STOCK PRICE PREDICTION MODEL BASED ON DUAL ATTENTION AND TCN

Yifeng Fu and He Xiao

Department of Software Engineering,
Jiangxi University of Science and Technology, Nanchang, China

ABSTRACT

The stock market is affected by many variables and factors, and the current forecasting models for time series are often difficult to capture the complex laws among multiple factors. Aiming at this problem, a stock price prediction model based on dual attention mechanism and temporal convolutional network is proposed. First, a convolution network more suitable for time series is used as the feature extraction layer. Feature attention is introduced to dynamically mine the potential correlation between the input factor features and closing prices. Second, based on Gated Recurrent Unit, on the other hand, a temporal attention mechanism is introduced to improve the model's ability to learn important time points and obtain importance measures from a temporal perspective. The experimental results show that the proposed model performs better than the traditional prediction model in the error index of stock price prediction and realizes the interpretability of the model in terms of index characteristics and time.

KEYWORDS

Time convolutional network, GRU, Temporary attention, Feature attention, Interpretability .

1. INTRODUCTION

Stock price forecasting refers to analyzing stock-related data and making predictions about the subsequent trend or fluctuation of the stock. The forecast results of stock prices can play a certain reference and guidance role for investors. However, because the time series data of the stock market have the characteristics of non-linearity, non-stationarity and high complexity, the prediction effect is not ideal.

We present the history of the development and some research results in the field of stock forecasting in Section 2, followed by the model structure and its definition in Section 3, the process of constructing the model, the experimental results and the analysis of the results in Sections 4 and 5. Finally, the experimental results are summarized in Section 6.

2. RELATED WORK

In the field of studying stock trends, early approaches mainly used machine learning algorithmic models to analyze based on numerical values. Wen Fenghua et al [1] proposed a method using a combination of SSA and SVM for prediction. They use singular spectrum analysis to decompose stock prices into trends, market fluctuations, and noise with different economic characteristics over different time horizons, and then introduce these characteristics into a support vector machine for price prediction [2].

David C. Wyld et al. (Eds): SIGEM, MLTEC, SEAPP, ITCON, NATL, FUZZY, CSEA - 2022
pp. 99-114, 2022. CS & IT - CSCP 2022 DOI: 10.5121/csit.2022.122007

It can be found that in early research, deep learning is not commonly used in the field of stock prediction, but due to its powerful learning ability, more and more researchers find it very suitable for learning complex data such as stock prediction, traffic prediction, and pedestrian flow prediction. Khaled Althelaya evaluates and compares LSTM deep learning architectures for short and long term forecasting of financial time series. They considered bidirectional and stacked LSTM prediction models in their experiments and benchmarked them with shallow neural networks and simple forms of LSTM networks [3][4].

As the study progressed, researchers found that most research methods using single-factor data had poor predictive effects because the stock trading data itself had limited information [5][6]. Jian Wang used the moving average convergence/divergence and the improved historical volatility index as the evaluation indicators for buying or selling, and constructed a forecast based on the improved MACD indicator[7].The results show that the model based on the index after the introduction of the panic index, the buy and sell operations made more profit[8][9].

Among the existing methods, the traditional RNN suffers from gradient explosion, and CNN is also considered not suitable for time series processing. To address this issue, this paper proposes a combined model based on Feature Attention TCN and Temporal Attention GRU.

The main features of the model are as follows.

1. In terms of data processing, a multi-source data fusion approach is used to fuse the numerical features of stock prices and the features of stock news, considering the influence of public opinion on stock prediction.
2. Since the neurons in RNN are invariant to replication, it brings the problem of gradient disappearance in training. We try to use CNN to process time series, but CNN is not suitable for time series learning. Therefore, this paper differs from the traditional use of RNN to process time series data by using a one-dimensional temporal convolution network to process the extracted features[10].
3. Based on TCN and GRU, a dual focus mechanism of features and time series is introduced to capture potential connections between different stock indicators and prediction targets for the relevant indicators of stock series. In terms of time nodes, the expression capability of output targets at key time points is enhanced.

3. MODEL DEFINITION

3.1. TCN Model

In general, CNNs are not considered suitable for solving time series problems. However, TCN, as a special kind of convolutional neural network, is more suitable for processing time series data. In this paper [11], the authors compare the performance of LSTM, GRU, RNN, and TCN using various sequence modeling tasks. In the sequence task on the MNIST dataset, TCN achieves an accuracy of 99%, at least 3 percentage points higher than the other models. There are three reasons for this: the existence of causality in the convolution in this architecture means that there is no information leakage from future to the past. Secondly, the convolutional architecture can map arbitrary length sequences to fixed length sequences. In addition, it uses residual modules and dilated convolution to construct long-term dependencies. In terms of performance comparison, TCN can parallelize the time series as vectors, which is faster than the point-by-point sequential computation of RNN. In addition, TCN can extend the input into a one-dimensional sequence, thus avoiding the need for time-point-by-time alignment of features [12].

