Credit Scoring for SME using a Manifold Supervised Learning Algorithm

Armando Vieira & Ning Chen
Instituto Superior Engenharia do Porto
Porto, Portugal
far.quasar@gmail.com, cng@isep.ipp.pt

Bernardete Ribeiro
Departamento Engenharia Informática
Universidade de Coimbra, Portugal
bribeiro@dei.uc.pt

***Abstract*— We propose a credit scorecard algorithm based on the supervised ISOMAP to rate SME. By projecting the companies balance sheet data into a one dimensional component we obtain a smoother distribution of ratings while increasing the discriminatory capability of each rate in terms of the probability of default. The method is applied to a large dataset of French SME.**

Keywords- Credit Risk, Credit Scoring, Supervised Learning, Isomap.

I. INTRODUCTION

Credit risk analysis is a very important and actual topic. In some cases the total loss due to bankruptcies can be as high as 5% of the nominal GDP. The causes of bankruptcy can broadly be assigned in two categories: intrinsic failure and network effects. The first can be originated by anemic sales; irresponsible management; accumulated deficit; low capitalization (lack of working capital or very high interest rates); random cause; build-up of inventory; excessive business investment. The network effects are caused by an aftershock of another company's bankruptcy (excess of bad debts) or a non-enforceable accounts receivable [1,2]. In this work we deal with the first type of bankruptcy causes .

However, in order to advice the analyst or the investor, more than a detecting the bankruptcy we need to have a credit scorecard to rate the company in terms of how close it is from default or the probability of becoming bankrupt. The difficulty of computing these ratings have several causes, but two are paramount: 1) all algorithms are retrospective, i.e, they project historical data into the future and 2) the only empirical evidence we have to gauge the accuracy of the models is when the company becomes bankrupt, i.e. the labeled data from the training set is either 0 – non-bankrupt or 1 - bankrupt. The algorithm has to infer the real status of the company between those extremes from a classifier that will learn a bimodal distribution picked around 1 and 0.

To address this problem we propose a new scoring algorithm based on semi-supervised manifold learning: the SSA. Many applications of data classification and data mining deals with a large number of unlabeled examples. Supervised nonlinear dimensionality reduction can be used as a preprocessing step before classification. The rationale here is to map the high-dimensional data space into a lower dimensional space where classification methods do not suffer from the curse of dimensionality. As the explicit mapping is not found by the algorithm some learning methodology must be used. Our approach uses a set of training labels in the data set to provide a better construction of features and improve learning.

In this paper we address the problem of building a smooth and reliable credit scorecard algorithm. We want to explore the intrinsic structure of the financial data of companies in order to effectively compare them using a meaningful metric. Our aim is to improve credit scoring and classification with unlabeled examples under the assumption that the data resides on a low-dimensional manifold within a high dimensional representation space. In this case we make the most drastic dimensionality reduction, namely to project data onto a 1-dimensional manifold.

This paper is organized as follows. In Section 2 we describe our dataset, Section 3 deals with the topic of manifold learning and describe the algorithm. Section 4 presents the results.

II. CHARACTERIZATION OF THE DATASET

We used a sample obtained from Diane, a database containing about 100 000 financial statements of French SME companies [9]. The sample consists of financial ratios of industrial French companies, for the years of 2002 to 2007, with at least 10 employees. From these companies, 1511 were declared bankrupted in 2007 and 2272 presented a restructuring plan to the court for approval by the creditors. We decided not to distinguish these two categories as both signal companies in financial distress. From these dataset we build a balanced sample with 600 financial distressed firms, most of them small to medium size, with a number of employees from 10 to 800, corresponding to the year of 2007 – thus we are making bankruptcy prediction one year ahead.

Our database contains many cases with missing values, especially for defaults companies. For this reason we sorted the default cases by the number of missing values and selected the examples with 10 missing values at most. A final set of 600 default examples was obtained. In order to obtain a balanced dataset we selected randomly 600 non-default examples resulting in a set of 1200 examples.

The remaining missing data was treated as follows. For the ratios of the years 2003 and 2006 each missing value was replaced by the value of the closest available year; for 2004 and 2005, if values of the next and previous years were available, each missing value was replaced by their mean, otherwise it was replaced by the remaining value. In some cases there was no data available for a ratio in any of the years. In this very few cases the missing data was replaced by the median value of the ratio in each year. Finally, all ratios were logarithmized and then standardized to zero mean and unity variance.

