
Neural Network based Bankruptcy Prediction System

Anany Dwivedi¹ Zili Ma¹ Jiexuan Sun¹

Abstract

The bankruptcy prediction is always one of the hottest topics in economics. A lot of statistical analysis has been made in the solving this problem. However, as far as we know, a very less amount of research is available in solving this problem using Artificial Intelligence (eg. Neural Nets, SVMs, Decision Trees etc). In this paper, we propose a Neural Network based approach to solve this problem. We design our network using Net2Net technique. We achieve an AUC value above 0.920 for all the five different dataset. In the end we compare the Net2Net trained network with a conventionally trained network of the same structure.

1. Introduction

Financial stability of a company/firm or a bank is really important for its proper functioning. Foreseeing the financial conditions of a firm based on various econometric measures is one of the most important mission concerned by every industry participant like investors shareholders, lawmakers, central banks, auditors and managers (Bredart, 2014). So, prediction of bankruptcy of an enterprise is of great importance. This is a problem that affects the economy on a global scale (Zhang et al., 2013). Therefore, various methods have been developed including statistical hypothesis testing, statistical modeling and machine learning techniques.

However, the complexity and diversity of the real business world makes such predictions very challenging. For example, the financial indicators describing the business conditions sometimes can not exactly reflect the true operation status of the firm, and as one possible extreme result, it may cause a rare, hard-to-predict bankruptcy with series of serious damages as consequence, which is known as a black swan event. To effectively identify those companies with higher financial risk based on these inaccuracy indi-

cators is one problem which needs to overcome (Altman & Hotchkiss, 2010). Moreover, the most concerned issue here associated with bankruptcy prediction is that the data set is typically skewed, because there are more companies with normal operation than companies going bankrupt. This problem has gained increasingly attention from researchers because it will cause classification models to favour the majority class over the minority class. It is also considered as a challenge by the Data Mining Community (Yang & Wu, 2006).

1.1. Dataset

In this work, we use the Polish companies bankruptcy dataset. The dataset is hosted on UCI's Machine Learning Repository (Tomczak, 2016). It is about bankruptcy classification of Polish companies. The dataset is collected over 2007-2012. The whole dataset is divided into five separate subsets namely: Year1, Year2, Year3, Year4, Year5. Each year forecast the company financial status for the year 2013.

1.2. Research contributions

In this paper we prove that the Neural Networks are capable of predicting if a company can go bankrupt with high accuracy. Our Neural Network outperforms the neural network designed by (Zikeba et al., 2016) and even their best network in a few cases. We also show that the knowledge gained from training a model for year1 dataset could be utilized while training networks on subsequent datasets for faster learning.

2. Related Work

Bankruptcy prediction is an important and widely studied topic in the business intelligence field. In brief, there are two approaches conceptually existing for bankruptcy prediction: structural and empirical. The former one is mainly based on the financial statement data and explain how these financial ratios worsened as long as firms faced bankruptcy. Compared to the former empirical approaches, neural networks do not require to make assumptions about the distribution of the data and allows nonlinear relations beside linear model. This allowance is especially important for

¹Worcester Polytechnic Institute. Correspondence to: adwivedi,zma2,jsun2 <@wpi.edu>.

bankruptcy predictions because the relation between the likelihood of default and the explanatory variables do not have to be linear (Iturriaga & Sanz, 2015).

From the late 1980s, artificial intelligence technique, particularly rule-based expert system, base-based reasoning systems has been applied to bankruptcy prediction. Since the methods of machine learning have been evolving quickly in recent years, many attempts have been made based on different techniques like evolutionary programming (Zhang et al., 2013), support vector machines (SVM) (Shin et al., 2005), neural network ((Kittler et al., 1998), (Angelini et al., 2008)), and ensemble classifier, which can take advantage of different learning paradigm (Chen & He, 2015). The problem of SVM comes from that only one specific kernel function determines the model and the tuning of this kernel function has to be completed carefully and manually. To alleviate the disadvantage, the neural network model with the capability of automatic feature extraction was proposed in recent years, some authors have shown the superiority of neural networks relative to other techniques (Lee & Choi, 2013), but the comparison between traditional models and neural networks remains an open question within the literature and has led to mixed results (Bernhardsen, 2001). Moreover, these studies, as far as we know, were almost reported before 2014, since many new techniques to improve the performance of neural network have been developed in recent years. The aim of this paper is to build a automatic prediction model based on the novel techniques reported recently and to improve the final performance in predicting bankruptcy.

3. Proposed Method

Most classification algorithms will only perform optimally when the number of samples of each class is roughly the same. Highly skewed datasets, where the minority is heavily outnumbered by one or more classes, have proven to be a challenge while at the same time becoming increasingly common. One way of addressing this issue is by re-sampling the dataset as to offset this imbalance with the hope of arriving at a more robust and fair decision boundary than one would otherwise.