The network structure of TCN consists of three main parts: causal convolution, dilated convolution and residual module.

3.1.1. Causal Convolution

Causal convolution is a strictly time-constrained structure. Each hidden layer has the same length as the input layer, and is padded with zeros to ensure subsequent layers have the same length. For the value of the previous layer at time t, it only depends on the value of the next layer at time t and its previous value. The causal convolution structure is shown in Figure 1.

$$F(t) = \left(x_{(*d)}f\right) = \sum_{i=0}^{k-1} f(i) \cdot x_{(t-di)} \quad (1)$$

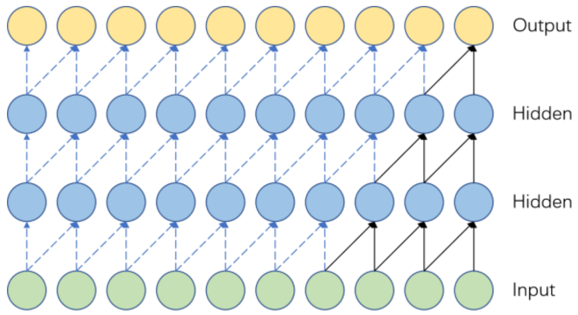

Figure 1. Causal Convolution

3.1.2. Dilated Convolution

There are still some problems with pure causal convolution, such as difficulty in capturing the dependencies between longer interval time points. Dilated convolution allows spaced sampling of the convolution time point input, and the sampling rate is controlled by the parameter Dilate [13]. The higher the level, the larger the Dilate used. Therefore, dilated convolution can expand the effective window, so that a larger receptive field can be obtained with a smaller number of layers.The dilate convolution structure is shown in Figure 2 and 3.

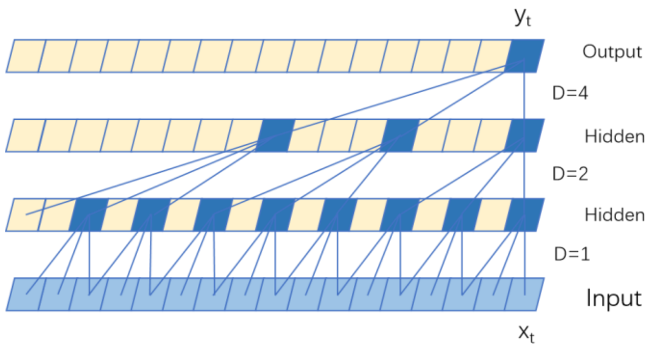

Figure 2. Dilated convolutional layer

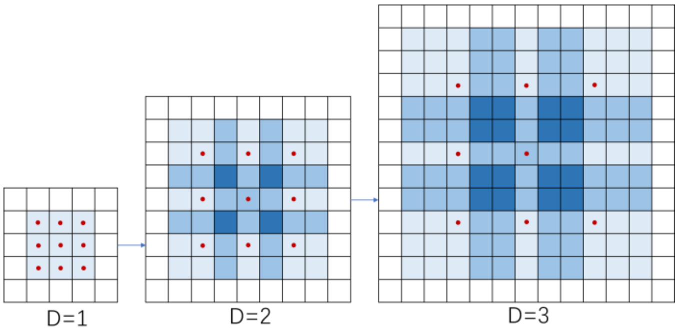

Figure 3. Feeling field for expanding convolution

3.1.3. Residual Module

The residual module of TCN consists of two layers of dilated convolution and ReLU function. The layer-hopping connection directly connects the feature maps of the lower layer to the upper layer[14]. The residual module ensures that each layer learns the relationship between mappings efficiently, which is very effective in networks with deeper layers[15]. As a result, the gradients of TCN are more stable and the problem of exploding or disappearing gradients can be effectively avoided. The structure of te residual block is shown in Figure 4.

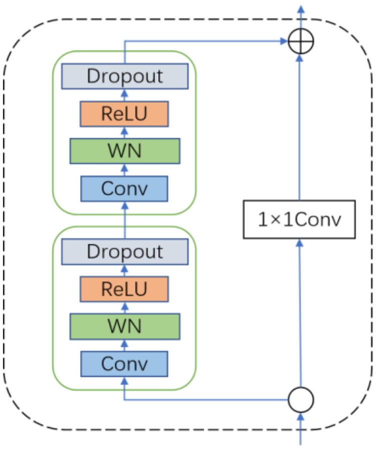

Figure 4. Residual Block

3.2. FATCN-TAGRU Model

By introducing the feature attention mechanism into the feature extraction process of the TCN model, the structure of feature attention is shown in Figure 5. Suppose the input time series is $X=[x_1,x_2,\cdots,x_T]=\left[x^{(1)},x^{(2)},x^{(3)},\cdots,x^{(6)}\right]^T$, The expansion can be represented as the following matrix:

$$X=\begin{bmatrix} x_1^{(1)} & x_1^{(2)} & \cdots & x_1^{(6)} \\ x_2^{(1)} & x_2^{(2)} & \cdots & x_2^{(6)} \\ \vdots & \vdots & & \vdots \\ x_T^{(1)} & x_T^{(2)} & \cdots & x_T^{(6)} \end{bmatrix} \in R^{T\times 6} \tag{2}$$