This dataset includes companies from a wide range of sectors - see for Figure 1. However, due to the large number of attributes available, we used several ranking algorithms to select the most relevant. From the initial 30 financial ratios defined by COFACE and included in the Diana database, we select only the 8 most relevant, namely: Number of employees, Liquidity ratio, Financial Debt/Equity, Financial Debt/Cashflow, Cashflow/Turnover, Working Capital Needs/Turnover, Total Assets/Turnover and EBITDA Margin. All ratios were normalized to zero mean and unity variance.

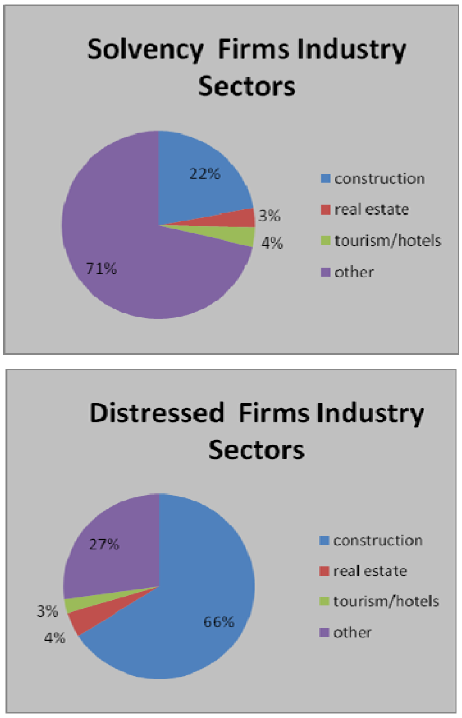


Figure 1. Characterization of the dataset in terms of sectors of activity.

III. Manifold Learning

In many real applications observational high-dimensional data can be cast into low-dimensional manifolds embedded in low-dimensional spaces, provided a suitable representation procedure is found. Instead of working with points in a high-dimensional space, classification and prediction algorithms can be easily used in these low-dimensional spaces sought from the embedded learning process [3, 5, 6].

Attempting to uncover this manifold structure in a data set is known as manifold learning. Manifold methods include a number of nonlinear approaches to data analysis that exploit the geometric properties of the manifold on which the data is supposed to lie. Manifold learning can be seen as an unsupervised feature extraction algorithm. We have previously applied it to the problem of bankruptcy prediction [4].

Although the structure inherent in many real world domains exhibit embedded low dimensional structures, as for example in image data or video frames, much research is going on in other areas.

A. *Unsupervised Map*

Unsupervised learning does not incorporate domain knowledge but its useful since in many cases we don't have label data. Unsupervised Isomap [3] consists of three main steps:

1) Estimates which points are neighbors on the manifold *M*, based on the distances $dX(i, j)$ between pairs of points *i*, *j* in the input space ***X*** by computing the weighted graph *G* of neighborhood relations given by the edges of weight $dX(i, j)$.

2) Estimates the geodesic distances between all pairs of data points in the manifold *M* by computing the shortest path distance on the *k*'s nearest neighbor graph built on the data set.

3) Applies classical Multi Dimensional Scaling (MDS) to the matrix of graph distances, constructing an embedding of the data in a *d*-dimensional Euclidean space *Y* that best preserves the manifolds estimated intrinsic geometry. Isomap assumes that there is an isometric chart that preserves distances between points.

```
The ISOMAP algorithm:
input: x1, · · · xn , k
1. Form the k-nearest neighbor graph with edge
   weights $W_{ij} := \|x_i - x_j\|$ for neighboring points
   $x_i, x_j$ .
2. Compute the shortest path distances between all
   pairs of points using Dijkstra's or Floyd's algorithm.
   Store the squares of these distances in $D$.
3. Return $Y := MDS(D)$
```

It is assumed that for nearby points in the high dimensional space the Euclidean distance is a good approximation of the geodesic distance whereas for distant points this is not true. Therefore, another technique described in ISOMAP is applied. It consists of building a weighted graph a *k*'s nearest neighbours where its edges are weighted by the Euclidean distances between nearby data points. Then a shortest path computation algorithm such as, Dijkstras or Floyds, will complete the calculus of the remainder geodesic distances.

