2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE 2021)

2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE) | 978-1-6654-1540-8/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICBAIE52039.2021.9390010

A Deep Neural Network Based Model For Stock Market Prediction

Yuye Liu[1]
The University of Warwick,
Coventry, England
Yuye.Liu@warwick.ac.uk

Chen Cao[1]
The University of Sydney
Sydney, Australia,
ccao9775@uni.sydney.edu.au

Weixin Huang[2]
Sun Yat-Sen University
GuanDong, China
tedhuangwx@hotmail.com

Shi Hao[3]
Australian National University
Canberra, Australia
dulyhao@163.com

Abstract—The stock market has gradually become an indispensable part of the securities industry and the entire financial industry, and has attracted more and more investors' attention. Therefore, the analysis and prediction of the stock market trend has great theoretical significance and considerable application value. In this paper, an algorithm based on a deep neural network is proposed to build a stock prediction model. The neural network model is a complex network system formed by a large number of simple neurons widely connected to each other. It is a highly complex nonlinear dynamic learning system that can effectively mine attributes of different dimensions for prediction. This model performs better than other comparative models in predicting the trend of stocks. Specifically, the return value of our neural network model is 1059 higher than the Xgboost algorithm and 2257 higher than the random forest algorithm.

Index Terms—Neural Networks, Deep learning algorithms, Data Mining,

I. INTRODUCTION

The stock market has been growing in recent years and has gradually become an indispensable part of the securities industry and even the entire financial industry. It has received more and more attention from investors. Therefore, the analysis and prediction of the stock market trend has great theoretical significance and considerable application value. With the continuous application of new theories and technical analysis methods to the technical analysis of stocks, the Chinese securities market has gradually become more rational, and people from inside and outside the industry have more urgent needs for the application of new technical analysis to the stock market. Therefore, this paper uses the neural network (NN) model to predict the trend of stocks and propose a decision-making plan.

With the rapid development of machine learning and deep learning, deep learning algorithms have been widely used in various industries to solve practical problems. In recent years, the application of deep learning prediction models in predicting stock price trends has received extensive attention from researchers, which can help investors make better investment decisions. Therefore, this paper aims to use the data set provided by Jane Street to build stock forecasting models and develop trading strategies. In this paper, an algorithm is proposed based on a deep neural network to build a stock prediction model, which performs better than other comparative models in predicting stock trends. The neural network model is a complex network system formed by a large number of simple neurons widely interconnected. It is a highly complex nonlinear dynamic learning system that can effectively mine attributes of different dimensions for prediction.

Experiments are conducted on the data set provided by Jane Street, which comes from major stock exchanges in the world, to predict stock trends and develop trading plans to maximize returns.

Python is the main tool for data processing, modeling and analysis. In the experimental stage, feature engineering processing is performed, including compressing memory, extracting temporal features, statistical features, and selecting important features. The experimental results show that the stock prediction model based on neural network is better than other machine learning models, such as Xgboost algorithm and random forest algorithm. Specifically, the return value of our neural network model is 1059 higher than the Xgboost algorithm and 2257 higher than the random forest algorithm. In addition, the stock return trends of different trading numbers are also studied and some important findings are obtained to guide future work. Experiments prove that the idea of using neural network algorithms is effective for stock prediction tasks.

In summary, the contributions of this paper are as follows:

1. A stock prediction model is proposed based on neural network algorithms to formulate trading strategies, which can effectively mine attributes of different dimensions.

2. During the experiment, detailed feature engineering processing were performed, such as memory compression, temporal feature extraction, statistical features, and important feature selection. The experiments show that feature engineering is effective for training models.

3. The neural network method is evaluated on the data set provided by Jane Street. Experiments show that the method is

978-1-6654-1540-8/21/$31.00 ©2021 IEEE

superior to comparison methods such as Xgboost algorithm and random forest algorithm. The stock return trends of different trading numbers show some interesting conclusions that can guide future work.

A. Related Work

Data mining refers to mining the important information hidden in the data from massive data through algorithms. The hidden information behind the data can be obtained through preprocessing, feature engineering and modeling. Accurate stock forecasts are highly helpful for investment decisions.

There are many machine learning and deep learning methods that have strong predictive capabilities.

The regression model is introduced in [1], [2] and [3]. The goal of function-to-function regression is to establish a mapping from function predictor variables to function response. A functional regression model based on pattern sparse regularization is proposed, and the modal sparse regularization method is used to automatically filter irrelevant functions. [4] [5] [6] introduced very classic machine learning methods, among which the linear regression, logistic regression and random forest methods mentioned in this paper have good performance and interpretability, and their application scenarios are also extremely good. In [7], three machine learning models of GBDT, Xgboost and LightGBM are used to predict monthly rent. By comparing and analyzing the prediction results under different data sets training, it is found that the Xgboost and LightGBM models are better than the traditional GBDT model.

