Discussion Paper, Series A, No.2021-361

ENSEMBLE NEURAL NETWORK USING A SMALL DATASET FOR THE PREDICTION OF BANKRUPTCY: COMBINING NUMERICAL AND TEXTUAL DATA

RASOLOMANANA Onjaniaina Mianin'Harizo

October.2021

Faculty of Economics and Business
Hokkaido University
Kita 9 Nishi 7, Kita-Ku, Sapporo 060-0809, JAPAN

ENSEMBLE NEURAL NETWORK USING A SMALL DATASET FOR THE PREDICTION OF BANKRUPTCY: COMBINING NUMERICAL AND TEXTUAL DATA

Rasolomanana Onjaniaina Mianin'Harizo

ABSTRACT

This paper presents an ensemble neural network using a small data set in the context of bankruptcy prediction. The individual models of the ensemble use different data of different types. We compare the performance of three neural network models: one using a single type of data, one using a combination of both data in a single data frame, and one using ensemble learning. The results show that the ensemble model outperformed the individual model and the combined model. This suggests that with scarce training data, especially when using different types of data, ensemble neural network can improve the level of prediction accuracy.

Keywords: ensemble neural network, small dataset, combined data, bankruptcy prediction

1. INTRODUCTION

This research focuses on the development of a bankruptcy prediction model using numerical and textual data simultaneously. Being able to predict bankruptcy accurately will allow managers, investors, and other stakeholders to spot the risk in advance, based not only on quantitative information, but also on qualitative information, and to take the necessary measures.

In bankruptcy prediction, although previous studies were primarily focused on quantitative data, namely financial ratios, several studies have also investigated the addition of qualitative data to their model (Zopounidis et al., 1992; Slowinski & Zopounidis, 1995; Park & Han, 2002; Zakharova, 2013; Boratyńska & Grzegorzewska, 2018). They found out that quantitative and qualitative data may be based on different assumptions but are complementary. Therefore, using both in the same model yields better prediction results than using only one of the two types.

However, these models were trained using a huge volume data in the development process. In practice, though, especially when it comes bankruptcy, the data may be limited

naturally (for instance, the number of companies which has gone bankrupt cannot be controlled) or expensive in terms of time and/or money to get. Therefore, there is a need to develop a bankruptcy prediction model with mixed data (quantitative and qualitative) using a small dataset. In this study, we are using data from Japanese listed companies, out of which only a few (about 200 companies) have filed for bankruptcy up until 2019, which is relatively small as a dataset. Here, a small dataset means that the number of examples used in the training is far smaller than the number of predictors.

In this study, we develop a neural network model using different types of data: numerical and non-numerical data. Before feeding the data to the model, non-numerical features, such as texts, are transformed beforehand. However, when the data are different in nature (categorical or continuous, sequential, or non-sequential, etc.), a plain concatenation may cause discrepancies, and hence, hurt the neural network model. Although, it has been proven possible to bypass this problem by using an end-to-end learning approach (creating a model with multiple input layers), it would still require a large amount of data.

Some studies have investigated the use of neural networks with small datasets, especially in the field of medicine to identify diseases.

Pasini (2015) has successfully developed a neural network tool for small datasets analysis using the leave-one-out procedure. The leave-one-out procedure is a cross-validation method commonly used in machine learning. Since this method sets the elements in the validation dataset randomly, to improve the robustness of the model, the author also used an ensemble neural network.

D'souza et al. (2020) found out that optimization task is more important when the data is scarce. The authors developed a convolutional neural network using a two-step method to find the optimal neural network structure. The first step consists of finding and listing all potential network structures, and the second step consists of finding the optimal layer dimension by picking the best performing models and permuting combinations of layer dimension.

However, both studies require a huge computing power. In fact, the leave-out-procedure can be seen as an extreme version of the classic k-fold cross-validation, where the value of k

corresponds to the number of examples in the dataset. Even with a small dataset, the computation behind the cross-validation would take a long time. Similarly, the method proposed by D'souza et al. (2020) may end up with a large list of potential neural network structures and training each model may take a long time without additional optimization or with a limited computing power.

Therefore, in the present study, we introduce another method using an ensemble neural network with 2 models: one with numerical data, and another with textual data. The purpose is then to investigate how well the ensemble neural network model can perform with a small dataset to predict bankruptcy, when using numerical and textual data. Thus, this paper attempts (1) to compare the performance of a neural network using combined data (numerical and textual data into a single data frame) with an ensemble of two individual networks with different data types, (2) to determine what combination of numerical and textual data gives the best bankruptcy prediction results.

