DECLARATION

I hereby declare that this seminar report entitled "**Attention-based CNN-LSTM and XGBoost hybrid model for stock prediction**" has been prepared by me in partial fulfillment of the requirement for the Master of Computer Science degree by University of Calicut. I also declare that this seminar report has not been submitted by me fully or partially for the award of any other degree, diploma, title or recognition.

PLACE:UNIVERSITY OF CALICUT

DATE:

ADITHYAN B ASOK

CUAVCMF002

Attention-based CNN-LSTM and XGBoost hybrid model for stock prediction

Adithyan b asok
MSc Computer Science
Department of Computer Science
University of Calicut,India
adithyanbasok@gmail.com

Abstract—Stock market plays an important role in the economic development. Due to the complex volatility of the stock market, the research and prediction on the change of the stock price, can avoid the risk for the investors. The traditional time series model ARIMA can not describe the nonlinearity, and can not achieve satisfactory results in the stock prediction. As neural networks are with strong nonlinear generalization ability, this paper proposes an attention-based CNN-LSTM and XGBoost hybrid model to predict the stock price. The model constructed in this paper integrates the time series model, the Convolutional Neural Networks with Attention mechanism, the Long Short-Term Memory network, and XGBoost regressor in a non-linear relationship, and improves the prediction accuracy. The model can fully mine the historical information of the stock market in multiple periods. The stock data is first pre processed through ARIMA. Then, the deep learning architecture formed in pretraining-finetuning framework is adopted. The pre-training model is the Attention-based CNN-LSTM model based on sequence-to-sequence framework. The model first uses convolution to extract the deep features of the original stock data, and then uses the Long Short-Term Memory networks to mine the long-term time series features. Finally, the XGBoost model is adopted for fine-tuning. The results show that the hybrid model is more effective and the prediction accuracy is relatively high, which can help investors or institutions to make decisions and achieve the purpose of expanding return and avoiding risk. Source code is available at https://github.com/ zshicode/Attention-CLX-stock-prediction.

Index Terms—Attention mechanism, Convolutional Neural Networks, Long Short-Term Memory, XGBoost, stock prediction

I. INTRODUCTION

STOCK market plays an important role in the economic development. Due to the high return characteristics of stocks, the stock market has attracted more and more attention from institutions and investors. However, due to the complex volatility of the stock market, sometimes it will bring huge loss to institutions or investors. Considering the risk of the stock market, the research and prediction on the change of the stock price can avoid the risk for the investors.

The traditional time series model ARIMA can not describe the nonlinear time series, and needs to satisfy many pre conditions before modeling, and can not achieve remarkable results in the stock forecasting. In recent years, with the rapid development of artificial intelligence theory and technology, more and more researchers apply artificial intelligence method to the financial market. On the other hand, the sequence mod eling problem, focusing on natural language sequences, protein sequences, stock price sequences, and so on, is important in the field of artificial intelligence research [8], [13]. The most representative artificial intelligence method is neural networks, which are with strong nonlinear generalization ability.

Recurrent Neural Network (RNN) was adopted for analyz ing sequential data via neural network architecture, and Long Short-Term Memory (LSTM) model is the most commonly used RNN. LSTM introduced gate mechanism in RNN, which can be seen as simulation for human memory, that human can remember useful information and forget useless information [6]. Attention Mechanism [7], [16] can be seen as simulation for human attention, that human can pay attention to useful information and ignore useless information. Attention-based Convolutional Neural Networks (ACNN) are widely used for sequence modeling [4], [10]. Combining Attention-based Con volutional Neural Networks and Long Short-Term Memory, is a self-attention based sequence-to-sequence (seq2seq) [15] model to encode and decode sequential data. This model can solve long-term dependency problem in LSTM, hence, it can better model long sequences. LSTM can capture particular long-distance correspondence that fits the sturcture of LSTM itself, while ACNN can capture both local and global corre spondence. Therefore, this architecture is more flexible and robust.

Transformer [16] is the most successful sequential learning self-attention based model. Experiments on natural language processing demonstrates that Transformer can better model long sequences. Bidirectional Encoder Representation Trans former (BERT) with pretraining [2] can perform better than the basic Transformer. Pretraining is a method to significantly improve the performance of Transformer (BERT).

This paper proposes a hybrid deep learning model to predict the stock price. Different from the traditional hybrid prediction model, the proposed model integrates the time series model ARIMA and the neural networks in a non-linear relationship, which combines the advantages of the two vanilla models, and improves the prediction accuracy. The stock data is first preprocessed through ARIMA. The stock sequence is put into neural networks (NN) or XGBoost after preprocessing via ARIMA(p=2,q=0,d=1). Then, the deep learning architecture formed in pretraining-finetuning framework [2], [5] is adopted. The pre-training model is the Attention-based CNN-LSTM

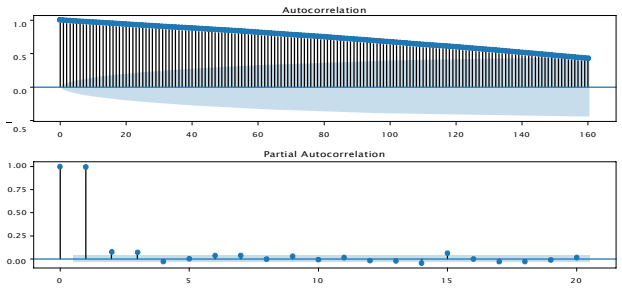

Fig. 4. The ACF and PACF of the original sequence.