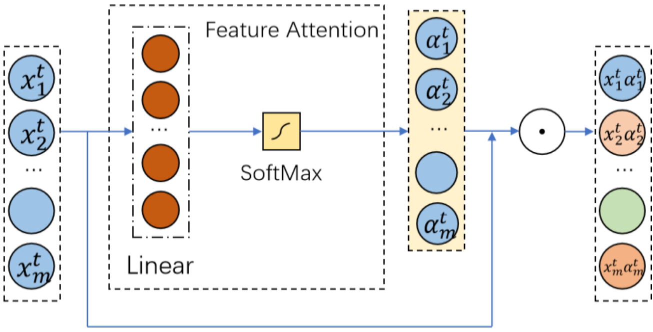

Figure 5. Feature Attention Model

In formula 2, $x_t = \left[x_t^{(1)}, x_t^{(2)}, \cdots, x_t^{(6)}\right](1 \le t \le T)$ is the set of 6 features at time t. $x^{(m)} = \left[x_1^{(m)}, x_2^{(m)}, \cdots, x_T^{(m)}\right](1 \le m \le 6)$ is the value of the m-th stock price-related variable at time t. Using a single-layer neural network to calculate the attention weight vector e_t, The formula is as follows:

$$e_t = tanh\left(W_e x_t + b_e\right) \qquad (3)$$

In formula 3, $e_t = \left[e_{1,t}, e_{2,t}, \cdots, e_{M,t}\right]$ is the attention weight coefficient combination corresponding to each input feature at the current time t; W_e is the training weight matrix; b_e is the bias vector for calculating the feature attention weight [16]. Since feature attention is located in the shallow layer of the model, and the input feature data is usually concentrated in a certain numerical range[17].

The structure of the time series attention mechanism is shown in Figure 6. The input $h_t = \left[h_{1,t}, h_{2,t}, \cdots, h_{k,t}\right]$ is the hidden layer state of the GRU network from the model iteration to time t, where k is the length of the input sequence time window. The time series attention weight vector l_t corresponding to each historical moment at the current time t is:

$$l_t = tanh\left(W_d h_t + b_d\right) \qquad (4)$$

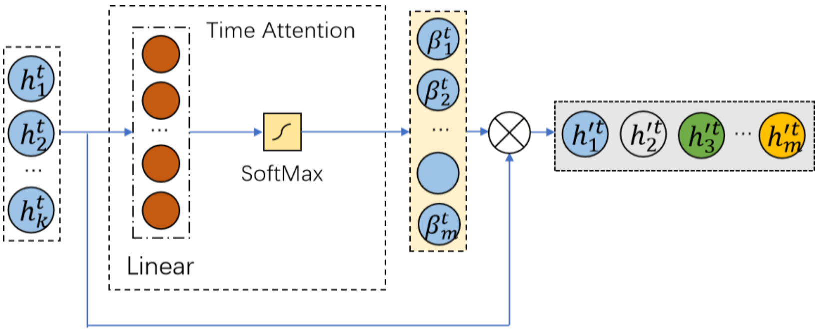

Figure 6. Time Attention Model

In the formula 4; W_d is the trainable weight matrix; b_d is the bias vector for calculating the temporal attention weight. The time attention weight $\beta_t = [\beta_{1,t}, \beta_{2,t}, \cdots, \beta_{\tau,t}, \cdots, \beta_{k,t}]$ is obtained, where $\beta_{\tau,t}$ is the attention weight at the τ th moment, which is weighted with the hidden layer state of each corresponding historical moment to obtain a comprehensive timing state $h_t^{'}$.

$$\beta_{\tau,t} = \frac{\exp(l_{\tau,t})}{\sum_{j=1}^{k} l_{j,t}} \quad (5)$$

$$h_t^{'} = \beta_t \otimes h_t = \sum_{\tau=1}^{k} \beta_{\tau,t} h_{\tau,t} \quad (6)$$

The FATCN-ATGRU network model is constructed by introducing the above feature attention and time series attention into the TCN layer and the GRU layer, respectively. The structure is shown in Figure 7. First, build the TAGRU network structure [18]. Compared with the TCN module, because the normalization part is already included in the attention mechanism, the Weight norm layer and the ReLU layer in the residual block are deleted. Use FATCN to mine the potential relationship between the input features and obtain the weighted input sequence; then extract the hidden time sequence correlation information from the weighted input feature sequence at the ATGRU layer, and mine the relevant feature time sequence information and current time data through the time sequence attention layer [19]. And assign time attention weights to it to enhance the expressive ability of key historical moment information, obtain weighted comprehensive time series information status, and finally send it to the fully connected layer to output future closing price predictions [20].