The Polish companies dataset is highly skewed. The majority (safe) class is over 97% in all the 5 dataset. To overcome the skew class problem we used Synthetic Minority Over-sampling Technique (SMOTE). We implement this using imbalanced-learn package of scikit-learn contrib (Lemaître et al., 2017).

For sampling using SMOTE, we first split our original data in 80% – 20% split. We use the 80% for data generation and keep the 20% for the validation of our network model. Figure 1-A. shows the first three principal components of

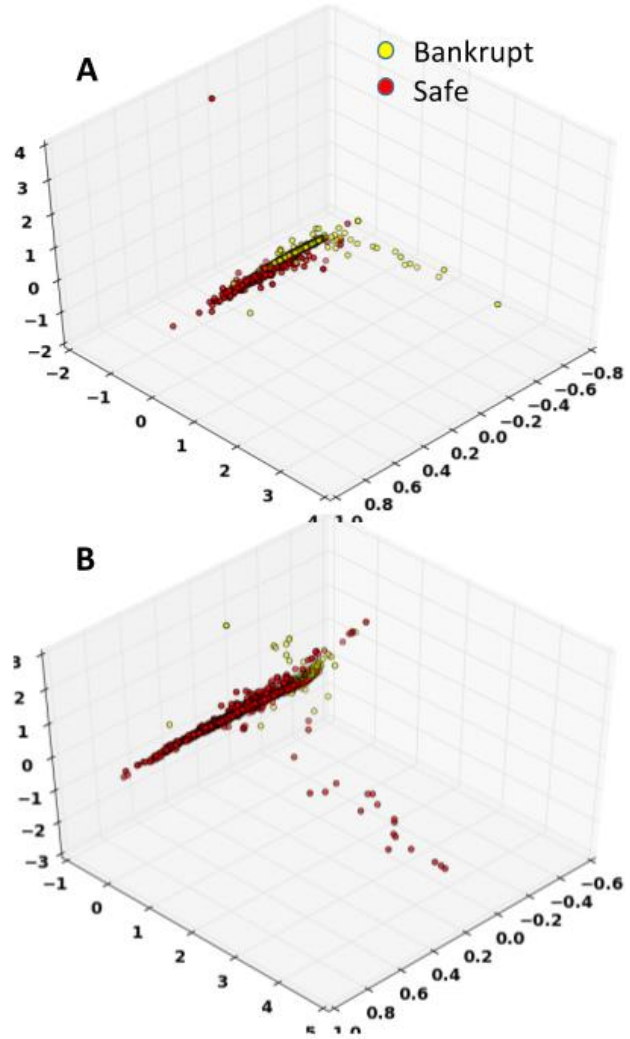


Figure 1. First three principal components of Original and Generated data. Figure A shows the First three principal components of the original data. Figure B shows the First three principal components of the generated data.

the original data and Figure 1-B. shows the first three principal components of the generated data.

To model the network, we try two different approach and then compare the results obtained from the two techniques. The first technique used for learning to predict if the company is safe or bankrupt was based on the Net2Net initialization technique (Chen et al., 2015), which performs learning in a sequential manner starting with small MLPs and then scaling the neural network up to a larger size in width and depth by using the previous smaller network as a teacher to the new larger student network. This approach avoids the usage of a very large initial network and then re-learning the entire network from scratch if the performance is not suitable. The scaling of the network is performed

Table 1. AUC score on validation data during widening operation [Year1]

Number of Neurons in First Hidden Layer	2	4	8	16	32	64	128
AUC Score	0.8318	0.8529	0.8843	0.8990	0.9212	0.9441	0.9452

Table 2. AUC score on validation data during deepening operation [Year1]

Number of Hidden Layers	1	2	3	4	5	6	7	8
AUC Score	0.9452	0.9503	0.9546	0.9603	0.9628	0.9659	0.9668	0.9670

by initializing the student network with the weights of the teacher network and then widening or deepening it. Widening involves adding additional neurons to a layer. Deepening involves adding a new hidden layer to the network. The advantage of the method is that it ensures that the student network improves upon the teacher network. The second approach is to train a FCN using the conventional method using the best architecture we obtain for the Net2Net approach. We will then compare the results of the two approach based on the performance metric discussed in section 4.

4. Metric of Performance

As we discussed in section 3 all of the five dataset are skewed with almost 97% of the majority class and rest being the other. So, in such a case, making Accuracy (that is checking % correct) as the Metric of Performance is not the right approach. The reason behind this is that the algorithm can end up predicting always the majority class and still achieve accuracy as ~97%. Due to this problem we decide to use Area under the ROC curves (AUC_ROC score) as the Metric of Performance.

