Analyzing Financial Statements with the Self-Organizing Map

Kimmo Kiviluoto
Helsinki University of Technology
Laboratory of Computer and Information Science
P.O. Box 2200
FIN-02015 Espoo, Finland
kimmo.kiviluoto@hut.fi

Pentti Bergius
Kera Ltd.
Enterprise Development and Financing
P.O. Box 559
FIN-33101 Tampere, Finland
pentti.bergius@kera.fi

Abstract

The Self-Organizing Map is used as a tool for analyzing financial statements, with the focus on bankruptcy prediction. The phenomenon of going bankrupt is analyzed qualitatively, and companies are also classified into healthy and bankrupt-prone ones.

In the qualitative analysis, the Self-Organizing Map is used in a supervised manner: both input and output vectors are represented in the weight vector of each unit, and during training, the whole weight vector is updated, but the best-matching unit search is based on the input vector part only.

In the quantitative analysis, three classifiers that utilize the Self-Organizing Map are compared to Linear Discriminant Analysis and Learning Vector Quantization. A modification of the Learning Vector Quantization algorithm to accommodate the Neyman-Pearson classification criterion is also presented.

1 Introduction

Assessing the probability of bankruptcy of an enterprise is one of the key issues in a credit granting decision. Besides analyzing the strategy, personnel etc. of the firm, the financiers usually perform an analysis of the financial statements. One standard approach has been to use a mathematical model based on Linear Discriminant Analysis, but a wide variety of other statistical techniques have also been proposed. Recently, models utilizing neural networks have been introduced and compared with the "traditional" techniques.

The importance of the problem has made it something of a benchmark test for different models. Usually, in these tests the problem has been reduced to a classification of companies into healthy and non-healthy ones. There are two characteristics common to many of the reported studies: they are based on fairly small data sets, and the proportion of the bankrupt firms is much higher in the data than in the total population from which the sample is selected. This makes the results somewhat difficult to interpret. With small data sets, especially when the results are not cross-validated, the differences in classifier performance cannot be clearly distinguished from statistical noise; with biased sample, one may also get an over-optimistic view of the classifier performance on the total population. In the present study, we have tried to avoid these problems by using a very large sample consisting of 4 898 financial statements, in which the ratio of healthy and bankrupt-prone firms is the same as in the base population.

2 The tools

The study consists of two parts: qualitative analysis and classification. Both parts utilize the Self-Organizing Map (SOM). In qualitative analysis, the SOM is used to form a "non-linear regression" from the input space into a plane; this makes it possible to visually examine the differences between firms that go bankrupt and those that do not. The idea is similar to that used in [1, 2, 3, 5]. In classification, SOM is used as a (possibly smoothed) vector quantizer.

2.1 Qualitative Analysis

The input vectors $\mathbf{x}^i$, $i = 1, \ldots, N$ consist of the normalized[1] financial indicators that are derived from the financial statements. In addition to these, an indicator variable b^i is associated with each input vector. The indicator variable is set to one if the company which gave the financial statement went bankrupt within certain time interval; otherwise it is set to zero. We shall denote the input vectors that are augmented with the indicator variable with $\hat{\mathbf{x}}^i \equiv [\mathbf{x}^{iT}, b^i]^T$.

In map training, the indicator variable is not used in searching the winner, and we have the familiar formula

$$c = \underset{j}{\operatorname{argmin}} \, ||\mathbf{x}^i - \mathbf{m}_j|| \tag{1}$$

for the winner unit index. However, after the winner has been found, also the "extra component" associated with each weight vector is updated, and the update rule becomes

$$\hat{\mathbf{m}}_j := \hat{\mathbf{m}}_j + \alpha(t)h(j,c)(\hat{\mathbf{x}}^i - \hat{\mathbf{m}}_j), \qquad \forall j \tag{2}$$

where $\hat{\mathbf{m}}$ is the augmented weight vector, $\alpha(t)$ is the learning rate, and $h(j,c)$ is the neighborhood function which here has the form of a Gaussian.

