PREDICTING BUSINESS FAILURE USING DATA-MINING METHODS

Sami BEN JABEUR
Youssef FAHMI

http://www.ipag.fr/fr/accueil/la-recherche/publications-WP.html

IPAG Business School
184, Boulevard Saint-Germain
75006 Paris
France
PREDICTING BUSINESS FAILURE USING DATA-MINING METHODS

Sami BEN JABEUR, Youssef FAHMI
Université de Bretagne-Sud (IREA)
benjabeursami@yahoo.fr, fyoussef@hotmail.com

Abstract: The aim of this paper is to compare two statistical methods in predicting corporate financial distress. We will use the PLS (Partial Least-Squares) discriminant analysis and support vector machine (SVM). The PLS discriminant analysis (PLS-DA) regression is a method connecting a qualitative variable dependent to a unit on quantitative or qualitative explanatory variables. The SVM may be viewed as non-parametric techniques. It is based on the use of so-called kernel function which allows optimal separation of data. In this work we propose to use a French firm for which some financial ratios are calculated.

Keywords: financial distress prediction, PLS discriminant analysis, Support Vector Machine

1. Introduction

This paper joins within the framework of the research works on the models of forecast of bankruptcies, which could be used to detect the money troubles of the SME (small and medium-sized enterprise).

The research works on the forecast of the financial difficulties of companies are of a big importance for all the partners of a company. From a manager point of view, have tools of forecast of the financial failures allows to take in time strategic measures and management corrective appropriate to prevent these failures (Hooted and al 2007). For other partners, such tools contribute to reduce the informative asymmetry with the company, to detect quickly the vulnerable companies and to optimize the allocation of their capital (financier, human being, social). The forecast of the financial difficulties so constitutes a means of diagnosis of the performance of companies and ends in a classification of these last ones among the failing or not failing companies. The forecast of Failing companies can help in particular to prevent the difficulties before they are translated by a financial crisis and to set up the necessary measures (assistants, restructuring) before it is too late.

The use of the models of forecast of the bankruptcy of companies can vary according to tools and methodologies used, the reserved sample, the explanatory variables chosen as well as the method of validation of the results.

Before presenting the main methods of forecast used in this research work, it is useful to call back(to remind) quickly the definitions of the financial distress.

2. Definition of the financial distress

During these last years, the annual flow of failures of companies did not stop growing and this tendency becomes more marked during the periods of crisis. According to the INSEE( NATIONAL INSTITUTE FOR STATISTICS AND ECONOMIC STUDIES), the number of failures of companies affected 52103 in 2010.

Table 1: Failures of French companies (published in the BODACC) between 1993 and 2010

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of failing companies</td>
<td>55174</td>
<td>57494</td>
<td>52640</td>
<td>56858</td>
<td>53261</td>
<td>47525</td>
<td>42132</td>
<td>38346</td>
<td>36941</td>
</tr>
<tr>
<td>Year</td>
<td>2002</td>
<td>2003</td>
<td>2004</td>
<td>2005</td>
<td>2006</td>
<td>2007</td>
<td>2008</td>
<td>2009</td>
<td>2010</td>
</tr>
<tr>
<td>Number of failing companies</td>
<td>38202</td>
<td>43068</td>
<td>42368</td>
<td>43138</td>
<td>40407</td>
<td>42840</td>
<td>51254</td>
<td>53739</td>
<td>52103</td>
</tr>
</tbody>
</table>
The movement of the failures concerns in particular the companies from 10 to 50 employees and those whose turnover is included between 500,000 euro and 2 million euro. In 2009 and in 2010, the fiscal measures of reflation saved between 15,000 and 30,000 companies. SME (SMALL AND MEDIUM-SIZED ENTERPRISE), defined as the companies of less than 250 employees and less than 50 million euro of turnover, benefited from 40% of these measures, that is 6.4 billion euro.

Even if the failure can have multiple reasons (fall in demand and difficulties of access to outlets, problem of management, bad strategic choices and errors of management, competence and formation of the team, the obsolescence of the production tool, the undercapitalization, the frauds), it is generally translated by financial distress and a deterioration of the financial situation of the company (reduction of the activity, the decrease of the profitability, the cash flow problems, the financial imbalance).

Wruck (1990) defines the financial distress as being a situation where the cash flows are insufficient to cover the current bonds. These obligations can include the debts suppliers, the expenses. According to Baldwin and Scott (1983), when the situation of a company degrades in the point where she cannot face her financial constraints, the firm enters a state of financial distress. These same authors assure that this situation is the result of a bad economic situation, a decline of their performances and a low quality of their management. From their part, Ooghe and Van Wyneersch (1996) explain the financial distress in particular by the problems of liquidity (because of an insufficient profitability and of a lack of resources) and of a high level of debts which pulls problems of solvency of the company.