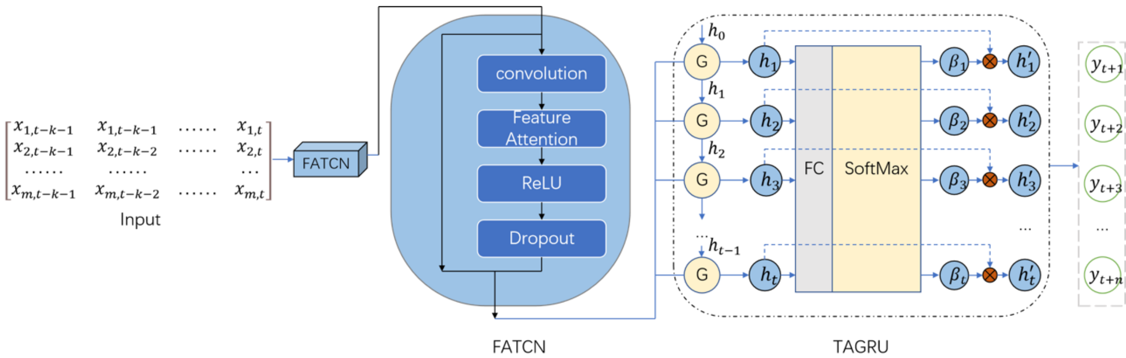

Figure 7. Model Structure

4. Experimental Modeling and Process

4.1. Data Collection and Pre-Processing

The stock data was downloaded from Yahoo Finance website. The time span is from June 26, 2017 to December 31, 2021. Stock market data for Apple, Google, Tesla, and Amazon are included. News text data was obtained by crawling the data of individual stock research reports from Eastern Fortune Website.

The downloaded stock numerical data includes six characteristic attributes: opening price, closing price, reweighted closing price, high price, low price, and trading volume. In this paper, closing price is used as the prediction label. News text features are obtained by processing research reports. Sentiment analysis is performed on the text data by using sentiment analysis API, and the processed sentiment indices are fused with the corresponding stock data as sentiment indices. The final fused input data has 7 feature attributes.

Data pre-processing consists of two specific steps.

1. Data cleaning: check the missing values and clean off the missing values, such as a day's data is missing a certain indicator, the data of that day will be deleted.
2. Normalization processing: This paper uses the MinMaxScaler normalization method, which is a linear transformation of the original data to eliminate the effects of differences between large and small units. The method deflates each feature to a given range, and after normalization, the indicators are in the same order of magnitude, which facilitates comprehensive comparison.

4.2. Model Parameter Settings

The initial time step size set in the experiment is 7, the learning rate is 0.01, and the parameters are updated with Adam optimizer and MSE loss function. Based on many previous studies and my replication and validation of experiments in the relevant literature, RNNs work best for time series tasks when the number of hidden layers is set to 2, and the number of neurons in each layer is 64. In addition, the effects of the number of residual blocks and the number of model iteration cycles in the TCN that need to be compared and analysed[21].

Table 1. Experimental results for some parameters Comparison.

Blocks	**Epoch**	**R2**	**MAPE**
2	500	0.981	0.0322
2	1000	0.974	0.0337
2	2000	0.983	0.0330
3	500	0.950	0.0338
3	2000	0.978	0.0378
4	500	0.974	0.0379
5	500	0.970	0.0382

It can be seen that when the residual block is set to 2 and the number of iterations is set to 500 or 2000, the prediction of the model is better. But when the number of iterations is set to 2000, the processing efficiency of the model is too low. So we choose 2 and 500.

5. Experimental Results and Analysis

5.1. Evaluation Indicators

The evaluation indicators use RMSE, MAE, R2 Score, and MAPE. where N is the total number of samples, and the parameter i represents the i-th sample. y_i is the predicted value, y_i' is the actual value. Calculated as follows:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - y_i')^2} \tag{7}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - y_i^{'}\right| \tag{8}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - y_i^{'})^2}{\sum_{i=1}^{N}(y_i - \overline{y})^2} \tag{9}$$

$$MAPE = \frac{100}{N}\sum_{i=1}^{N}\left|\frac{(y_i - y_i^{'})}{y_i}\right| \tag{10}$$

5.2. Comparative Analysis of Different Models

5.2.1. Single Model Comparison

In order to better compare and analyse the effect of the model, the traditional RNN and its variants and combined models are used for test comparison. The single model prediction effect is shown in Table 2 and Figure 8,9,10,11.

Table 2. Single Model Scoring Comparison.