2.2 Classification

The classification of companies into healthy and non-healthy ones is done in two different ways: trying to minimize the total number of misclassifications, and using the Neyman-Pearson criterion, i.e. fixing the type I error (classifying a bankrupt company erroneously as a healthy company) to a suitable value, and within this constraint minimizing the type II error (classifying a healthy company erroneously as a bankrupt company). In practice, a classifier that is based on the Neyman-Pearson criterion would be the preferred one: type I error is much more costly than type II error, but because the proportion of non-bankrupt companies is higher, a classifier that minimizes the total number of misclassifications would pay more attention on minimizing the type II errors.

The classifiers used in this study are the following: Linear Discriminant Analysis (LDA), Learning Vector Quantization (LVQ), Self-Organizing Map (SOM-1 and SOM-2), and SOM-based Radial Basis Function Network (RBF-SOM). The SOM-based classifiers are only used here with the Neyman-Pearson criterion.

The SOM-based classifiers are briefly discussed below, as is also the modification of the LVQ for the Neymann-Pearson criterion. The LDA classifier is used in the usual straighforward manner.

2.2.1 SOM Classifiers

The SOM is used for classification in two different ways. In both of these, all the financial statements that are mapped on the same neuron are assigned the same class label, ie. the map consists of "bankruptcy units" and "non-bankruptcy units". The difference between the SOM models is in the labeling method of map units.

[1] The normalization scheme used in this study is histogram equalization for each indicator.

In the first model, which we shall call SOM-1, a simple voting scheme is used: for each unit n, $P(\text{bankruptcy}|n)$ is estimated to be the proportion of financial statements given by companies that had gone bankrupt. This conditional probability in turn is used as the classification criterion.

In the second model, SOM-2, we utilize a strategy that is similar to that used in qualitative analysis (see section 2.1). The map unit weight vectors consist of the four financial indicators, which are used in determining the winner unit, and an additional component that reflects the conditional probability of bankruptcy given the unit. Again, the classification decision is based on that conditional probability. The winner search and weight vector update formulas are as in equations (1) and (2).

In addition to these, there is also a third model that utilizes the SOM, here referred to as the RBF-SOM. It is a standard RBF network, in which the kernel centers are placed on the SOM weight vectors; the kernel widths β_j are set to

$$\beta_j = \rho \min_k ||\mathbf{m}_j - \mathbf{m}_k|| \tag{3}$$

where the optimal value for the parameter ρ is searched iteratively.

2.2.2 Modifying LVQ for the Neyman-Pearson Criterion

The LVQ is usually used to minimize the total number of misclassifications, but it can also be used with the Neyman-Pearson criterion. Here we present only the algorithm; for its justification, see [4]. Basically, the modification consists of changing the LVQ update rule to the form

$$\mathbf{m}_c := \begin{cases} \mathbf{m}_c + \alpha\beta(\mathbf{x}^i - \mathbf{m}_c) & \text{when } \mathbf{x}^i \text{ and } \mathbf{m}_c \text{ are from the same class,} \\ \mathbf{m}_c - \alpha(1-\beta)(\mathbf{x}^i - \mathbf{m}_c) & \text{otherwise} \end{cases} \tag{4}$$

where the correct value for the parameter β, $0 < \beta < 1$ is searched empirically. The initialization of the codebook vectors should also be suitably modified; e.g., one may use an initialization based on a weighted kNN classification.

3 Material

The material used in the present study represents a certain segment of Kera Ltd.'s customer companies. The segment consists of small and medium-sized industrial enterprises, from which the sample has been selected using the line of business, age, and size as the pruning criteria. It was also required that the history and state of the enterprise is known well enough: if there was no data available for a longer period than two years before the bakruptcy, or if the last known financial statements were very poor, the company was rejected from the sample. However, in excess to these criteria, no data was rejected because it was "atypical", or looked like an outlier.

The total number of financial statements used is 4 898; these have been given by 1 137 companies, of which 304 have gone bankrupt.

4 Results

The maps used for the qualitative analysis are presented in figures 1 and 2.