Area Under the ROC Curve (AUC_ROC): A Receiver Operating Characteristic (ROC) curve is the most commonly used way to visualize the performance of a binary classifier (Markham, 2014). In statistics, a receiver operating characteristic curve, or ROC curve, is a graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is varied. The curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings (Wikipedia, 2016). Figure 2 shows the ROC curve for year1 data of the dataset. The area under the ROC curve equals the probability that a randomly chosen positive example ranks above (is deemed to have a higher probability of being positive than) a randomly chosen negative example.

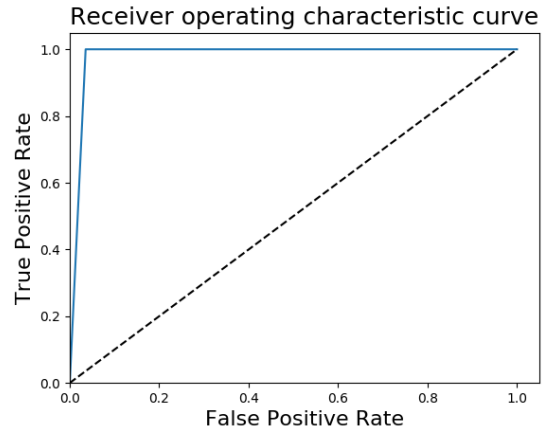


Figure 2. Receiver Operating Characteristic Curve for year 1 data with the AUC score of 0.9670

5. Experiment

5.1. Wellness Analysis of SMOTE Data

To confirm if the classes in the generated dataset (by SMOTE) are distributed similar to the original dataset, we implement a basic anomaly detection algorithm using Auto-Encoder with Neural Networks. For this test, we normalize all the features of the safe class and the bankrupt class separately. Then, we train the Auto-Encoder using 80% safe company data of the original dataset. Then for the remaining original dataset (20% of the safe class and all of the bankrupt class) we check the reconstruction error of the dataset. Based on this error we distinguish it as safe or bankrupt. The algorithm can classify with AUC score of 0.995. Next we pass all of the generated data (SMOTE data) through the Auto-Encoder (trained on original data) and classify based on the reconstruction error. The algorithm classifies the generated data with an AUC score of 0.998. So, we confirm that both the original and generated dataset have similar distribution. We do not use this algorithm as our one of our results because we normalize the safe company data and bankrupt company data separately,

¹(Zikeba et al., 2016)

Table 3. Comparison of AUC score obtained using each technique

Year	1	2	3	4	5
Net2Net	0.967	0.944	0.927	0.930	0.959
FCN	0.959	0.899	0.903	0.899	0.945
FCN¹	0.543	0.514	0.548	0.596	0.699
Ensemble Boosted Trees¹	0.959	0.944	0.940	0.941	0.955

thus having a prior knowledge of the data classification, to train the Auto-Encoder algorithm.

5.2. Model Training

We first implement a Network model using the Net2Net technique. The technique used for training all the Net2Net networks utilized the following hyper parameters. All the models were trained with Adam as the optimizer and binary_crossentropy was used as the cost metric to train the network. PReLU were used as the activation units as better results were obtained with this over ReLu, sigmoid and tanh. The learning rate was scheduled with initial value being the default $5e-3$ (for Adam optimizer) and a reduction in learning rate by a factor of 10 was effected whenever the loss failed to reduce in 3 consecutive epochs. Training was performed for 2000 Epochs for each teacher network. Scaling of the network was done by widening the network first and then deepening it. Widening operation was done to a maximum of 128 neurons and then deepening operations were performed up to 8 hidden layers (discussed in detail in section 6). We implemented our network models using the software package Keras (Chollet, 2015). For year1 data, we start to train the network with 1 neuron and scale it to 8 hidden layers with 128 neurons in each layer. For the subsequent years, we use some of the knowledge gained in training network for year1. While training for year2, 3, 4 and 5 we chop off the network at layer4 and add layers 5, 6, 7 and 8. The intuition behind choosing to chop off at layer 4 is discussed in Section 7

After developing an optimized model using the Net2Net technique, we train the same network architecture using conventional method. We compare the results of the two techniques in section 6.

6. Results

The AUC scores obtained during training is shown in Table 1 (for widening operation) and Table 2 (for deepening operation). Data from Table 1 indicate that the AUC score does not have a significant improvement from 64 neurons to 128 neurons during the widening operation. This is the primary reason why the widening operations were not pursued after 128 neurons. As hidden layers are added, the

improvement in AUC score reduces as we add layer8. This is why we stop at layer8. All the AUC scores are reported on the 20% of the original dataset which was not shown to SMOTE while sampling.

