Credit Bank Default Prediction Based on Machine Learning Approaches

Runqi Jiang(✉)

Department of Math, The Ohio State University, Columbus, OH, USA
jiang.1962@osu.edu

Abstract. With the rapid development of Internet, data science is playing a more and more important role in all fields. Especially in the financial industry, the application level of big data has become the embodiment of enterprise competitiveness. Contemporarily, among many types of data science, deep learning has developed most rapidly, where plenty of relevant achievements have been achieved accordingly. It's a combination of data science and human neural networks. This paper will first introduce the basic descriptions and principles of deep learning and model description. Subsequently, a specific application of bank default prediction combined will be demonstrated with prediction variables, common models, existing disadvantages, and prospects. Based on the analysis, it should be noted that a clear limitation appears from both the perspectives of theoretical knowledge and construction of mode, i.e., scholars still could not solve the problem of high requirement of training sample and computation sources. These results shed light on guiding financial application in terms of machine learning concepts.

Keywords: Machine Learning · Stock Price Prediction · Credit Bank Default · Deep Learning · Bigdata Analysis

1 Introduction

Big data refers to the database that the volume of it is too enormous to be processed by conventional data processing tools [8, 11, 13]. It has four main characters: huge volume, high velocity, wide variety, low value density [4, 9]. Nowadays, Financial business has become the major application area of bigdata. It mainly focuses on the function of Credit Investigation, Information Verification, Credit pre-judgement, Providing credit decision for third-party business.

As one of many branches of Machine learning, deep learning refers to making the processing, judging of machine be like human, in another words, more intelligent. Contemporarily, as the rapidly evolution of the state-of-art machine learning scenarios, many of new models have been created. In addition to deep convolutional networks (e.g., AlexNet [5]) that performs object recognition tasks. In addition, Convolutional Neural Networks (CNN) also include many excellent models for plenty of applications [3, 7, 14]. For example, previous researchers apply the convolution process in different ways for different tasks and produce very good results on these tasks. Basically, convolution

D. Qiu et al. (Eds.): ICBBEM 2022, AHIS 5, pp. 859–867, 2023.
https://doi.org/10.2991/978-94-6463-030-5_85

has many excellent properties compared to the original fully connected network. For example, it only partially connects with the neurons in the previous layer, and the same convolution kernel can be reused on the input tensor, that is said that the feature detector can repeatedly detect the presence or absence of this local feature on the input image. This is an excellent property of convolutional networks, which greatly decreases the quantity of parameters between two layers.

Recurrent neural networks are an important part of deep learning, which allows neural networks to process sequence data such as text, audio, and video [1, 10]. They can be used for high-level semantic understanding of sequences, sequence labeling, and even to generate new sequences from a fragment. There are many artificial intelligence applications that rely on recurrent deep neural networks, and RNNs can be found in products from Google (voice search), Baidu (Deep Speech), and Amazon. The basic RNN structure is difficult to deal with long sequences, however a special RNN variant, the "Long Short Term Memory (LSTM)" network, can handle long sequences well [1]. This powerful model has achieved landmark results in tasks such as translation, speech recognition, and image description. Therefore, recurrent neural networks have been widely used in recent years. A deep generative model can demonstrate its understanding of the data by generating completely new samples, although these generated samples are very similar to those training samples. Many of these models are related to the idea of the previous autoencoder, which has an encoder function that maps data to a representation, and a decoder function (or generator) that maps this abstract representation to the original data space.

In addition, many generative models are also applied to the idea of GAN, i.e., the generator is prompted to generate very realistic images through the confrontation between the discriminator and the generator [6, 15]. In variational autoencoders, an encoder and decoder should be trained through samples. In the process, intermediate hidden variables are got. If one need to generate a new image, the only need is to sample the hidden variable and put it into the decoder to complete the generation. In Generative Adversarial Networks, a discriminative model and a generative model are defined. First, generated samples mixed with real samples will be delivered to the discriminative model to train its ability to distinguish between true and false, and then fix the discriminative model and train the generative model to generate more realistic images.

The rest part of the paper is organized as follows. The Sect. 1 will introduce some of the common machine learning approaches. Subsequently, the Sects. 2 and 3 will demonstrate the applications of machine learning in credit default and stock price prediction, respectively. Eventually, a brief summary and future outlook will be given in Sect. 4.