The classification results using the Neyman-Pearson criterion are shown in table 1, and the results minimizing the total number of misclassifications in table 2. In the latter case, the SOM-based classifiers were not used, because the results with LVQ and LDA classifiers show that minimizing the total number of misclassifications yields an unacceptably high error I rate.

The performance of the SOM-1 and SOM-2 classifiers is depicted in figure 3.

5 Discussion

In the qualitative analysis, the SOM was a very valuable tool, with which several different regions can be found from the maps. In classification, there was no big difference between most classifiers. The brute-force-approach SOM-1 was clearly outperformed by other classifiers, as might have been expected. However, the SOM-2 worked quite well: the smoothing by the Gaussian neighborhood function seems to make it surprisingly insensitive to the number of neurons. SOM with RBF performed better than any other classifier considered here, although its generalization properties appear to be highly sensitive to the value of the kernel width parameter ρ.

Acknowledgements

This project was initiated when the first author was with the Neural Networks Research Centre of the Helsinki University of Technology. Both authors wish to express their gratitude to professor Teuvo Kohonen, head of the Research Centre, for his guidance and support during the project.

References

[1] B. Back, G. Oosterom, K. Sere, and M. van Wezel. A comparative study of neural networks in bankruptcy prediction. In *Multiple Paradigms for Artificial Intelligence (SteP94)*. Finnish Artificial Intelligence Society, 1994.

[2] B. M. del Brío and C. Serrano-Cinca. Self-organizing neural networks for the analysis and representation of data: Some financial cases. *Neural Computing & Applications*, 1:193–206, 1993.

[3] S. Kaski and T. Kohonen. Structures of welfare and poverty in the world discovered by the self-organizing map. Technical Report A24, Helsinki University of Technology, Neural Networks Research Centre, Espoo, Finland, February 1995.

[4] K. Kiviluoto. Analyzing financial statements with the self-organizing map. Master's thesis, Helsinki University of Technology, Espoo, Finland, May 1996.

[5] C. Serrano-Cinca. Beyond Z-analysis: Self-organizing neural networks for financial diagnosis. Discussion Papers in Acconting and Management Science 94-92, University of Southampton, December 1994.

Table 1: Classification results using Neyman-Pearson criterion with different error I values (per cent), based on financial statements given 2 ... 0 years before bankruptcy

Classifier	error I target	total error	st.dev.	error I	st.dev.	error II	st.dev.
LDA	0,20	19,0	(1,6)	21,0	(6,2)	18,8	(2,3)
	0,25	15,7	(1,0)	25,7	(5,4)	14,6	(1,5)
	0,30	14,1	(1,0)	29,5	(5,1)	12,5	(1,4)
LVQ	0,20	20,5	(2,0)	21,9	(4,1)	20,3	(2,4)
	0,25	15,9	(0,8)	25,7	(5,4)	14,9	(1,6)
	0,30	14,3	(1,0)	30,3	(4,5)	12,5	(1,5)
RBF-SOM	0,20	18,3	(1,2)	20,7	(5,6)	18,1	(1,7)
	0,25	15,8	(0,8)	26,4	(6,1)	14,7	(1,3)
	0,30	13,5	(1,0)	30,5	(6,4)	11,7	(1,6)
SOM-2	0,20	20,1	(1,9)	19,9	(6,3)	20,1	(2,6)
	0,25	16,6	(1,3)	25,4	(6,7)	15,7	(2,0)
	0,30	14,8	(0,4)	30,4	(6,8)	13,2	(0,9)
SOM-1	0,25	18,9	(2,8)	24,7	(6,3)	18,4	(3,4)
	0,30	18,2	(1,6)	30,8	(9,2)	16,9	(2,4)

Table 2: Classification results when minimizing the total number of misclassifications (per cent), based on financial statements given 2 ... 0 years before bankruptcy