Model	**RMSE**	**MAE**	**R2**	**MAPE**
RNN	0.0614	0.0380	0.8375	0.0598
LSTM	0.0484	0.0306	0.9161	0.0516
GRU	0.0479	0.0304	0.9141	0.0477
TAGRU	0.0365	0.0212	0.0934	0.0338

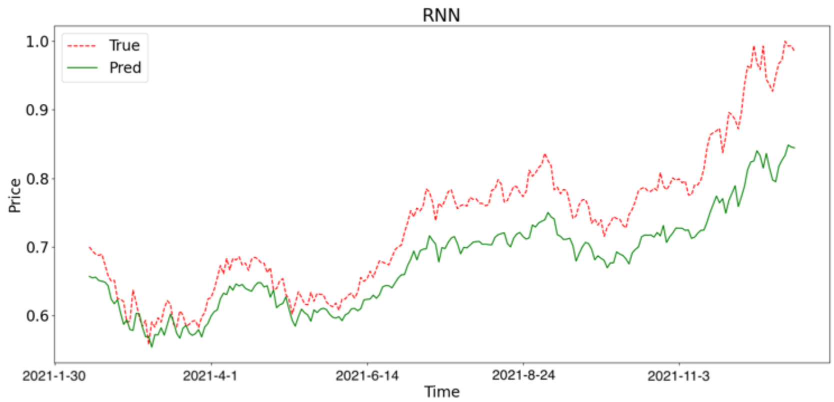

Figure 8. Prediction results of the RNN model

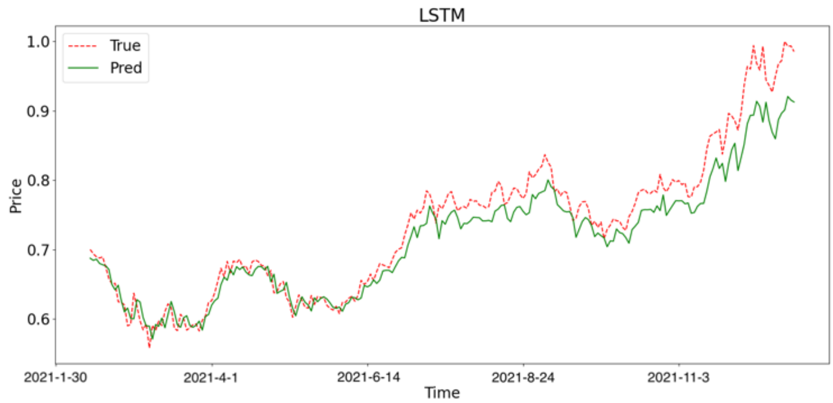

Figure 9. Prediction results of the LSTM model

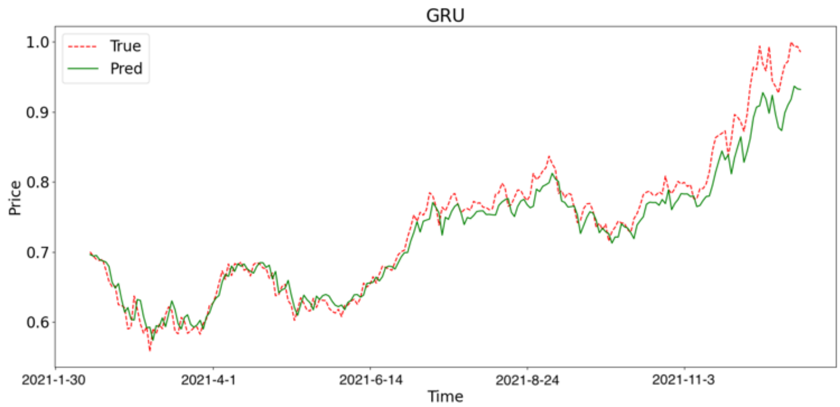

Figure 10. Prediction results of the GRU model

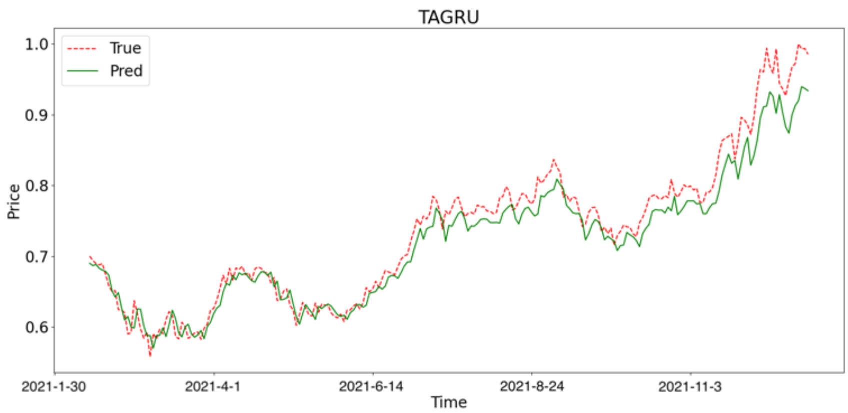

Figure 11. Prediction results of the TAGRU model

It can be found that the experimental results without using TCN network are poor, the fitting degree does not exceed 94 precent, and the predicted results have a certain lag phenomenon. The reason is that a model with a gated structure has the ability to retain memory to capture long-term relationships, and has a mechanism to reduce vanishing gradients, while a simple RNN erases and rewrites the entire memory at each update. Between LSTM and GRU, the gap between the two is small, and the fitting effect of GRU is only 0.2 precent better than that of the LSTM network. The reason is that GRU has the advantage of simple structure and less computational effort than LSTM. The above experimental results illustrate that the attention mechanism weights can help the model to better capture the complex patterns between different time points and help to deal with irregular time series.