B. *The supervised map*

The supervised version of the algorithm is based on the assumption that different features of the data can be captured by different dissimilarity measures. The algorithm builds up from a dissimilarity matrix to uncover the manifold embedded in the data. The dissimilarity matrix $D(x_i, x_j)$ between two sample points x_i and x_j is defined as [6]:

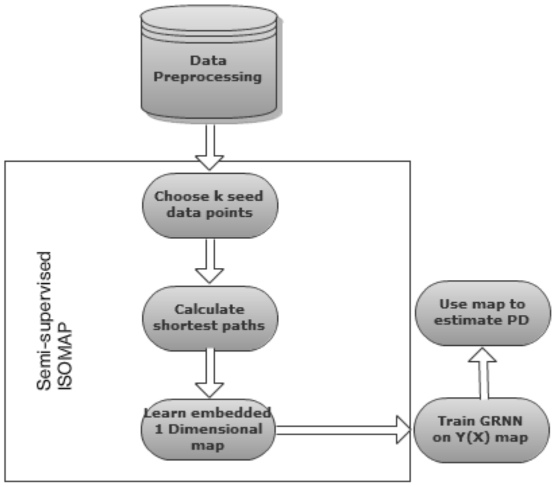
$$D(x_i, x_j) = \begin{cases} \sqrt{(a-1)/a} & \text{if } c_\text{i} = c_\text{j} \\ \sqrt{a} - d_0 & \text{if } c_i \neq c_j \end{cases} \qquad (1)$$

where $a = 1/e^{-d_{ij}/\sigma}$ with d_{ij} a distance measure (in our case Euclidean), σ a smoothing parameter (set according to the data 'density'), d_0 a constant ($0 \leqslant d_0 \leqslant 1$) and c_i, c_j are the data class labels. If dissimilarity between two samples is less than 1, points are in the same class, otherwise points are in different classes. The parameter d_0 allows that points in different classes to have a smaller value of dissimilarity, than those in the same class. In general the inter-class dissimilarity is larger than the intra-class dissimilarity (depending on the parameter d_0) conferring a high discriminative power on the method, which is very powerful for further classification. This approach was recently applied to our dataset with good results [4].

C. The semi-supervised algorithm

Since bankruptcy prediction is a high-dimensional unbalanced problem with incomplete information, both supervised and unsupervised approaches have its drawbacks. Supervised manifold learning is a powerful approach but it does not incorporate all the data since there is a lack of information about the real situation of a non-bankrupt company. Furthermore, if we use all points to train the Isomap, we may obtain a crispy map.

Our objective is to build a map that smoothly relates companies; from the worst possible situation to the healthier one. In order to accomplish this, we need a metric system to effectively compare companies through their financial situation. However, supervised Isomap may not be suitable because it forcefully distorts the original structure of the input data. This is particularly severe in the presence of noisy data.



Our approach uses labeled data (seeds) to constrain the learning of the Isomap [7] much in the same way as other seeded semi-supervised algorithms. We consider prior information in the form of manifold coordinates of certain data points. As a first step we choose a fraction of seed points at random. We run the supervised Isomap to constrain the map on this subset of points. Then we run the unsupervised Isomap to calculate the new MDS distances that does not violate the membership condition imposed by the seed points. Finally, after the implicit mapping Y is created, we train a generalized neural network (GNN) to retrieve an explicit map for fast recover of results.

IV. Results

We run the SSA algorithm using $k = 4$ neighbors and setting the parameter $d_0 = 0.5$. We used two different percentages of seeds for semi-supervised algorithm: 10% and 30%. Finally the general regression neural network (GNNR) was trained using a Matlab code.

TABLE I. Performance comparison of the algorithms. SSA – 10 and SSA – 30 corresponds to using 10% and 30% of seed points respectively

	Precision (%)	**Recall (%)**
S-Isomap	87.94	86.79
SSA – 30	85.77	84.44
SSA – 10	84.03	83.05

We test our algorithms using data from 2006 to make one year ahead prediction, i.e., failures in 2007. First we run the algorithm in order to determine the level of degradation in the accuracy of classification for not using all the labeled data - Table 1. We see that, even with 10% of seed points, the accuracy is not substantially compromised when compared with the full supervised version of Isomap - see [4]. From now on, we use 30% of seed points, so refer to SSA as SSA – 30. Note that our focus is not in the accuracy per se, but in obtaining a smooth and stable map to feed the credit score card algorithm.