Classifier	total error	error I	error II
LVQ	8,6	65,2	2,7
LDA	10,5	47,1	6,6

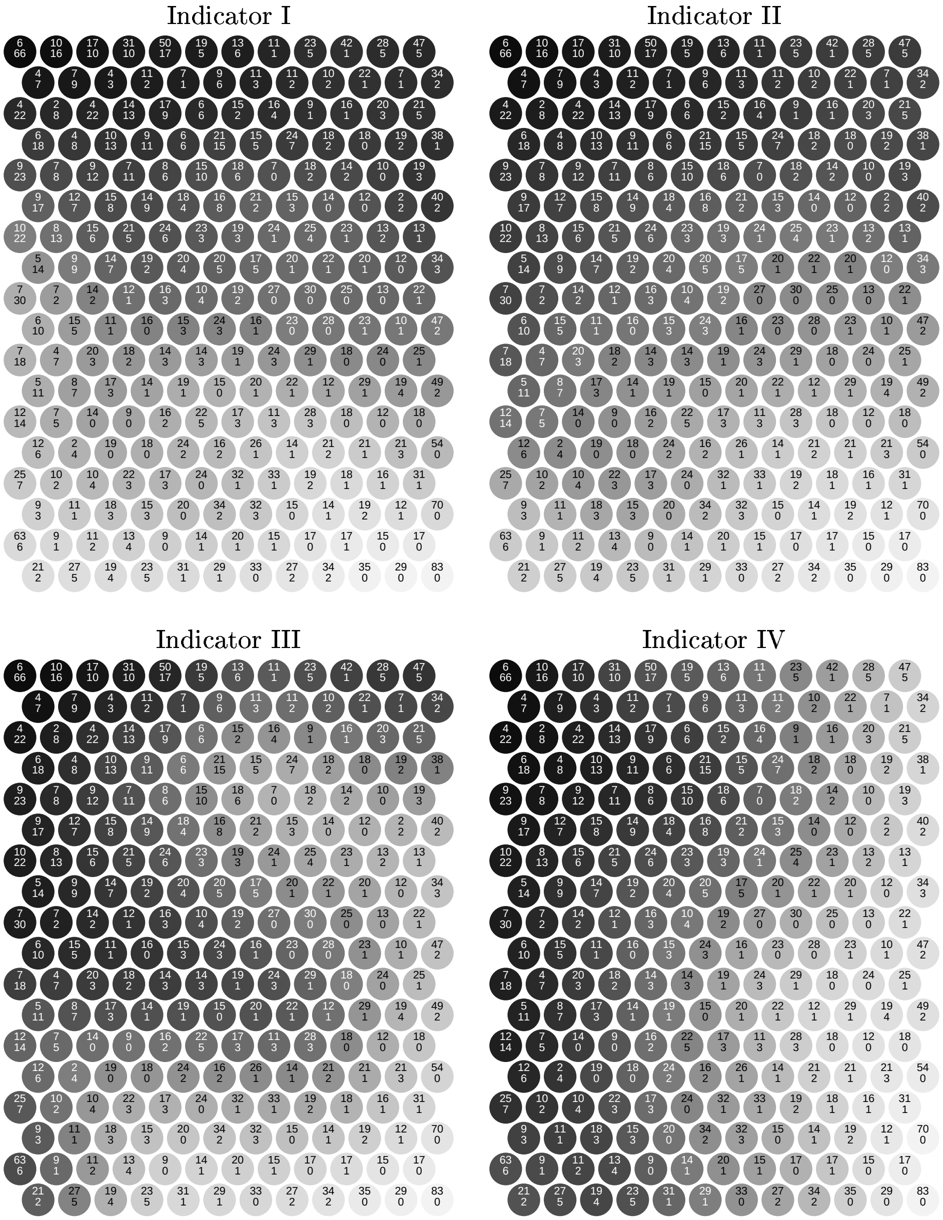

Figure 1: The relative values of the financial indicators – the lighter the color, the better the relative value. The number of the healthy (upper figure) and bankruptcy (lower figure) companies that have been mapped to each map unit is also shown; here a company has been considered as a healthy one, if it has not gone bankrupt within five years.