Several surrounding areas of forecast of the financial distress of companies were proposed since the end of the sixties, using financial ratios and accounting data. The purpose is to differentiate the failing companies and the not failing companies and to build an explanatory model of the failure of companies. The most used techniques are the univariate approach, the multivariate approach, the discriminating analysis multiple shelf space, the multiple regression, the logit regression, the models based on the artificial neuronal approach. These methods of forecast present a joint representation concerning the evaluation of the risk of defect of companies, in the sense that they base themselves on the financial analysis and the exploitation of the knowledge ex comment of the future of companies (Redone 2004).

In this work, we are going to apply the discriminating approach and the SVM approach in the detection of the failing companies. Both approaches will lean on a sample of 800 companies during period 2006 in 2008, as well as on an recourse to 33 financial ratios.

3. Analysis method

3.1. The discriminating analysis PLS

There are several versions of the algorithm of univariate regression PLS1. They different at the level of the normalizations (standardizations) and the intermediate calculations, but they give quite the same regression. According to Bastien, Esposito and Tenenhaus (2005), the algorithm of the discriminating regression PLS (PLS-DA) can decompose as follows: we can repeat this process by using in the same way residues Y2, X21, X2k of the regressions of Y, X1, Xk on t1, t2.

The algorithm PLS1 spells then:

Stage 1: X0=X; Y0=Y (initialization)
Stage 2: for h=1,…, r :
Stage 2.1: \( w_h = \frac{X'_{h-1}Y_{h-1}}{Y'_{h-1}Y_{h-1}} \)
Stage 2.2: Standards \( W_h \) in 1
Stage 2.3: \( t_h = \frac{X'_{h-1}W_h}{W'_{h}W_h} \)
Stage 2.4: \( p_h = \frac{X_{h+1}}{t_h} \)

Stage 2.5: \( X_h = X_{h-1} - t_h p_h \)

Stage 2.6: \( c_h = \frac{y_{h+1}}{t_h} \)

Stage 2.7: \( u_h = \frac{y_{h+1}}{c_h} \)

Stage 2.8: \( y_h = y_{h-1} - c_h f_h \)

Once we retained the number of the constituents, we make a discriminating analysis on these constituents and not on the variables of origin.

3.2. The principles of the method SVM

Without going into the technical details, we expose very briefly the principles of the method "Support Vector Machines".

Support Vector Machines (SVM) are techniques of classification based on the statistical theory of the learning (Vapnik on 1995, 1998). The SVM can be used to resolve problems of regression, that is predict the numerical value of a variable, or a discrimination, that is decide to which class belongs a sample.

As method of classification, the approach SVM represents a kind of discriminating analysis generalized, made in a space of rather big dimension so that exists a linear separation. She takes place in two stages. In the first one, a not linear transformation makes pass of the space of origin in a space of bigger dimension but endowed with a scalar product. Assist stage: in we look for a linear separator:

\[ f(x) = a \cdot x + b, \]

which is a hyperplane to separate the whole learning points so that all points of the same class are on the same side of the hyperplane. The latter must simultaneously fulfill two conditions:
- it separates the groups (accuracy of the model), in the sense \( f(x) > 0 \Rightarrow \text{class A} \) and \( f(x) \leq 0 \Rightarrow \text{class B} \);
- it is the farthest from all observations (model robustness), knowing that the distance of an observation \( x \) to the hyperplane is \( \frac{a \cdot x + b}{\|a\|} \), the margin being \( 2 \|a\| \)

Given points \((x_i, y_i)\) with \( y_i = 1 \) if \( x_i \) is in A and \( y_i = -1 \) if \( x_i \) is in B, find the linear separator \( f(x) = a \cdot x + b \) equivalent to finding a pair \((a, b)\) simultaneously satisfies two conditions:
- For all \( i \), \( y_i (a \cdot x_i + b) \geq 1 \) (good separation).
- \( \|a\|^2 \) is minimum (maximum margin).

SVMs seek, among all possible hyperplanes, the one that maximizes the distance between the hyperplane decision and the nearest points of each class.

When the two populations are not perfectly discriminated or separated but overlapping, a terme must be added to to each of the two previous expressions.

The solution \( f(x) \) expressed as a function of inner products \( x \cdot x' \). After transformation \( \Phi \), it is expressed in terms of inner products \( \Phi(x) \cdot \Phi(x') \). The amount \( k(x, x') = \Phi(x) \cdot \Phi(x') \) is called the nucleus. In the algorithm, the kernel \( k \) and not \( \Phi \) chosen and we can calculate \( k(x, x') \) without showing \( \Phi \). The calculations are then made in the original space and become simpler and faster. This is why it is called core or kernel machine.