2 Model Descriptions

2.1 Tree Algorithms

2.1.1 Decision Tree

Decision tree (DT) is a decision support tool in tree structure, where the every subbranch represents a test output, each leaf node represents a category, and each internal node denotes for a test on an attribute. Normally, circles are used to represent chance nodes,

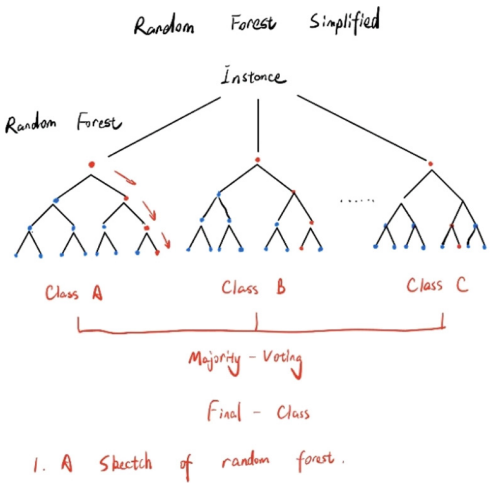

Fig. 1. A sketch of random forest. (photo credit: Original)

squares are used to represent decision nodes, and triangles are used to represent end nodes. To give a mathematical description, the function can be given as follows:

$$Igain(s) = H(t) - H(s,t) \tag{1}$$

$$\Phi(s,t) = 2P_L P_R Q(s|t) \tag{2}$$

Decision Tree is a rather simple concepts but the fitting performances are limited in most of the cases.

2.1.2 Random Forest

Random Forest is an ensemble method by constructing large amounts of decision tree at training times. A sketch of the algorithms is given in Fig. 1. To give a mathematical description, the function can be given as follows:

$$\left|m_{M,n}(\boldsymbol{x}) - \tilde{m}_{M,n}(\boldsymbol{x})\right| \leq \frac{b_n - a_n}{a_n}\tilde{m}_{M,n}(\boldsymbol{x}) \tag{3}$$

where

$$a_n \leq N_n(\boldsymbol{x}, \boldsymbol{\theta}) \leq b_n; a_n \leq \frac{\sum_{m=1}^{M} N_n \boldsymbol{x}}{M}, \theta_m \leq b_n \tag{4}$$

Although random forests typically achieve higher accuracy than individual decision trees, they sacrifice the interpretability inherent in decision trees. In contrast, decision trees are a fairly small family of machine learning approaches that are as easy to interpret as linear models, rule-based models, and attention-based models, which can be regarded as one of the most desirable features of decision trees.

2.1.3 Xgboost

Xgboost is an improvement approaches based on the concepts of gradient boosting. To give a mathematical description, the flow of algorithms can be given as follows. Primarily, the model with constant value is initialized:

$$f_0(x) = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{N} L(y_i, \theta) \tag{5}$$

Subsequently, for m = 1 to M, one computes the gradients and hessians as:

$$g_m(x_i) = \left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)} \right]_{f(x)=f_{m-1}(x)} \tag{6}$$

$$g_m(x_i) = \left[\frac{\partial^2 L(y_i, f(x_i))}{\partial f(x_i) 62} \right]_{f(x)=f_{m-1}(x)} \tag{7}$$

Afterwards, one uses the training set by solving the optimization problem:

$$\phi_m = \underset{\phi \in \Phi}{\operatorname{argmin}} \sum_{i=1}^{N} \frac{h_m(x_i)}{2} \left[-\frac{g_m(x_i)}{h_m(x_I)} - \phi(x_i) \right]^2 \tag{8}$$

where the final results can be described as:

$$f_m(x) = \alpha \phi_m(x) \tag{9}$$

The training set can be described as:

$$\left\{ x_i, -\frac{g_{m(x_i)}}{h_m(x_i)} \right\}_{i=1}^{N} \tag{10}$$

The model update will follow the form of:

$$f_{(m)}(x) = f_{m-1}(x) + f_m(x) \tag{11}$$

and the output can be described as:

$$f(x) = f_M(x) = \sum_{m=0}^{M} f_m(x) \tag{12}$$

As for the advantages of the Xgboost is that the performances will be improved greatly. However, the training requires a large number of computer sources as well as memeory.

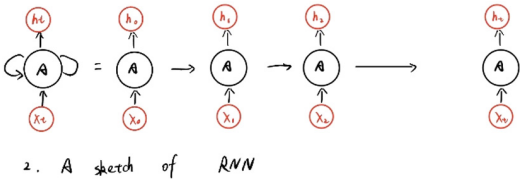

Fig. 2. A sketch of RNN. (photo credit: Original)

2.1.4 LightGbm

LightGbm is Light Gradient Boosting Machine. It is a distributed boosting framework which is totally free and famous for its open sources. LightGBM is initially designed and developed by Microsoft. The engineers utilized the decision tree algorithms as the basic theory to build up the structure of the LightGBM. The primary functions of it are ranking, classification, and machine learning. To give a mathematical description, the flow of algorithms can be given as follows. The function can be given as:

$$V_{j|0}(d) = \frac{1}{n_0}\left[\frac{\left(\sum_{\{x_i \in O: x_{ij} \le d\}} g_i\right)^2}{n^j_{l|0}(d)} + \frac{\left(\sum_{\{x_i \in O: x_{ij} > d\}} g_i\right)^2}{n^j_{r|0}(d)}\right] \tag{13}$$

As for the pros of the model, the time cost of the training and efficiency are much larger compared to other algorithms with the smaller memory requirements and better accuracy. However, it still faces the limitation of overfitting as the tree leaf-wise can cause overfitting in much complex and compatibility with datasets.