2. METHODOLOGY

This section will outline the setup and steps towards the construction of the ensemble neural network. An ensemble model combines several "weak" models with relatively low performance into a single stronger model with a high predictive performance. In the model presented in this manuscript, we combined two independent neural networks, one using numerical data and one using textual data.

2.1. Input data

In this study, we used both numerical and textual data. The numerical data is composed by a set of 22 financial ratios, which are the mainly used variables in literature and in practice, when it comes to analyzing the financial situation of companies or predicting their failure (Altman, 1968; Ohlson, 1980; Shimerda & Kung, 2012; Liang et al., 2016).

The values of the financial ratios were normalized between 0 and 1, because despite being expressed in percentages, the scales are very different (maximum value= 12029.10%, minimum value= -4889.69%). Normalization ensures that the model will not attribute

higher weights to high value input, as it may hurt the learning of the model (Bhanja & Das, 2018).

As for the textual data, we extracted segments of text from Japanese listed companies' securities reports and "*kessan tanshin*" reports[1], that directly or inherently refer to the business profitability and/or competitive advantage of the company, which are the two most common and important qualitative factors of bankruptcy according to literature (Boratyńska & Grzegorzewska, 2018; Kim & Han, 2003; Park & Han, 2002; Slowinski & Zopounidis, 1995; Zakharova, 2013). The sections that refer to the business profitability and/or competitive advantage of the company are the following:

- Risks factors (from securities report)
- Overview of the performance (from securities report)
- Issues to be addressed (from securities report)
- Performance prediction (from "*kessan tanshin*" report)

Each segment of text was preprocessed so all stop words or meaningless characters are removed, then transformed into numerical values using Term Frequency – Inverse Document Frequency (TF-IDF)[2] scores.

The total sample of 137 companies used in this study comprises 68 cases for the bankrupted class and 69 cases for the non-bankrupted class. The financial ratios and the reports of each company were downloaded from the EOL database[3], ranging from Fiscal Year 2004 to Fiscal Year 2018, and no industry were excluded.

2.2. Neural network development

We developed 3 types of neural networks models: individual model (using a single type of data), combined model (using numerical and text features combined into a single data frame), and ensemble model (2 individual networks using different types of data).

[1] "*Kessan tanshin*" reports are exclusive to Japanese companies, which is why we kept the original appellation. They report financial results of Japanese listed companies, that are disclosed to the market and investors in a timely manner.
[2] TF-IDF is a technique to transform words in a document into numerical values. Roughly speaking, if a rare word appears many times, it increases the importance of that word when classifying the document.
[3] EOL is a database that provides financial and non-financial information on Japanese listed companies.

After the data has been preprocessed, it is split into 3 sets: training set, validation set and test set. As a rule of thumb, when a huge volume data is available, the data is split randomly according to a certain distribution. But in the case of a small dataset, the same method of splitting would create a great imbalance amongst the 3 sets, as our dataset is comprised of companies from different subclasses (different industries and data periods). Therefore, it is necessary to preserve a certain level of control when splitting the data by preserving a minimum of balance amongst the sets. Ultimately, the elements in each set were selected randomly, however, we made sure that each set had elements from each subclass.

A total of 15 different models were built, each model being a combination of financial ratios and the segments of texts (see Table 1), to find out which texts are the most descriptive of a nearly coming bankruptcy.

Table 1. Input information in each model

Model	Numerical data	Textual data			
	Financial ratios	Risks factors	Overview of the performance	Issues to be addressed	Performance prediction
A	○	○			
B	○	○	○		
C	○	○		○	
D	○	○			○
E	○	○	○	○	
F	○	○	○		○
G	○	○		○	○
H	○	○	○	○	○
I	○		○		
J	○		○	○	
K	○		○		○
L	○		○	○	○
M	○			○	
N	○			○	○
O	○				○

Each neural network model is then trained and cross-validated. We closely monitored the training loss and the validation loss to ensure that the model was learning appropriately. If the loss does not converge, the settings of the network are iteratively changed until the optimal settings are found (until the loss is minimized). Moreover, to avoid overfitting, we applied some regularization techniques to the models.

For the individual models, the training set is directly fed to the network. For the combined model, after each data has been preprocessed, they are first concatenated into a single data frame before the training. For the ensemble model, instead of concatenating the inputs into a single data frame, two individual neural networks were built. The output from each network is then weighted proportionally to its performance to calculate the output of the ensemble model (See Figure 1).