Examples of cores include:
- Linear \( k(x, x') = x \cdot x' \);
- polynomial \( k(x, x') = (x \cdot x')^d \); if \( d = 2 \), \( x = (x_1, x_2) \) and \( \Phi(x) = (x_1^2, \sqrt{2}x_1x_2, x_2^2) \) then \( \Phi(x) \cdot \Phi(x') = (x_1x_1' + x_2x_2')^2 = (xx')^2 \).
- Gaussian (RBF) \( k(x, x') = e^{-\frac{(x-x')^2}{2\sigma^2}} \); one of the most commonly used
- Sigmoid \( k(x, x') = \tanh(\kappa(x \cdot x') + \theta) \) where \( \kappa \) is the gain and \( \theta \) the threshold.

4. Methodological choice

4.1. The construction of the sample

Our approach to data collection involves several steps: the choice of the database, the selection of companies and the indicators of financial failure.

4.1.1. Presentation of data:

We use the database DIANE (instant access to data from French companies for economic analysis) for our sample. This database provides access to a fund composed of more than one million businesses, as these are the companies that publish their annual accounts within the registries of commercial courts. Collected by Coface Services, information on company accounts and is enriched with lots of related information. With ten years of historical accounts, DIANE is the basis of financial information and general on French enterprises the most comprehensive.

4.1.2. The survey sample:

A company was considered deficient if it has been a first event declaration to the judicial tribunal of commerce during 2009. The data studied is organized so that the accounting year is available for 2008, 2007 and 2006. The final sample obtained upon completion of this rigorous selection process consists of 800 companies split into two sub-samples: 400 healthy firms and 400 failing firms. This choice was dictated by itself, because of constraints arising from the availability of identifying information of companies.

4.2. The selection of indicators and financial ratios

Selecting financial ratios has been a more logical and methodological choice in order to provide a relevant and credible battery capable of meeting the objectives and expectations of external reviews. Financial ratios were selected according:
- their recurrence in French literature (Bardos, 1995, Bank of France’s work) and international literature (Altman, 1968, 1984, Conan and Holder 1979; Rose and Giroux,1984, Remade, 2004).
- their relevance to financial analysis, incorporating the basic ratios in most existing models of detecting bankruptcy: liquidity ratios, profitability management, productivity and financial structure. Thus, a series of 33 ratios (R01 to R33) was retained among those commonly used in literature and which have a significant informational content in the analysis of the financial situation of the company.

5. Results

5.1. The classification of entreprises by the method of PLS discriminate analysis

PLS-DA analysis allowed retaining six components, the classification results are presented in the following table:

\( \begin{align*}
\text{ Component 1 } & \quad \text{Component 2} \\
\text{ Component 3 } & \quad \text{Component 4} \\
\text{ Component 5 } & \quad \text{Component 6}
\end{align*} \)
According to this table, we can say that the results obtained using the PLS-DA method are attractive compared to discriminate analysis (DA1). Indeed, the rate of correct classification of companies is about 96.50%, an error of type 1 (percentage of failing firms considered healthy) of 5.75% and a type 2 error (percentage of non-failing companies considered risky) of 1.25%. According to this analysis PLS-DA was able to identify 377 healthy firms, a rate of 94.25%, with only five failing firms are classified as healthy, a rate of 98.75%.

In the group of healthy companies, the model holds 126 companies as being dysfunctional; thus they are really not. Also in the group of failing firms, 137 of them are considered healthy, which is not the case. The study of the “two years before failure” model, reveals a decrease in classification rate of healthy firms from 94.25% to 68.5%; and 98.75% of companies falling to 65.75%. So the more the forecast is, the less the accuracy of the model is.

Looking at this table, we can say that the results obtained using the PLS method are attractive compared to discriminate analysis. Indeed, the rate of correct classification of companies is about 60.5%, a second type error of 38.75% and a first type error is about 40.25%.

Regarding the predictive power of the model in the span of time from two to three years before the failure, we find a drop in clearance rates in healthy firms from 68.5% to 61.25% and 65.75% of failing companies to 59.75%. Thus, the greater the forecast horizon is far, the more the ability of the model is reduced. In conclusion, we detect a superiority of PLS regression in one, two and three years before failure compared to traditional discriminate analysis. Indeed, one year before failure, the rate of correct classification of firms is 95.9% for AD1 and 96.5% for the analysis PLS-DA1. Two years before the failure, the rate of correct classification is 64.8% to 67.125% for AD2 and PLS-DA 2. Regarding the third year, the rate of correct classification is 59.25% for AD 3 and 61.25% for PLS-DA3. These results were confirmed by previous work applied to scientific fields other than the prediction of failure by Nguyen and Roke (2004), Bastien et al. (2005), who believe that the role of restrictive assumptions of discriminate analysis from the PLS regression can be successfully applied to financial data.