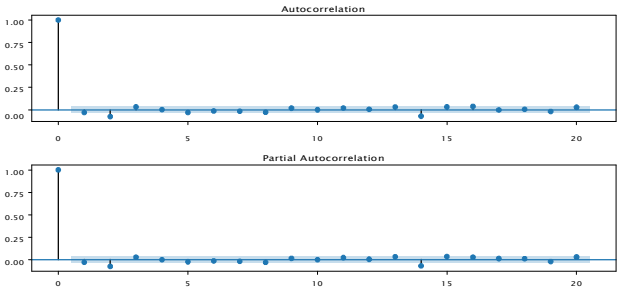

Fig. 5. The ACF and PACF of the first-order difference.

Seq2seq suppress the effect of noise through encoder-decoder architecture. Based on deep learning, hidden information of state is depicted more effectively, while the model would not satisfy the assumptions of linear property of stock price. LSTM receives the context from Attention-based CNN (ACNN)encoder.

The ACNN encoder block consists of self-attention layer and CNN. Q, K, V are computed through Eq. (9) after self-attention layer, and H is computed through Eq. (12). This is the input of LSTM decoder block. Encoder-decoder layers depicts relationship between current sequence and previous sequence, and relationship between current sequence and embedding. Encoder is still with multi-head mechanism. When the k-th embedding is being decoded, only $k-1$-th and previous decoding can be seen. This multi-head mechanism is masked multi-head attention.

Attention-based CNN (ACNN) can capture both global and local dependency that LSTM may not [4], which enhance the robustness. In our proposed encoder-decoder framework, we can adopt a ACNN-LSTM structure. Attention is usually before memory in human cognitive system. The reason why ACNN can capture long-term dependency, is that it integrates multi-head self attention and convolution. Combining LSTM and ACNN can enhance both structural advantages and ability for time-series modeling. Integrating multi-head attention and multi-scale convolutional kernel, ACNN enocder can capture saliency that LSTM may not, while LSTM can better depict time-series property.

1) Fine-tuning: After decoding, the output is obtained through a XGBoost regressor for precise extraction of features and fine-tuning. Our proposed Attention-based CNN-LSTM and XGBoost hybrid model is so called AttCLX, which is shown on Fig. 6.

As the fine-tuning model, XGBoost [1] is with strong expansion and flexibility. It integrates multiple tree models to build a stronger learner model. Based on the pre-training, we propose fine-tuning based on XGBoost, and establish a regression prediction model for stock data. XGBoost fine-tuning model also achieves better predictive ability and generalization ability.

II. EXPERIMENTS

A. Modification of model

After preprocessing by ARIMA ,the nueral networks,is a two- dimensional matrix of data at intravals of aperiod of ttime, with a size of TimeWindow Features. In Empirical studies on stock prediction, features include the basic stock market data (opening price, closing price, highest price, lowest price, trading volume, trading amount). The ARIMA-processing sequence along with the residual sequence are also concatenated as features.

We adopted look back trick for time-series forecasting, and the look back number is 20, i.e. o_t can be obtained through $o_{t-1}, \cdots, o_{t-5}$. This means that the TimeWindow width is 20. The layer number of LSTM is 5, and the size is 64. The epoch number is 50. Model is trained by introducing dropout [14], and the dropout rate is 0.3. The head number is 4.

The experiments are on an NVIDIA GTX2070 GPU with 8GB memory. The model is trained through Adam optimizer [9], and learning rate is 0.01.

The data used in this article comes from the open and free public dataset in Tushare (https://www.tushare.pro/) for the research of stock market in China, which has the characteristics of rich data, simple use, and convenient implementation. It is very convenient to obtain the basic market data of stocks by calling the API.

The implementation details of this paper can refer to source code of this paper at https://github.com/zshicode/Attention-CLX-stock-prediction. We conduct empirical study on the stock price of Back of China (601988.SH) in Chinese stock market. The data is downloaded from Tushare(www.tushare.pro). The stock price data on Tushare is with public availability. The data is selected from the data from January 1, 2007 to March 31, 2022, the data in one day denotes a point of the sequence. The train set and test set was divided on June 22, 2021, as shown on Fig. 7. This means that there are 3500 training samples and 180 testing samples. The batch size is 32.

A. Prediction performance

The stock price prediction result of ARIMA model is shown on Fig. 8. The resudial and residual density plot are shown on Fig. 9.

The stock price prediction result of ARIMA+XGBoost model is shown on Fig. 10. The resudial and residual density plot are shown on Fig. 11.

The loss curves of original sequence and residual sequence of ARIMA+SingleLSTM are shown on Fig. 12 and Fig. 13.