2.2 Neural Networks

2.2.1 RNN

RNN is Recurrent neural network. In the directed and undirected graph, RNN is a class of artificial neural networks that connect nodes. The internal state of RNN can be used to process multiple length sequences of inputs which enhance the applicability. Since the high applicability of RNN, lots of applications and implications become applicable. A sketch of the RNN is given in Fig. 2.

Generally speaking, RNN has plenty of advantages, e.g., less sensitive to the input length, model size does not increase along the input size increases, internal memory of RNN can be utilized for processing the arbitrary series of inputs. On the other hand, it should be noted that it has advantages including low computation speed, difficulty for parameter searching and pronging to problems such as exploding and gradient vanishing.

2.2.2 LSTM

LSTM, which is different from regular and common neural networks, possesses the feedback mechanisms according to connections. It can process single data points and entire sequences of data includes speech and video and widely used in plenty of situations

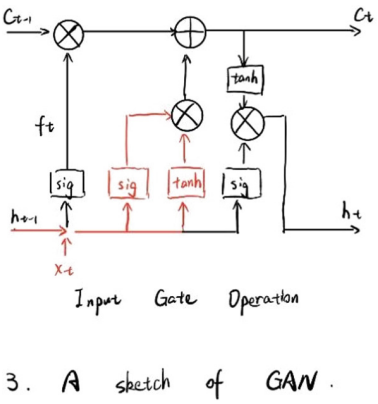

Fig. 3. A sketch of GAN. (photo credit: Original)

on account of the complexity of the models. However, it dose have some disadvantages, e.g., longer training time, larger memory, easily to be overfitted, etc.

2.2.3 GAN

GAN is Generative adversarial network which was originally developed and built by Goodfellow in 2014. It builds up a system of competing neural network models. Those models are competitors to each other and can be utilized to analyze, capture and copy the variations in a certain dataset. In the application of GAN, there are three parts-generative, adversarial, and networks. Due to the shaper ad clearer, GAN can bring better modeling and data distribution. In theory, GANs can train any kind of generator network. There is no need to use the Markov chain to repeatedly sample, without inferring in the learning process, without complicated variational lower bounds, avoiding the difficulty of approximating the difficult probability of calculation. As for disadvantages, the instability makes the GAN hard to train. Besides, uses often face the problem of mode collapse issue. There could be missing pattern in the learning process of GANs. A sketch of GAN is illustrated in Fig. 3.

3 Credit Default Forecast

Data is generated and stored in large quantities based on the network. According to related research, with the entry of the information age, the amount of information has experienced explosive growth, and the compound annual growth rate of the total global data is as high as 50%. More than 5 terabytes of data are produced per person. In finance, too, data is so vast and scattered across different systems, each enterprise tries to process and manage those data since the core of financial agent is risk management. In the loan default prediction, data science is used for predicting the factors which will affect default probability and predicting if a person will be defaulted in his loan behavior [2]. Then, the choice of method or model is an important step in the prediction. Thereinto, deep learning scenarios are a really famous and widely applied approach for modeling

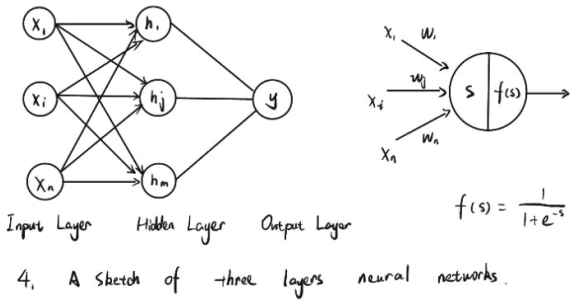

Fig. 4. A sketch of three layers neural networks. (photo credit: Original)

business data [16]. ANN is a framework of machine learning algorithms that provide an opportunity to learn nonlinear patterns from complex datasets. (Baghdasaryan 2021). There are some other popular models in deep learning which can be helpful for this task, such as Convolutional Neural Network (CNN), Deep Neural Network (DNN) other than some normal classification models like Logistic Regression (LR), Decision Tree (DT), Naïve Bayes (NB) and Support Vector Machines (SVM). After choosing model, pre-processing data is required. To deal with large amounts of missing values, visualization of missing values is a good idea. And data cleaning is also a usual method to deal with missing value. After managing all data, a meaningful actioned called feature engineering must be done. There are 4 parts: Feature derivation, Feature abstraction, Feature scaling and Feature selection. Finally, by the Figs. 4 and 5 illustrated, XGBoost behave better than all other models in the experiment [12].