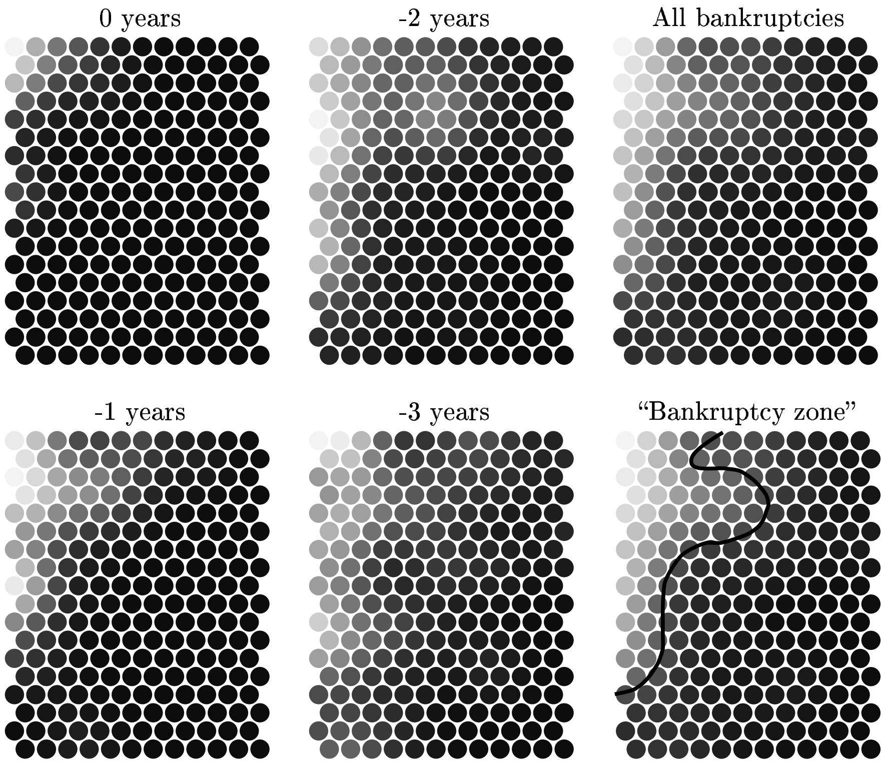

Figure 2: Bankruptcies, depicted 3 . . . 0 years before the bankruptcy; the corresponding financial indicators are shown in figure 1. Light color corresponds to higher proportions of bankruptcy companies. In the lower right corner, a “bankruptcy zone” is drawn: more than one third of the companies that are projected on the left side of the line have gone bankrupt within five years.

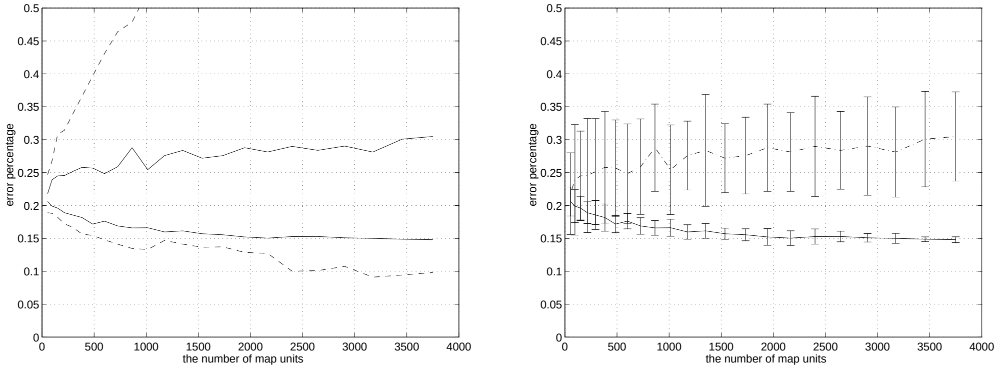

Figure 3: On the left: the classification accuracies of the SOM-1 (dashed line) and SOM-2 (solid line) classifiers vs. the number of map units – the lower lines represent the percentage of all misclassifications, the upper lines type I misclassifications (classifying a bankrupt company erroneously as a healthy company). On the right: SOM-2 classifier performance standard deviations, using 5-fold cross-validation.