Table 3 compares the results obtained by Net2Net technique with the conventional training method. It can be observed that the network trained using Net2Net technique performs better than the conventional method in all the datasets. The Table also shows the results obtained by (Zikeba et al., 2016) on the same dataset using FCN and Ensemble Boosted trees. Their best results for this dataset were obtained by using Ensemble Boosted Trees. We outperform their FCN for all the years and their Ensemble boosted trees for years 1 and 5.

7. Discussion

High AUC score for year 2,3,4 and 5 for the Net2Net in Table 3 signifies that the data has some correlation over different years, thus utilizing the information learnt about the two classes during year1 helps in learning for subsequent years.

We conducted experiments by chopping the year1 network at layer 7,6,5 and 4. We obtained best results for the years 2, 3, 4 and 5 when we chop at layer4. An intuition for this could be that the features learnt in layer5 and onwards of network 1 were highly specific to year1 which did not contribute much to the other years. But the features learnt till layer 4 were generic to all the years.

8. Conclusions and Future Work

In this paper we show that the neural networks can be used for predictions related to the financial state of the companies with high AUC score. We also show that knowledge of one of the year's could be used to learn a model for other years. To best of our knowledge, the model presented in this paper is among the best performing kernel on this dataset.

For future works we propose to approach this problem as an anomaly detection problem where the companies with safe functioning are non-anomalous. This approach will be

able to deal with the skew class issue. Also, we feel that this problem might be able to red-flag any other financial issue and not just future bankruptcy.

References

- Altman, Edward I and Hotchkiss, Edith. *Corporate financial distress and bankruptcy: Predict and avoid bankruptcy, analyze and invest in distressed debt*, volume 289. John Wiley & Sons, 2010.
- Angelini, Eliana, di Tollo, Giacomo, and Roli, Andrea. A neural network approach for credit risk evaluation. *The quarterly review of economics and finance*, 48(4):733–755, 2008.
- Bernhardsen, Eivind. A model of bankruptcy prediction. Master’s thesis, 2001.
- Bredart, Xavier. Bankruptcy prediction model using neural networks. *Accounting and Finance Research*, 3(2):124, 2014.
- Chen, Tianqi and He, Tong. Higgs boson discovery with boosted trees. In *NIPS 2014 Workshop on High-energy Physics and Machine Learning*, pp. 69–80, 2015.
- Chen, Tianqi, Goodfellow, Ian, and Shlens, Jonathon. Net2net: Accelerating learning via knowledge transfer. *arXiv preprint arXiv:1511.05641*, 2015.
- Chollet, François. Keras. <https://github.com/fchollet/keras>, 2015.
- Iturriaga, Félix J López and Sanz, Iván Pastor. Bankruptcy visualization and prediction using neural networks: A study of us commercial banks. *Expert Systems with applications*, 42(6):2857–2869, 2015.
- Kittler, Josef, Hatef, Mohamad, Duin, Robert PW, and Matas, Jiri. On combining classifiers. *IEEE transactions on pattern analysis and machine intelligence*, 20(3):226–239, 1998.
- Lee, Sangjae and Choi, Wu Sung. A multi-industry bankruptcy prediction model using back-propagation neural network and multivariate discriminant analysis. *Expert Systems with Applications*, 40(8):2941–2946, 2013.
- Lemaître, Guillaume, Nogueira, Fernando, and Aridas, Christos K. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17):1–5, 2017. URL <http://jmlr.org/papers/v18/16-365.html>.
- Markham, Kevin. *ROC Curve*, 2014. URL <http://www.dataschool.io/roc-curves-and-auc-explained/>.
- Shin, Kyung-Shik, Lee, Taik Soo, and Kim, Hyun-jung. An application of support vector machines in bankruptcy prediction model. *Expert Systems with Applications*, 28(1):127–135, 2005.
- Tomczak, Sebastian. *Polish companies bankruptcy data Data Set*, 2016. URL <https://archive.ics.uci.edu/ml/datasets/Polish+companies+bankruptcy+data>.
- Wikipedia. *Receiver operating characteristic*, 2016. URL https://en.wikipedia.org/wiki/Receiver_operating_characteristic#Area_under_the_curve.
- Yang, Qiang and Wu, Xindong. 10 challenging problems in data mining research. *International Journal of Information Technology & Decision Making*, 5(04):597–604, 2006.
- Zhang, Yudong, Wang, Shuihua, and Ji, Genlin. A rule-based model for bankruptcy prediction based on an improved genetic ant colony algorithm. *Mathematical Problems in Engineering*, 2013, 2013.
- Zikeba, Maciej, Tomczak, Sebastian K, and Tomczak, Jakub M. Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction. *Expert Systems with Applications*, 2016.