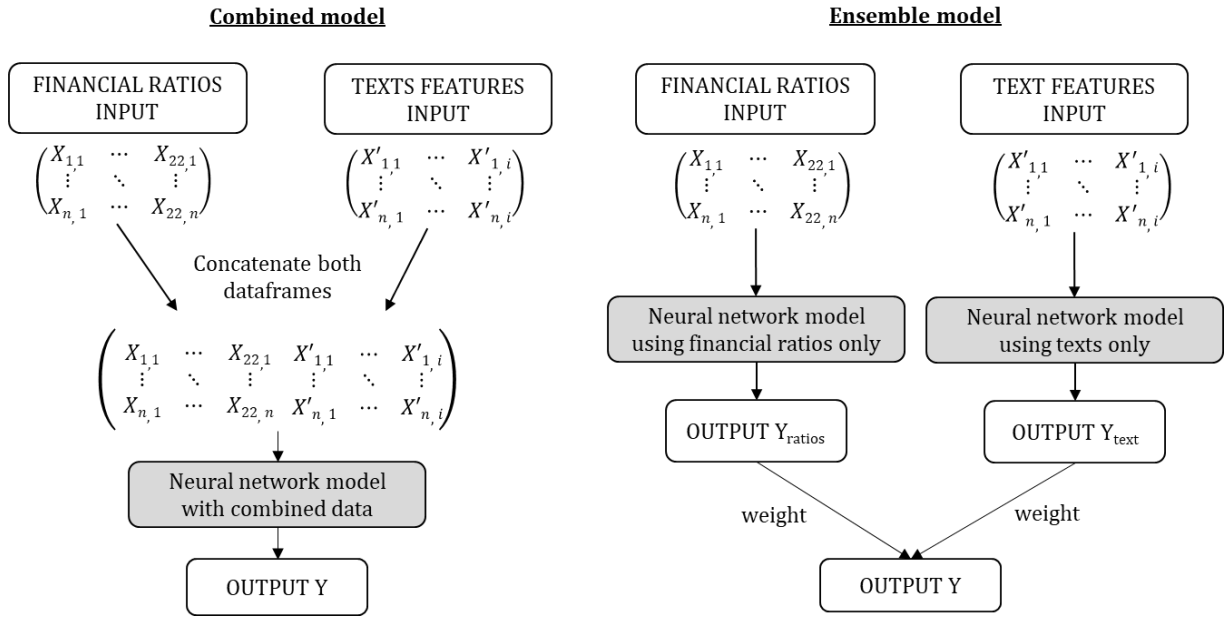

Figure 1. Schematic representation of the Combined model and Ensemble model used in this paper

The weighted average of the individual inputs is calculated as follows:

$$weighted\ average = \frac{\sum_{i=1}^{n}(x_i \times w_i)}{\sum_{i=1}^{n} w_i}$$

with
x = financial ratio or text predictions
w = weight factor (between 0 and 1)

3. RESULTS

After training, the performance of the models is evaluated using the test set. The accuracy rates of the individual model, combined model and ensemble model are summarized in Figure 2.