5.2.2. Combination Model Comparison

The prediction results of introducing the FATCN module are shown in Figure 12,13,14,15. By comparison, the model with the introduction of FATCN performs better in each error index, indicating that the features processed by TCN have a better effect in prediction. By comparing the prediction effect with or without introducing attention, it can be found that the model predicts better by introducing the attention mechanism. The reason is that different time points and different features in the stock data have different effects on the predicted time points. The stock data closer to the prediction point has a greater impact on the prediction point.

Table 3. Combination Model Results Comparison.

Model	RMSE	MAE	R2	MAPE
FATCN-RNN	0.0479	0.0269	0.9496	0.0357
FATCN-LSTM	0.0361	0.0274	0.9584	0.0478
FATCN-GRU	0.0300	0.0196	0.9840	0.0318
FATCN-TAGRU	0.0180	0.0129	0.9908	0.0222

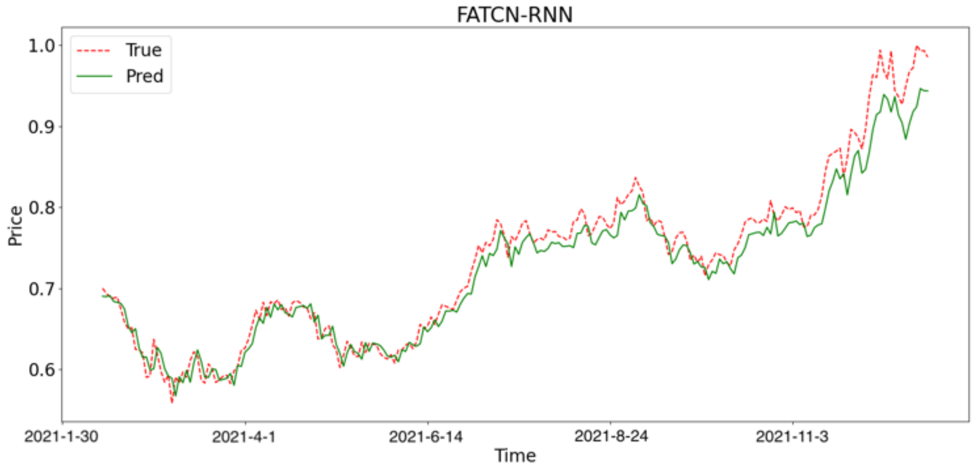

Figure 12. Prediction results of the FATCN-RNN model

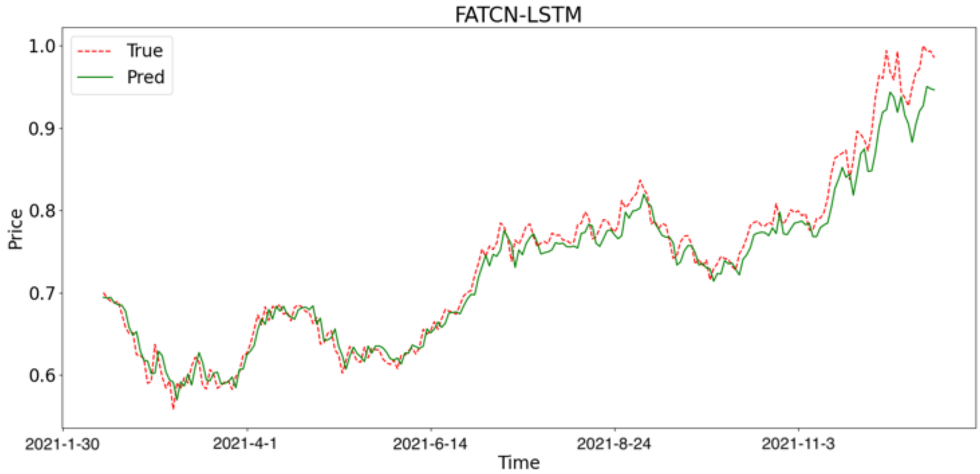

Figure 13. Prediction results of the FATCN-LSTM model

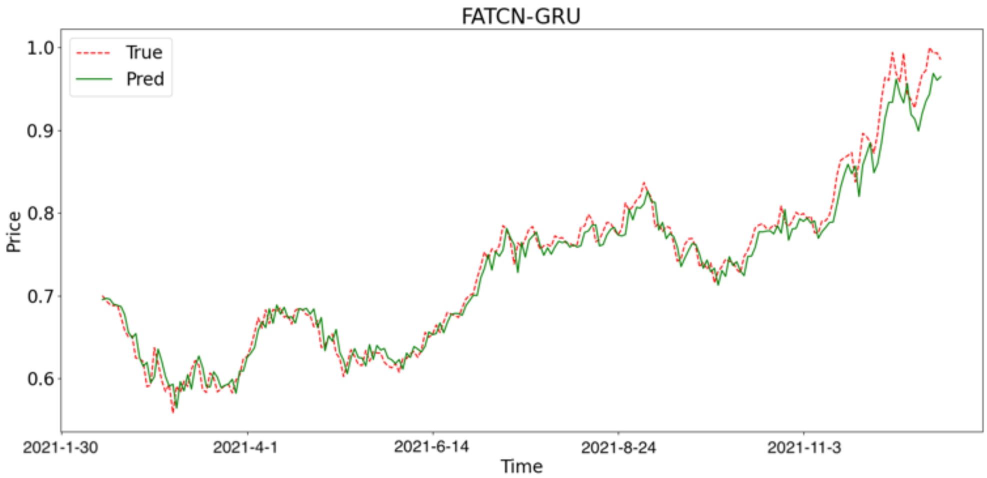

Figure 14. Prediction results of the FATCN-GRU model

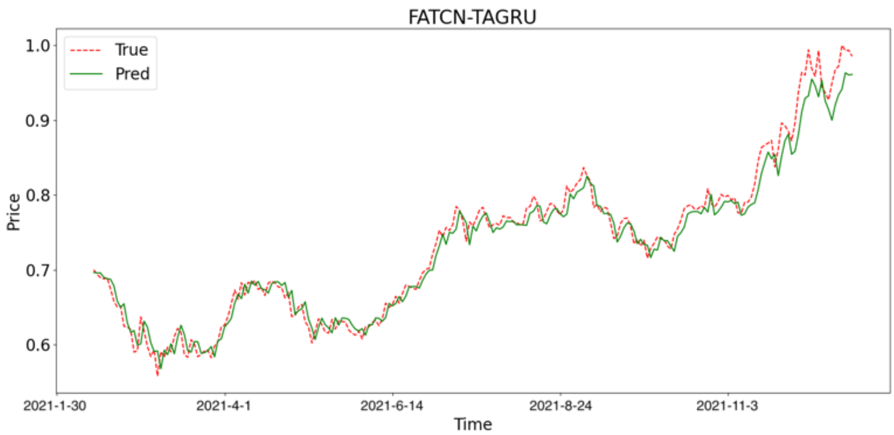

Figure 15. Prediction results of the FATCN-TAGRU model

5.3. Comparative Analysis of Different Stocks

In order to further examine the generalization ability of the model, three stocks of Apple, Tesla, and Amazon and their related news data are used to test the model.