In figure 2 we present the distribution of outputs over our dataset for the Logistic regression. Note that the distribution is strongly bimodal and with a strong overlap. In figure 3 we plot the same distribution for our algorithm - SSA. Note that in both categories we obtain a smooth Gaussian distribution – being the bankrupted distribution somehow skewed towards the right. From the fits of figure 3 we can estimate the probability of default, defined as

$$PD(x) = \frac{f(x)}{f(x) + g(x)}. \qquad (2)$$

where $f(x)$ is distribution non-bankrupt (healthy) companies and $g(x)$ the distribution of distressed (bankrupted) companies for the balanced dataset. In this figure we also plot the total expected number of defaults occurring at a given level:

$$N(x) = (f(x) + g(x))PD(x) \sim f(x)PD(x). \qquad (3)$$

In practice the most important aspect of the classifier algorithm is to use its discriminatory capabilities to bind a credit rating scorecard in order to classify the companies in terms of risk. We have already develop a stable rating algorithm rating which is more reliable than traditional approaches [10]. In this case we want to produce a smoother algorithm – in the sense of projecting the true distribution of companies ratings - even if we sacrifice the accuracy of the classifier.

Figure 4 we plot the distribution of ratings according to our algorithm compared to the widely used Logistic regression approach. Note that in this case we are using the full dataset according to Section 2. Rating levels were set by slicing the outputs into $M = 7$ equidistant levels. Note that the distribution of ranks is more uniform and the highest risk rates have a steeper decline.

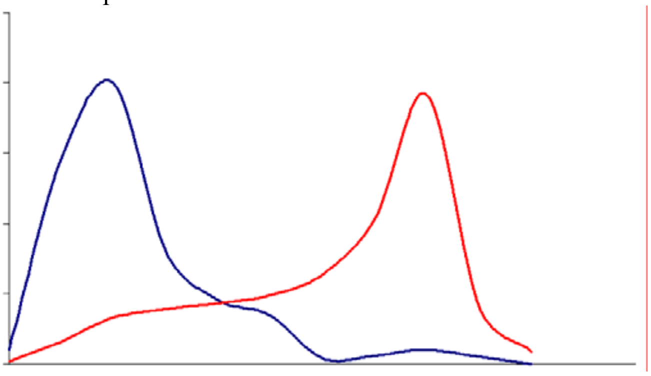


Figure 2 Output distributions for the Logistic regression. Red corresponds to bankrupt and blue to healthy companies. Arbitrary unities are used.

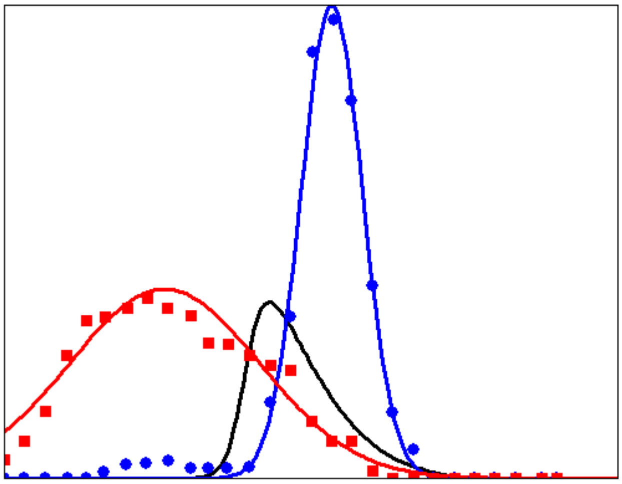


Figure 3. Output distributions of the classifier for SSA with k = 4. The continuous lines are Gaussians fits to the healthy (blue) and bankrupt (red) companies. The black curve is the number of bankrupcies $N(x)$ – Eq. (3).

V. Conclusions

We presented a scorecard algorithm for rating SME that is smoother without compromising the accuracy. In future work we want to test the stability of the algorithm over time.