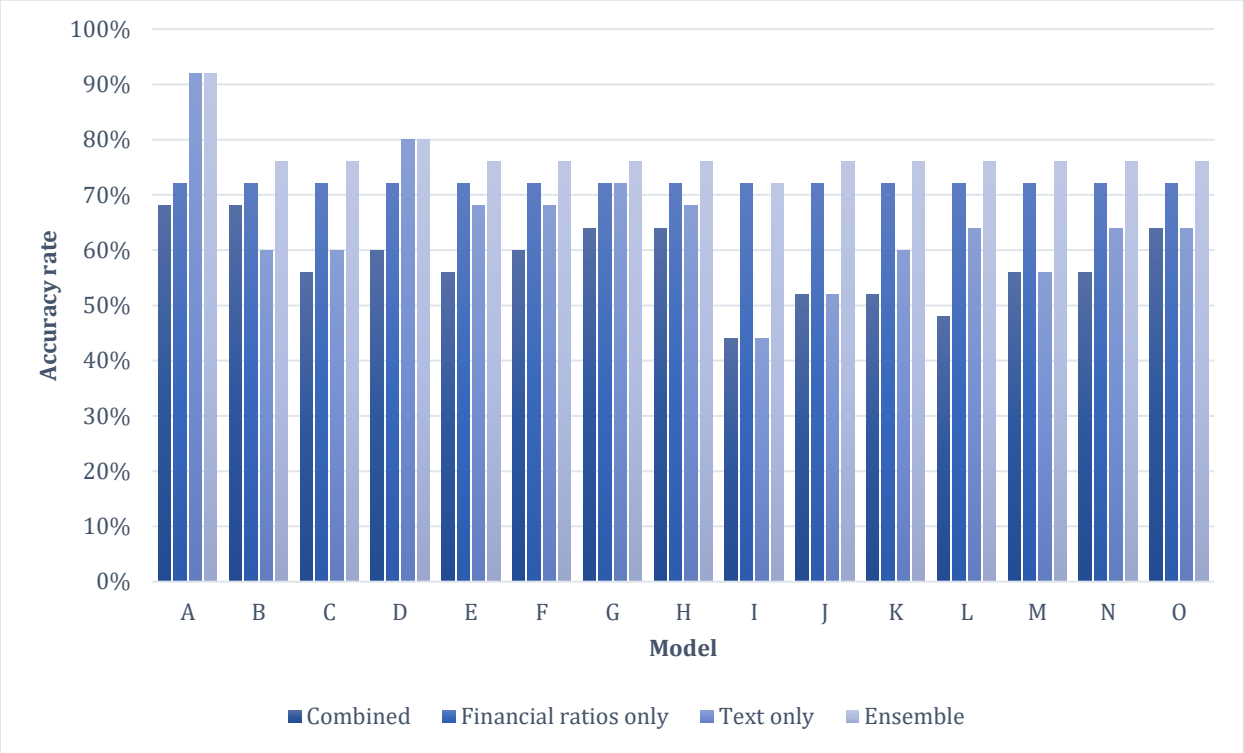

Figure 2. Graph showing the performance of the different models: combined model, individual models and ensemble model

The ensemble model showed the highest accuracy for all the models (model A to model O). The model A (financial ratios and risk factors) showed the best performance with an accuracy rate of 92%.

We compared the performance of the ensemble with other machine learning models The results are summarized in Table 2.

Table 2. Performance comparison with other machine learning models. NN_{ens}: Ensemble Neural Network, RDC: Random Forest, TREE: Decision Tree, SVM: Support Vector Machine, KNN: K-Nearest Neighbor, LOG: Logistic Regression, NB: Naive Bayes

Model	NN_{ens}	RDC	TREE	SVM	KNN	LOG	NB
A	92%	88%	64%	64%	80%	64%	56%
B	76%	84%	72%	64%	80%	64%	52%
C	76%	80%	64%	64%	80%	64%	56%
D	80%	84%	64%	64%	84%	64%	56%
E	76%	80%	72%	64%	80%	64%	52%
F	76%	84%	72%	64%	80%	64%	56%
G	76%	84%	76%	64%	76%	64%	52%
H	76%	84%	72%	64%	80%	64%	52%
I	72%	76%	72%	60%	80%	64%	56%
J	76%	80%	68%	60%	80%	64%	52%
K	76%	80%	68%	60%	80%	64%	56%
L	76%	84%	72%	68%	80%	64%	52%
M	76%	80%	64%	64%	80%	64%	60%
N	76%	84%	68%	64%	80%	64%	52%
O	76%	84%	64%	64%	80%	68%	64%

In overall, the Random Forest demonstrated the best performance. And model A showed the highest accuracy rate (88%) here too. K-Nearest Neighbor model also showed about 80% of accuracy amongst all models, however, the results are too stable regardless of the input information, making it less reliable than the Random Forest. The structure of the data indeed has an important impact on the results when using K-Nearest Neighbor, meaning that it is sensible to the dimensionality of the data. Therefore, as the dimension increases, more training examples is required. The remaining machine learning models are even more sensible to this matter, as they produced poorer results (52 to 72% accuracy).

4. DISCUSSION

Combining quantitative data and qualitative data did not give better prediction results than using only one type of data. However, when creating individual networks for each type of data, then averaging their respective results in an ensemble network, the performance was better than when using only one type of data. These confirm the fact that plain combination of different data creates discrepancies within the data, making the learning

difficult for the model. The results varied depending on the input information, and the model with financial ratios and risk factors (model A) showed the best result. Thus, we can deduce that amongst all the texts, the information contained in the section about risk factors is the most valuable when it comes to bankruptcy prediction.

The results also show that the performance of some other machine learning model was better than the ensemble neural network. Amongst all the models, Random Forest model yielded the best accuracy rates. This suggests that, although our ensemble neural network can be used with small datasets, it is still too complex when fed with a small size of samples. What is interesting is that Random Forest is also an ensemble learning method which uses decision trees. This is consistent with Dietterich (2000) stating that when the amount of available training data is too small, ensemble learning can help find a good approximation and improve prediction accuracy by averaging the outputs of the individual models. In fact, those constitute the benefits of using ensemble models. Since the base models are trained separately and do not have to be individually highly performant, this method requires much less computation power and model tuning.