At the present stage, there are still a lot of disadvantages in deep learning, which is mainly in the aspect of theoretical knowledge as well as the construction of the model. The nonlinear model used in deep learning need to be combined by shallow neural network and deep neural network. A deep network undoubtedly has better ability to a represent a nonlinear model, but it also means more complicated learning samples. The unknown of the training sample and computing resources required, in addition, the property of non-convex functions which the models of deep learning usually are, greatly increases the difficulty of theoretical research on deep learning. Moreover, extremely huge training samples and the demand of the computing resource modeling difficulty greatly increase. Although deep learning has great potential, under the limited existing technology and high time cost trade-off, a prediction error of the model by DNN (deep neural networks) sometimes even higher than an ordinary linear model does, this is because in the big data contain abundant information dimension, Thus, such a high-capacity complex model as DNN is in an under-fitting state. In the future, by technological innovation of Artificial Intelligence, deep learning will play a more important role in financial region.

4 Conclusion

IN summary, this paper introduces development of deep learning and its related application in a specific financial area (bank credit default). Based on the analysis, the performances of the state-of-art machine learning approaches are appealing while still have

lots of drawbacks and defects. Contemporarily, there are plenty of issues needed to be addressed, though some outstanding results have been achieved. In detail, the over-fitting effects and parameter selection are tougher issues to be solved. In the future, better approaches ought to be proposed to resolve the restrictions of current approaches. Overall, these results offer a guideline for machine learning implementation in financial field.

References

1. Sherstinsky, "Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network," Physica D: Nonlinear Phenomena, vol. 404, 2020, 132306.
2. Vardan, et al, "Comparison of Econometric and Deep Learning Approaches for Credit Default Classification." Wiley Online Library, John Wiley & Sons, Ltd, 10 May 2021,
3. Chen, et al. "Multi-domain gated CNN for review helpfulness prediction," The World Wide Web Conference, 2019.
4. A. Mauro, M. Greco, and M. Grimaldi, "A formal definition of Big Data based on its essential features," Library Review, vol. 1, 2016.
5. N. Iandola, et alB "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model sizeB" arXiv preprint arXiv:1602.07360, 2016.
6. Liang, and Y. Zhou, "A review: generative adversarial networks," 2019 14th IEEE conference on industrial electronics and applications (ICIEA). IEEE, 2019.
7. L. Alzubaidi, et al, "Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions," Journal of big Data, vol. 8(1), 2021, pp. 1–74.
8. M. Hilbert, "Big data for development: A review of promises and challenges," Development Policy Review, vol. 34(1), 2016, pp. 135–174.
9. N. F. Hordri, et al. "A systematic literature review on features of deep learning in big data analytics," International Journal of Advances in Soft Computing & Its Applications, vol. 9(1), 2017.
10. P. Dhruv and N. Subham, "Image classification using convolutional neural network (CNN) and recurrent neural network (RNN): A review," Machine Learning and Information Processing, 2020, pp. 367–381.
11. P. Jain, G. Manasi, and N. Khare. "Big data privacy: a technological perspective and review," Journal of Big Data, vol. 3(1), 2016), pp. 1–25.
12. S. Orkun, "A Deep Neural Network (DNN) based classification model in application to loan default prediction," Theoretical and Applied Economics Volume XXVI, vol. 621(4), 2019 Winter, pp. 75–84.
13. S. Sagiroglu and D. Sinanc, "Big data: A review," 2013 international conference on collaboration technologies and systems (CTS), IEEE, 2013.
14. T. Kattenborn, et al, "Review on Convolutional Neural Networks (CNN) in vegetation remote sensing," ISPRS Journal of Photogrammetry and Remote Sensing, vol. 173, 2021, pp. 24–49.
15. Y, Chika, and O. Ugot. "A review of generative adversarial networks and its application in cybersecurity," Artificial Intelligence Review, vol. 53(3), 2020, pp. 1721–1736.
16. Z. Kai, et al, "A Deep Metric Learning Approach for Weakly Supervised Loan Default Prediction." Journal of Intelligent & Fuzzy Systems, IOS Press, 1 Jan. 2021.