In [8], a deep learning model is used to predict stocks. The model first extracts news events, then expresses them as dense vectors, and uses Neural Tensor Networks to train events. Finally, a deep convolutional neural network is used to model the short-term and long-term effects of the event.

II. FEATURE ENGINEERING

In this part, data preprocessing is introduced.

First of all, the data are standardized and normalized. Data normalization is to limit the data that needs to be processed to a certain range after processing. Data standardization can not only facilitate subsequent data processing, but also speed up the convergence of the model during run time.

Then, the discrete values in the data set are encoded. In many machine learning tasks, features are not always continuous values, but may be categorical values. To facilitate model training, coding can be performed. Among them, one method is to use One-Hot Encoding. One-Hot encoding is also known as one-bit effective encoding. The method is to use N-bit status registers to encode N states. Each state has its own independent register bit, and at any time, only one valid. One-hot encoding solves the problem that the classifier is not good at processing attribute data, and to a certain extent also plays a role in expanding features. Its values are only 0 and 1, and different types are stored in the vertical space. When the number of categories is large, the feature space becomes highly large. The discrete numerical coding method used in this experiment is one-hot coding.

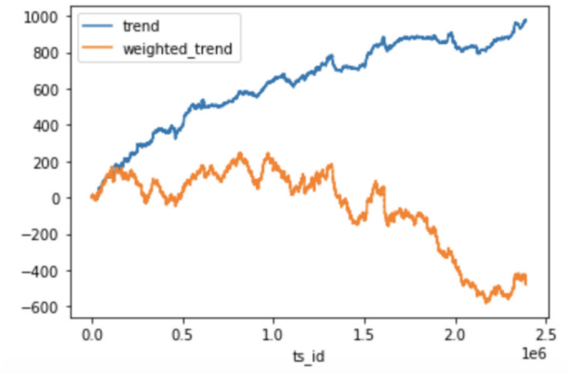

Figure 1. Stock return trends of different trading numbers

III. NEURAL NETWORK-BASED PREDICTION MODEL

In this section, the neural network-based stock prediction model is introduced.

Neural network is a complex network system formed by a large number of simple neurons widely connected to each other. It is a highly complex nonlinear dynamic learning system. Neural networks have large-scale parallelism, distributed storage and processing, self-organization, self-adaptation and self-learning capabilities, and are particularly suitable for handling inaccurate and fuzzy information processing problems that require consideration of many factors and conditions at the same time.

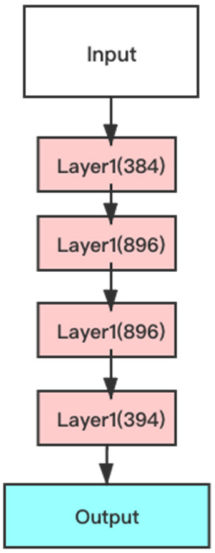

Figure 2. Neural Network Architecture

The neural network model is described based on the mathematical model of neurons. It has the following advantages:

1. Parallel distributed processing;
2. Highly robust and fault-tolerant;
3. Distributed storage and learning capabilities;

4. It can fully approximate complex nonlinear relationships.

There are now dozens of artificial neural network models. Typical neural network models that are more commonly used include BP neural network, Hopfield network, ART network and Kohonen network. In the experiment of this paper, BP neural network is used.

BP neural network is a multi-layer network that trains the weights of nonlinear differentiable functions. It is a kind of machine learning technology that simulates the working principle of the human brain to realize artificial intelligence, supports the processing of multiple types of data of images, text, speech and sequences, and can realize classification, regression and prediction. Since the neural network is built on the basis of many neurons, each neuron is a learning model, and these neurons are called Activation Units. Take the logistic regression model as an example, adopt some features as input and give a logical output. A neural network is a collection of multiple neurons combined together.

Neural network has four basic characteristics.

1. Non-linearity. Non-linear relationship is a universal characteristic of nature. The wisdom of the brain is a nonlinear phenomenon. Artificial neurons are in two different states of activation or inhibition. This behavior is mathematically expressed as a nonlinear relationship. The network consisting of threshold neurons has better performance, and can improve fault tolerance and storage capacity.

2. Non-limiting. A neural network is usually composed of multiple neurons connected extensively. The overall behavior of a system not only depends on the characteristics of a single neuron, but may be mainly determined by the interaction and interconnection between the units. Simulate the non-limitation of the brain through a large number of connections between units.

3. Very qualitative. Artificial neural network has the ability of self-adaptation, self-organization and self-learning. Not only does the information processed by a neural network change in various ways, but the nonlinear dynamic system itself is constantly changing as it processes the information.

4. Non-convexity. The evolution direction of a system depends on a particular state function under certain conditions. For example, the energy function, its extreme value corresponds to the relatively stable state of the system. Non-convexity means that this function has multiple extreme values, so the system has multiple stable equilibrium states, which can lead to the diversity of system evolution.