Another possible reason why Random Forest showed high levels of accuracy is the fact that we used only two base networks for the ensemble, while the Random Forest model used dozens of trees.

5. CONCLUSION

In a context where the available data is scarce, and predictive variables are far more numerous, the parameter tuning becomes too complex and the neural network ends up with a poor learning. In this paper, we demonstrate the viability of using an ensemble neural network with different data – numerical data in one network and textual data in another – in such context.

This study also confirms that quantitative and qualitative data are indeed complementary, and ensemble learning can bring out that complementarity, by giving a weighted average of the individual models, and yielding a higher level of prediction accuracy.

One limitation to this study is the need of a lot of computation power for Grid Searching all the parameters, so for efficiency reasons, it was done using an iterative process (trial-and-

error). Therefore, although the current settings of the model have showed acceptable results, it may not be the optimal ones.

As for further research, we intend to explore other ensemble learning approaches with neural network by using bagging, boosting, stacking methods, and integrating more than 2 networks.

6. ACKNOWLEDGEMENTS

Our sincere gratitude goes to our academic advisors for their insightful comments and guidance regarding the preparation of the present manuscript.

7. REFERENCES

Altman, E. I. (1968). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. *Source: The Journal of Finance*, *23*(4), 589–609.

Bhanja, S., & Das, A. (2018). *Impact of Data Normalization on Deep Neural Network for Time Series Forecasting*. 5–10. http://arxiv.org/abs/1812.05519

Boratyńska, K., & Grzegorzewska, E. (2018). Bankruptcy prediction in the agribusiness sector: Lessons from quantitative and qualitative approaches. *Journal of Business Research*, *89*(February), 175–181. https://doi.org/10.1016/j.jbusres.2018.01.028

D'souza, R. N., Huang, P. Y., & Yeh, F. C. (2020). Structural Analysis and Optimization of Convolutional Neural Networks with a Small Sample Size. *Scientific Reports*, *10*(1), 1–13. https://doi.org/10.1038/s41598-020-57866-2

Dietterich, T. G. (2000). Ensemble methods in machine learning. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *1857 LNCS*, 1–15. https://doi.org/10.1007/3-540-45014-9_1

Kim, M. J., & Han, I. (2003). The discovery of experts' decision rules from qualitative bankruptcy data using genetic algorithms. *Expert Systems with Applications*, *25*(4), 637–646. https://doi.org/10.1016/S0957-4174(03)00102-7

Liang, D., Lu, C. C., Tsai, C. F., & Shih, G. A. (2016). Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study. *European Journal of*

Operational Research, *252*(2), 561–572. https://doi.org/10.1016/j.ejor.2016.01.012

Ohlson, J. A. (1980). Financial Ratios and the Probabilistic Prediction of Bankruptcy. *Journal of Accounting Research*, *18*(1), 109. https://doi.org/10.2307/2490395

Park, C.-S., & Han, I. (2002). A case-based reasoning with the feature weights derived by analytic hierarchy process for bankruptcy prediction. *Expert Systems with Applications*, *23*(3), 255–264. https://doi.org/10.1016/S0957-4174(02)00045-3

Pasini, A. (2015). Artificial neural networks for small dataset analysis. *Journal of Thoracic Disease*, *7*(5), 953–960. https://doi.org/10.3978/j.issn.2072-1439.2015.04.61

Shimerda, K. H. C. and T. A., & Kung. (2012). An of Empirical Analysis Useful Financial Ratios. *Financial Management Association İnternational*, *10*(1), 51–60.

Slowinski, R., & Zopounidis, C. (1995). Application of the Rough Set Approach to Evaluation of Bankruptcy Risk. *Intelligent Systems in Accounting, Finance and Management*, *4*(1), 27–41. https://doi.org/10.1002/j.1099-1174.1995.tb00078.x

Zakharova, A. A. (2013). Fuzzy swot analysis for selection of bankruptcy risk factors. *Applied Mechanics and Materials*, *379*, 207–213. https://doi.org/10.4028/www.scientific.net/AMM.379.207

Zopounidis, C., Pouliezos, A., & Yannacopoulos, D. (1992). Designing a DSS for the assessment of company performance and viability. *Computer Science in Economics and Management*, *5*(1), 41–56. https://doi.org/10.1007/BF00435281