Application of Deep Learning Networks to Asses corporate Credit Rating

**Dataset size**

The dataset size severely compromises the hypothesis under testing. In statistical learning, each methodology is to be applied according to the dataset availability (size, format, panel data), the target and features nature (quantitative, nominal, image matrix, text), and according to the main interest of the analysis (analytical, binary classification, probabilistic).

In the case of the paper, the dataset size compromises the use of the techniques MLP, CNN, CNN2D, LSTM to validate the hypothesis stated at section 4 (pag. 6). Since the performance of deep learning neural networks often improves with the amount of data available, and given the huge number of features reported, the good performance measures can even be attributed to a possible overfitting.

In case of small data sets, there is some data augmentation techniques reported by Brownlee (2019) and Le Guennec, Malinowski and Tavenard (2016). This is done by applying domain-specific techniques to examples from the training data that create new and different training examples. This can be useful to artificially create new training data from existing training data.

But, there no way around the small sample size limitations. The amount of needed data depends on the complexity of the chosen algorithm. If a linear algorithm achieves good performance with hundreds of examples per class, you may need thousands of examples per class for a nonlinear algorithm, like random forest, or an artificial neural network.

For a discussion on the sample size requirements see Trevor Hastie, Robert Tibshirani, Jerome H. Friedman (2008) and Jain and Chandrasekaran (1981).

**Deep Learning architectures and Credit Score Databases**

In the case of the paper, the dataset size compromises the use of the techniques MLP, CNN, CNN2D, LSTM to validate the hypothesis stated at section 4 (pag. 6).

The hypothesis 3 could be addressed using simulated datasets with many observations. Even if the authors even if they had a data set with more than 1 million degrees of freedom, would not suffice to test the hypothesis 3.

For instance, Panda (2016) compares Deep Convolutional Neural Network performance for three datasets: Bioinformatics datasets, Handwritten dataset (*mnist*) and the Japanese vowel datasets**.** All those datasets are available in the internet platforms like Kaggle, TensorFlow Datasets and UCI Machine Learning Repository. And the criterium for choosing those datasets is that if consists of more observations than the number of instances.This way the author can provide a sound evidence that the proposed algorithms and architectures generalizes to different situations: Computer Vision, Image processing, Speech recognition.

The analogy with credit risk is to provide the performance measures for completely different datasets: retail credit, microcredit in developing countries, Interbank lending market.

Some free available credit score datasets for different countries are provided in this github:

<https://github.com/JLZml/Credit-Scoring-Data-Sets>

**Train/Test Split**

I found the train/test split choice to be very problematic and in dissonance with the literature on machine learning. Classical train/test split choices are 75%-25%, 70%-30%, 80%-20%. But even this has been replaced for k-fold evaluations, using out-of-sample cross validations.

Of course, this stumbles again in the dataset size.

**Justifying Architectures**

The literature in machine learning usually spends a great deal of time explaining and justifying the architectural choices behind each one of the chosen techniques. How many dropout layers is used? What are the dropout percentages at each dropout layer? How many convolutions? Are there fully connected layers? Are there Pooling layers? What is done to prevent overtiffing? What is the learning rate used in backpropagation? What and why a specific activation function were employed at the final layer? What is the possible consequence of choosing another activation function?

I didn’t find enough discussion on the why as a specific deep learning used besides quoting in the literature that is has been done before. More important, for the choses deep learning algorithms o didn’t found an exposition on the particularities of the architectures employed.

**LSTM**

Time series adds the complexity of a sequence dependence among the input variables. The Long Short-Term Memory (LSTM) is a type of neural network designed to handle sequence dependence is called recurrent neural networks. The pros of the LSTM network it allows one to use very large architectures characterized as deep learning. The cons is that the same very large architectures requires very large databases. Beside the already mentioned problem of sample size, the authors didn’t comment on how they when from a time series technique to a panel data datasets, that has on important additional dimension, the sectional dimension.

**Features Selection**

Machine learning algorithms don’t rely on lack of multicollinearity assumptions. As such, you can introduce as many features as possible. However, when you have more features then observations, then overfitting can be an issue. In practice one can concatenate all variables and create a unique identification by individual, and with this identification I would be able the exactly identify the target status. Given those considerations, I consider the hypothesis 1 to be irrelevant.

The more appropriate question, what are the consequences regarding overfitting when one chooses as much features as the number of observations.

**References**:

Mrutyunjaya Panda. Towards the effectiveness of Deep Convolutional Neural Network based Fast Random Forest Classifier, 2016. <https://arxiv.org/abs/1609.08864>

Jain and Chandrasekaran. 39 Dimensionality and sample size considerations in pattern recognition practice, 1982. <https://doi.org/10.1016/S0169-7161(82)02042-2>

Jason Brownlee. How to Configure Image Data Augmentation in Keras, 2019.

Arthur Le Guennec, Simon Malinowski and Romain Tavenard. Data Augmentation for Time Series Classification using Convolutional Neural Networks, 2016. <https://halshs.archives-ouvertes.fr/halshs-01357973/>

Trevor Hastie, Robert Tibshirani and Jerome Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics). 2008. <https://web.stanford.edu/~hastie/Papers/ESLII.pdf>