In the complex financial market, the trend of stocks can be affected by many factors, and market turbulence makes a single forecast full of uncertainty. These characteristics of neural network are suitable for such prediction environment, which is a highly targeted model application.

IV. MODEL TRAINING

In model training, it is important to choose an appropriate loss function. In this experiment, the loss function chosen is the binary cross-entropy loss function because in this task, there is a need for judging the rise and fall of the stock based on the information of the sample. Its essence is a binary classification problem, therefore, binary cross entropy function is selected.

For the choice of optimizer, the Adam optimizer is chosen. Adam has the following remarkable characteristics:

(1) It is simple to implement, computationally efficient, and requires less memory. (2) Its parameter update is not affected by the scaling transformation of the gradient. (3) Its hyperparameters have good interpretability, and usually do not need to be adjusted or only require plenty of less fine-tuning. (4) Its update step size can be limited to a rough range. (5) It can naturally realize the step size annealing process. (6) It is highly suitable for large-scale data and parameter scenarios. (7) It can be used for unstable objective function. (8) It is suitable for the problem of sparse gradient or very noisy gradient

In summary, the optimizer selected in this experiment is Adam.

In the training process, a suitable learning rate helps the model learn more efficiently. In this experiment, we choose the learning rate to be 1e5, and the amount of data in each batch is 4096.

V. EXPERIMENTS

In this part, experimental results and experimental analysis are shown.

In order to comprehensively evaluate the performance of the proposed method, classic machine learning methods is chosen and use the same evaluation indicators and data sets for comparison. In this paper, the calculation formula used to evaluate model performance is as follows:

Each date i is defined:

$$p_i = \sum_j (weight_{ij} * resp_{ij} * action_{ij})$$

$$t = \frac{\sum p_i}{\sqrt{\sum p_i^2}} * \sqrt{\frac{250}{|i|}},$$

where |i| is the number of unique dates in the test set. The utility is then defined as:

$$u = min(max(t, 0), 6) \sum p_i.$$

Table 1 shows the experimental results of other machine models and this model. It is easy to find that our neural network model has the highest return value. Specifically, the return value of our neural network model is 1059 higher than the Xgboost algorithm and 2257 higher than the Random Forest algorithm. Experiments show that neural network algorithms used in this paper and feature engineering methods are effective for stock prediction tasks.

Table 1. The results of different models used for stock prediction tasks.

Models	Unity
NN	**6684**
Xgboost	5625
Random Forest	4427

VI. CONCLUSIONS

In this paper, a stock prediction model is proposed based on neural network algorithms to formulate trading strategies, which can effectively mine attributes of different dimensions. First, detailed feature engineering processing is performed, such as data normalization and standardization, time feature extraction, statistical features, and important feature selection. Neural network methods are evaluated on the data set provided by Jane Street. Experiments show that method in this paper is superior to comparison methods such as Xgboost algorithm and random forest algorithm. The stock return trends of different trading numbers show some interesting conclusions that can guide future work.

ACKNOWLEDGEMENT

We sincerely thank Jane Street for the data set.

REFERENCES

[1] Tibor Kiss,Claudia RochJan,Jan Strunk.A logistic regression model of determiner omission in PPs.2010

[2] Mohiuddeen Khan,Kanishk Srivastava. Regression Model for Better Generalization and Regression Analysis.2016

[3] Autcha Araveeporn, Choojai Kuharatanachai. Comparing Penalized Regression Analysis of Logistic Regression Model with Multicollinearity.2018.

[4] Breiman, Leo. Random Forests. Machine Learning 45 (1), 5-32, 2001.

[5] David W Hosmer, Stanley Lemeshow. Applied logistic regression. Technometrics. 2000.

[6] Breiman, L., Friedman, J. Olshen, R. and Stone C. Classification and Regression Trees, Wadsworth, 1984.

[7] Xiang Wei, Mengzhong Ji, Jun Peng.The application analysis of forecasting housing monthly rent based on Xgboost and LightGBM algorithm.2019.

[8] Deep Learning for Event-Driven Stock Prediction,Xiao Ding, Yue Zhang, Ting Liu, Junwen Duan, 2015, IJCAI.

[9] Yue Zhang and Stephen Clark.Syntactic processing using the generalized perceptron andbeam search. Computational Linguistics, 37(1):105–151,2011.

[10] Boyi Xie, Rebecca J. Passonneau, LeonWu, and German G. Creamer. Semantic frames to predict stock price movement. In Proc. of ACL, pages 873–883,2013.

[11] William Yang Wang and Zhenhao，Hua. A semiparametric gaussian copula regression model for predicting financial risks from earnings calls. In Proc.of ACL, pages 1155–1165, 2014.

[12] Paul C Tetlock. Giving content to investor sentiment: The role of media in the stock market. The Journal of Finance, 62(3):1139–1168, 2007.