CLEANING DATA WITH SELF-ORGANIZING MAP TO IMPROVE BANKRUPTCY RISK ASSESMENT

PENTTI BERGIUS
Finnvera Ltd.
P.O. Box 1010
FIN-00101 Helsinki, Finland
Pentti.Bergius@finnvera.fi

KIMMO KIVILUOTO
Helsinki Univ. of Technology
P.O. Box 5400
FIN-02015 HUT, Finland
Kimmo.Kiviluoto@hut.fi

JYRKI MAARANEN
Finnvera Ltd.
P.O. Box 1010
FIN-00101 Helsinki, Finland
Jyrki.Maaranen@finnvera.fi

ABSTRACT

We first outline a model to assess the bankruptcy risk of an enterprise, based on the Self-Organizing Map (SOM). Then, we use our model to find those data vectors that are atypical or do not fit the model well. These atypical observations are removed from the data, so that a "cleaner" picture of the predictable bankruptcies is obtained. The removed observations are further analyzed, to identify whether they share some common properties that could be used to filter out similar observations before analyzing future data with the model.

1. INTRODUCTION

The risk of bankruptcy is probably the most important single factor that a financing institution must consider, when some company is applying for a loan with inadequate or no collaterals. The estimated bankruptcy risk affects the pricing of the loan, and when the risk grows too large, the financing institution usually decides not to grant the loan at all. However, estimating the risk is not a trivial task. Although several quantitative techniques based on analysis of financial statements have been proposed – Altman's Z-analysis [1] being a prominent example – it may be argued that they are sensitive to the company size, industry etc. and thus not universally applicable.

In this paper, we focus on the following question: when is our model applicable and when not, and would it be possible to use these findings to better classify new data? We start by outlining a method to first estimate the probability of bankruptcy in section 2; then, in section 3 we propose a simple scheme to filter out those observations for which our model is not applicable. This is followed by analysis of the removed data in section 4 and discussion in section 5.

2. SOM FOR BANKRUPTCY PREDICTION

The method is we use to to estimate the bankruptcy probabilities is based on a popular neural network model called the Self-Organizing Map (SOM). The SOM has been widely used for data visualization, but can also be applied for classification, regression etc. [5]. For the latter purposes, it resembles vector coding algorithms such as LBG, which are based on partitioning the input space into several cells and then building a separate model for each of these cells. In quantitative models, the advantage of SOM over most vector coding algorithms is that it is a topographic mapping, and thus the metric in the input space has a counterpart also on the output space, i.e. on the lattice consisting of the units of the SOM. This, in turn, can be utilized in smoothing the model, in analyzing trajectories (system states at several successive time instants), and in building hierarchical models.

We and a number of other authors have earlier applied the SOM on the bankruptcy prediction problem; see e.g. [7, 6, 3, 4]. Here, our first step follows a similar strategy as in these earlier works: we start by training a SOM with the data derived from the corporate financial statements. Then, we map the data onto the SOM, so that each data vector representing the financial statement of a company at a given year gets mapped to one unit of the map. Now, each map unit represents a part of the whole data set, a part that consists of financial statements that are similar to each other.

The next step is to estimate the probability of bankruptcy for each map unit, in other words, the conditional bankruptcy probability of a company given that its financial statement gets mapped to that particular unit, $\hat{P}(\text{bankruptcy} \mid \text{unit})$. The estimate $\hat{P}$(bankruptcy | unit) is formed simply by calculating the relative frequency of bankruptcies among all the data mapped to the unit. Here we labeled a financial statement as given by a bankrupt company, if the company did not survive for longer than three years after publishing that particular financial statement.

Note that the method we use here to obtain the bankruptcy probability estimates is a simplification from our earlier works, in which we have smoothed the relative frequencies [3, 4].

An example of a bankruptcy probability map is depicted in panel A. of Figure 1. Here we use a three-dimensional SOM, which sometimes captures the true structure of the data better than the more commonly used two-dimensional SOM [2].