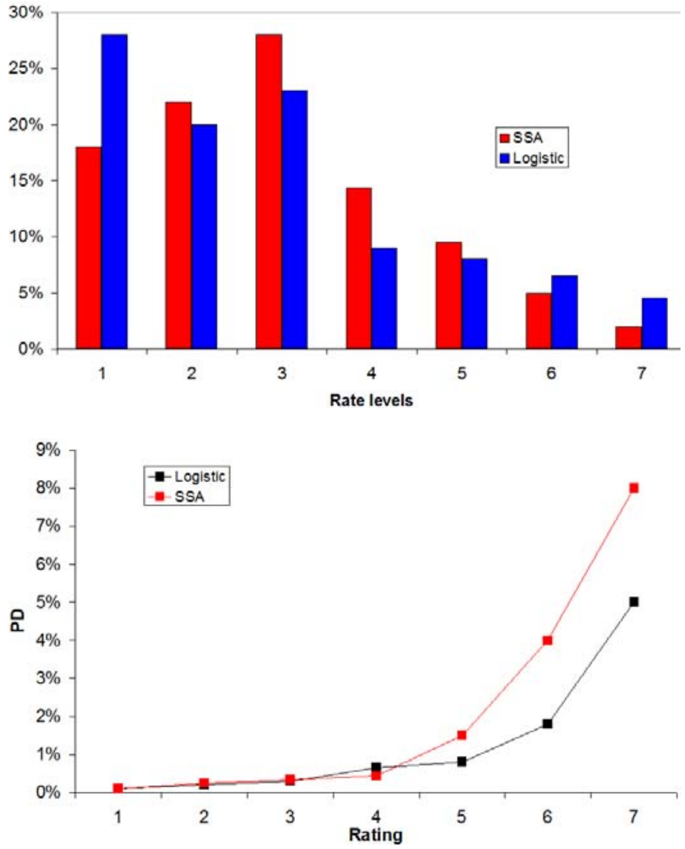


Figure 4. Ratings distribution using the Logistic algorithm and SSA (top graph) over the full dataset and the corresponding probability of default (PD) - bottom.

References

[1] A. Vieira and Joao Carvalho das Neves: "Improving Bankruptcy Prediction with Hidden Layer Learning Vector Quantization", European Accounting Review 15 (2), 253-271 (2006).

[2] A.F. Atiya Bankruptcy prediction for credit risk using neural networks: A survey and new results. IEEE Trans. Neural. Net. 12(4) (2001).

[3] J.B. Tenenbaum, de Silva, V., Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction. Science 290(5500) 2319–2323 (2000).

[4] B. Ribeiro, A Vieira, J Duarte, C Silva, J Carvalho das Neves, S Mukkamala, and A H Sung,"Bankruptcy Analysis for Credit Risk using Manifold Learning", ICONIP, Lecture Notes on Computer Science, Springer, Auckland (2008).

[5] S. Roweis, Saul, L.: Nonlinear dimensionality reduction by locally linear embedding. Science 290(5500) 2323–2326 (2000).

[6] X. Geng, Zhan, D.G., Zhou, Z.H.: Supervised nonlinear dimensionality reduction in visualization and classification. IEEE Transactions on Systems, Man and Cybernetics - Part B: Cybernetics 35(6) 1098–1107 (2005).

[7] X. Yang, Haoying Fu, Hongyuan Zha, and Jesse Barlow: Semi-supervised nonlinear dimensionality reduction. In Proceedings of the 23rd international conference on Machine learning (ICML '06). ACM, New York, NY, USA, 1065-1072 (2006).

[8] R. Chatpatanasiri and B. Kijsirikul, "A Unified Semi-Supervised Dimensionality Reduction Framework for Manifold Learning", Neurocomputing (2010)

[9] A. S. Vieira, João Duarte, B. Ribeiro and J. C. Neves: Accurate Prediction of Financial Distress with Machine Learning Algorithms", Proceeding of the International Conference on Artificial Neural Networks and Genetic Algorithms, Springer Lectures on Computer Science, ICANNGA 2009, Kuopio, Finland (2009).

[10] N. Chen, Armando Vieira, Bernardete Ribeiro, João Duarte, and João Neves. 2011. A stable credit rating model based on learning vector quantization. Intell. Data Anal. 15, 2, 237-250 (2011).