5.2. The classification of enterprises by the SVM method

The results of the classification of enterprises by the SVM method are presented in the following table:
Table 2: Validation of SVM model

<table>
<thead>
<tr>
<th>Group Trust real</th>
<th>T-1</th>
<th>T-2</th>
<th>T-3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1, 2 or 3 year Before the failure</td>
<td>Nature of the firm</td>
<td>S</td>
<td>D</td>
</tr>
<tr>
<td>number</td>
<td>healthy</td>
<td>370</td>
<td>30</td>
</tr>
<tr>
<td>Defaulting</td>
<td>11</td>
<td>389</td>
<td>400</td>
</tr>
<tr>
<td>%</td>
<td>healthy</td>
<td>92.5</td>
<td>7.5</td>
</tr>
<tr>
<td>Defaulting</td>
<td>2.75</td>
<td>97.25</td>
<td>100</td>
</tr>
</tbody>
</table>

Table 2: Validation of SVM model

T-1: One year before failure, the rate of correct classification of firms is around 94.875%, a type 2 error of 7.5% and a type 1 error of 2.75%. The SVM method was able to identify 370 of healthy companies, a rate of 92.5% while 30 failed businesses are classified as healthy firms. Similarly, it was able to identify 389 firms failed, a rate of 97.25%. Healthy companies classified as failing firms are 11 in number.

T-2: Two years before the failure, the SVM approach gave a rate of correct classification of the order of 65.625% with a type 2 error of 38% and a type 1 error of 30.75%. This method was able to identify 248 healthy 248 healthy firms, a rate of 62% while 152 companies are ranked among failed business. In addition, it was able to identify 277 failed companies, a rate of 69.25%. Healthy companies classified as failing firms are 123.

T-3: Three years before the failure, the SVM approach gave a rate of correct classification of approximately 59.25% with a type 2 error of 29% and a type 1 error of 52.5%. This method was able to identify 284 healthy companies, a rate of 71% while 116 companies are ranked among failed businesses. In addition, it was able to identify 190 failed companies, a rate of 47.5%. Healthy companies classified as failing firms are 210.

5.3. Comparison of two methods of classification

The following table compares the performance in anticipation of the PLS-DA and SVM approaches.

<table>
<thead>
<tr>
<th>Healthy firms</th>
<th>Failing firms</th>
<th>Good ranking</th>
</tr>
</thead>
<tbody>
<tr>
<td>PLS-DA</td>
<td>SVM</td>
<td>PLS-DA</td>
</tr>
<tr>
<td>T-1</td>
<td>94.25 %</td>
<td>92.5 %</td>
</tr>
<tr>
<td>T-2</td>
<td>68.5 %</td>
<td>62 %</td>
</tr>
<tr>
<td>T-3</td>
<td>61.25 %</td>
<td>71 %</td>
</tr>
</tbody>
</table>

Table 3: PLS-DA and SVM Performance approaches

The rates of correct classification provided by the two methods are very close. The rate of correct classification of non-failing firms is the share of non-failing firms correctly classified all non-failing firms. Similarly, the rate of correct classification is failing firms from failing firms correctly classified in all the failing companies.

The comparison of the overall performance of the two methods shows a clear superiority to the PLS-DA. A year before the failure, the PLS-DA is more effective than SVM approach since it leads to a correct classification rate of 96.5%, against 94.875 for the SVM method. Two years before the failure rates were 67% for the PLS-DA and 65.625% for the SVM method. Finally, three years before failure, the rates were 60.5% for the PLS-DA and 59, 25% for the SVM method.
The SVM method is superior to PLS-DA method only in terms of the classification of two failed businesses before failure (69.25% against 65.5% for the PLS-DA), and for classification non-failing firms three years before failure (71% against 61.25% for the PLS-DA).

6. Conclusion

As has been demonstrated by the previous studies and recent research in this field, we find that the predictive power is weakened for models that use less information, or to recent models that are intended to be used to predict business failure over a more distant period of time. In conclusion, we detected a clear strength in the PLS-DA method in terms of one and three years before failure-analysis compared to the SVM method. We have shown that the principles of PLS regression could be without difficulty discriminate analysis, PLS regression has an advantage in terms of classification, taking into account important variables in the analysis, it can also solve the problem of correlation between different variables while the majority of conventional methods grounded before such a problem.

7. Bibliography

Banque de france, Gourioux C., Foulcher S., Tiamo A. (2003), La structure par termes des taux de défaits et ratings, Cahiers Etudes et recherchesd'Observatoiredesentreprises, Direction desEntreprises, Banque de France, 1.34.
Li H., Jie Sun J. (2009), Predicting business failure using multiple case-based reasoning combined with support vector machine, Expert Systems with Applications 36, 10085–10096.
Pages J., Tenenhaus M. (2001), Multiple factor analysis combined with PLS path modeling, Revue de Statistique Appliquée, XLIV (2), 35-60.