3. CLEANING THE DATA

At this point, we already have a method for predicting bankruptcies: given a financial statement, we map it onto the SOM and take the bankruptcy probability as $\hat{P}$(bankruptcy | unit the financial statement is mapped to) we estimated above. However, these estimates are blurred by two kinds noise:

- Some of the companies that have gone bankrupt may have done so "without their own fault". For instance, a company may be a very well managed subcontractor of a larger firm, and if this larger firm moves its operations to another country or goes bankrupt, a small subcontractor may not be able to find new customers or change its product line quickly enough. Another example is a company that is very dependent on its owner/manager – for such a company, the health problems of the owner/manager may be fatal. Although a skilled analyst takes this kind of factors into account, there is no way to infer them from the financial statements, and therefore it is not reasonable to try to teach the corresponding cases to the neural network.

- Some of the companies we considered as healthy may actually be going bankrupt in the near future, but we do not have knowledge of this – many companies in our data are known for only a couple of years, and the average is just sligthly over four years.

If we re-build the model with the data where the noise is removed, it seems probable that the bankruptcy prediction accuracy might improve.

To get rid of the first kind of noise, we first find those "non-bankruptcy units" where the estimate $\hat{P}$(bankruptcy | unit) is smaller than some threshold T_1. Then, we look at the two last financial statements of companies that have failed, and only if both of these are mapped to the non-bankruptcy units, all the financial statements from that company are removed from the data. (The reason for looking at two last financial statements, not just the last one, is that it is rather common for a failing company to try to look good at any cost. It may even be able to give one good-looking statement after a bad one, but cannot do this more than once; therefore, the company does not publish any more annual statements and eventually enters the bankruptcy proceedings.)

To get rid of the second kind of noise, the above procedure is reversed. We set a threshold T_2, and select those SOM units where $\hat{P}$(bankruptcy | unit) > T_2. Then, we look at the last known statements of the (supposedly) healthy companies, and when these get mapped to the selected SOM units, all the financial statements from the corresponding companies are deleted from the sample.

Now we have much less noise in our sample, and we can re-evaluate the estimates $\hat{P}$(bankruptcy | unit), improving the accuracy of the map. Still better results are achieved, if we re-train the SOM from the start, and then re-evaluate the estimates $\hat{P}$(bankruptcy | unit) on the new map.

In panel B. of Figure 1 is displayed the same map as in panel A. but the bankruptcy probabilities are evaluated using the cleaned data. In panel C. the map is also re-trained with the clean data, and finally in panel D. the re-trained map is evaluated using the original data.

4. ANALYZING THE RESULTS

The data we have used in the present study consists of the financial statements of the customer companies – very small and startup companies excluded – of Finnvera Ltd, a Finnish risk financing company. In the original sample, there were 30 593 financial statements from 7 028 companies, of which 1 244 eventually failed; for this data, the classification error was 28.7%. In the cleaned data, with cleaning thresholds $T_1 = 0.1$ and $T_2 = 0.3$, there remained 27 274 financial statements, and the classification error decreased to 23.5% with the original map but re-evaluating the probability estimates, and further to 21.2% with re-training the map.

