانلود فوری مقاله پس از پرداخت آنلاین
- ✓ پشتیبانی کامل خرید با بهره مندی از سیستم هوشمند رهگیری سفارشات