A. The original map

B. The map A. evaluated using the cleaned data

C. The map A. retrained using the cleaned data

D. The map D. evaluated using the original data

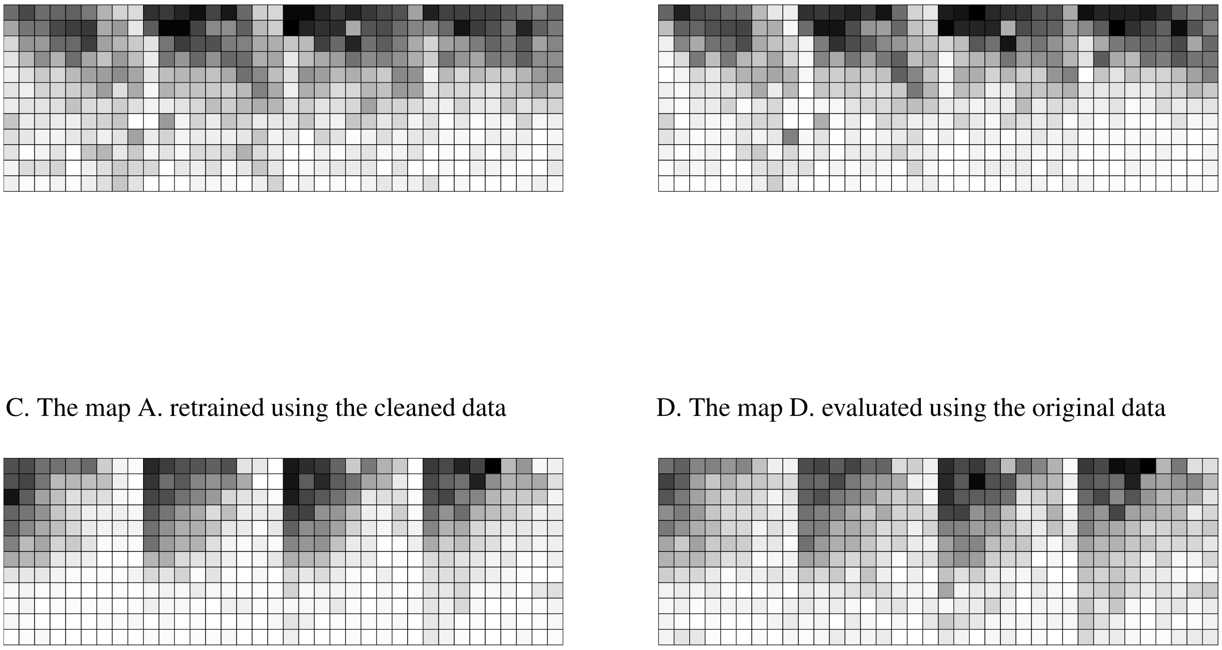

Figure 1: The four bankruptcy probability maps. On each map, the dimensions of the SOM grid are $12 \times 9 \times 4$, but for better visibility the four layers of the map are presented here side by side. On the dark areas of the map the bankruptcy risk is high.

Next, we analyzed those companies that went bankrupt on the "non-bankruptcy region" of the map and were thus removed from the data set. A sample of 32 bankrupt cases was formed, on the basis of geographical location; of these, we were able to examine 29 cases. The main reasons for each bankruptcy were distributed in the following way:

(a) 10 cases: unexpected circumstances due to the economic depression of the early 90's

(b) 4 cases: health problems of the owner/manager

(c) 3 cases: sudden crisis within one industry sector in the late 80's

(d) 2 cases: inadequate behavior of the management

(e) 1 case: unsuccesful change of the management

(f) 1 case: failure in a major product development project

(g) 1 case: unsuccesful attempt to restart an enterprise after bankruptcy

(h) 7 cases: no clear reason

Thus far, we haven't been able to carry out a corresponding analysis on the supposedly healthy companies that were in the bankruptcy region, and were thus removed from the data. However, an initial examination suggests that these cases are concentrated around the overheated economy of the late 80's.

5. DISCUSSION

The results show that cleaning the data gives a clearer picture of the bankruptcies. To some extent, it also makes it easier to recognize susceptible enterprises, although many atypical bankruptcies also seem to be unpredictable – for instance, in practice it would be difficult to monitor the health of the manager of a particular company, or to forecast a sudden crisis within some industry sector. Still, removing these cases from the data helps to improve the recognition of the predictable bankruptcies.

6. REFERENCES

[1] E. I. Altman. *A complete guide to predicting, avoiding, and dealing with bankruptcy*. John Wiley & Sons, Inc., 1983.

[2] K. Kiviluoto. Comparing 2D and 3D self-organizing maps in financial data visualization. In T. Yamakawa and G. Matsumoto, editors, *Methodologies for the Conception, Design and Application of Soft Computing – Proceedings of the 5th International Conference on Soft Computing and Information/Intelligent Systems (IIZUKA'98). Iizuka, Fukuoka, Japan.*, volume 1, pages 68–71, Singapore, Oct. 1998. World Scientific.

[3] K. Kiviluoto. Predicting bankruptcies with the self-organizing map. *Neurocomputing*, 21:191–201, 1998.

[4] K. Kiviluoto and P. Bergius. Two-level self-organizing maps for analysis of financial statements. In *Proceedings of the 1998 IEEE International Joint Conference on Neural Networks (IJCNN'98)*, volume 1, pages 189–192, Piscataway, New Jersey, USA, May 1998. IEEE Neural Networks Council.

[5] T. Kohonen. *Self-Organizing Maps*. Springer Series in Information Sciences 30. Springer, Berlin Heidelberg New York, 1995.

[6] B. Martín-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.

[7] S. A. Shumsky and A. Yarovoy. Neural network analysis of Russian banks. In *Proceedings of the workshop on self-organizing maps (WSOM'97)*, Espoo, Finland, June 1997. Neural Networks Research Centre, Helsinki University of Technology.