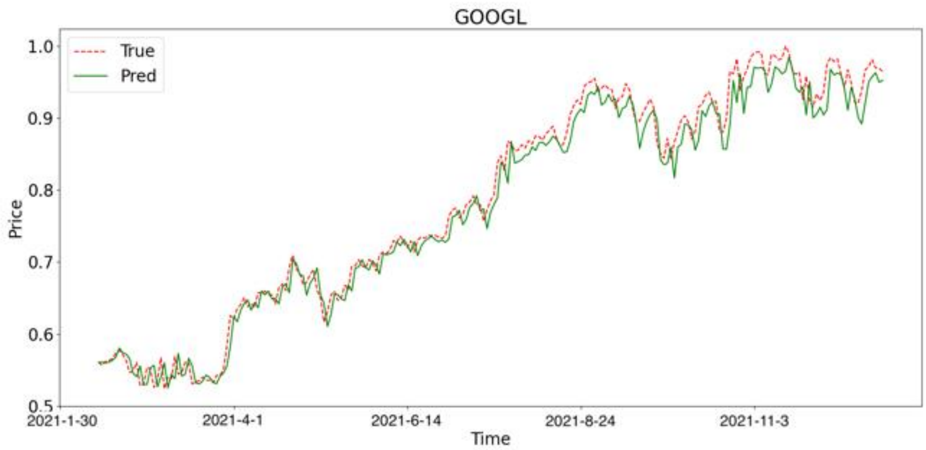

Figure 16. Results of an experiment using Google stock

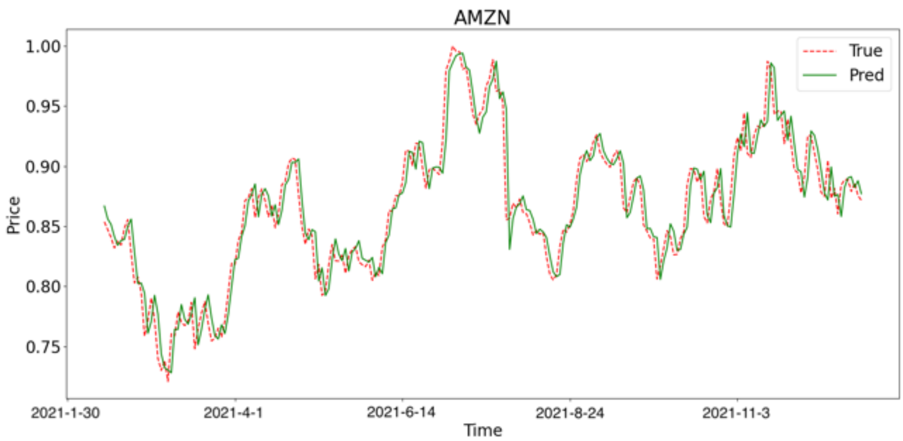

Figure 17. Results of an experiment using Amazon stock

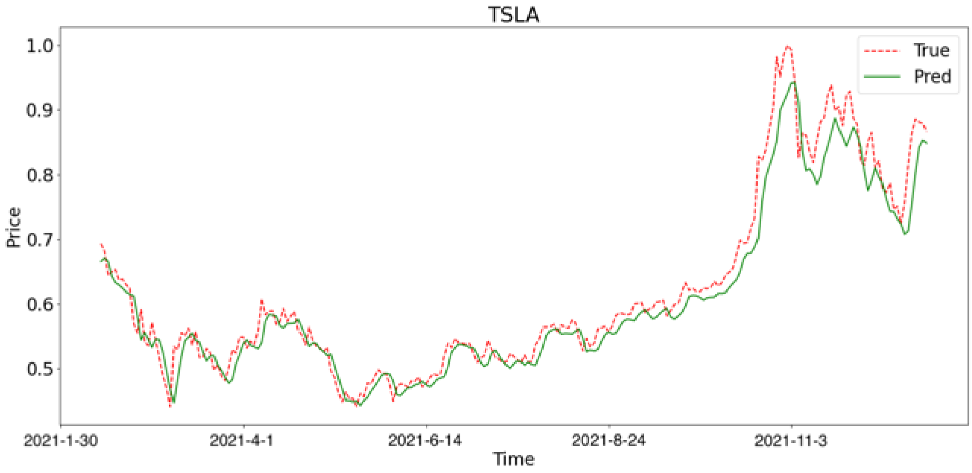

Figure 18. Results of an experiment using Tesla stock

It can be seen that after selecting the data of Apple, Amazon, and Tesla, the coefficient of determination of the model has reached more than 0.97. The three evaluation indicators also obtained good prediction results under other comparison models.

5.4. Interpretation Of Attention Weights

5.4.1. Feature Attention Weights Explained

Figure 19 shows a heatmap of the feature weights as the model converges. The vertical axis represents the characteristics of the stock data, and the horizontal axis represents the historical time point, in which time point 7 is the closest to the prediction time, and the depth of the color block represents the attention weight. In the convergence process, the weight of "closing price after resumption of rights" rises to 0.048 and the weight of "opening price" rises to 0.046, indicating that the closer to the prediction time point, the greater the contribution of the stock opening price and weighted closing price to the prediction result. The characteristics of "highest price", "minimum price" and "volume" have lower weights and have less influence on the prediction results. In the real stock trading market, the previous day's opening price and the reweighted closing price indicate the initial and final performance of the stock in the previous day's market, both of which directly affect the next day's movement [22]. The high and low prices represent only the two extreme points during the opening period and do not represent the final trend and performance of the stock. Therefore, the domain knowledge is consistent with the above experimental results and explains the rationality of the model from the perspective of feature importance.

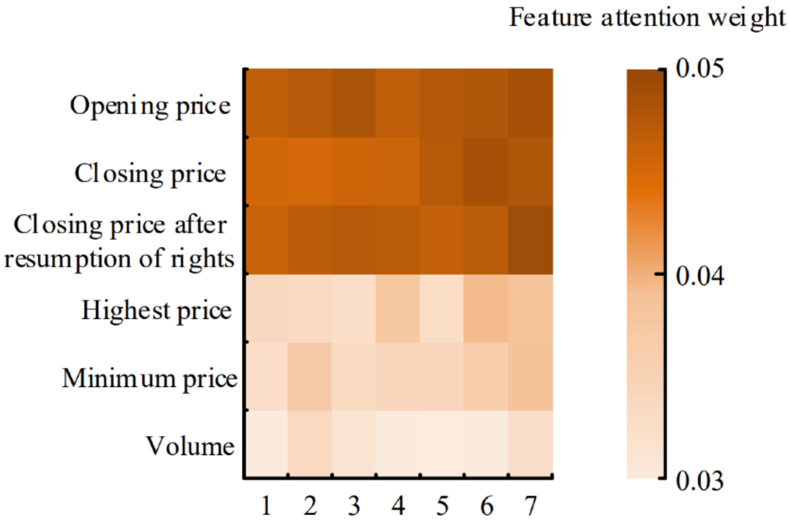

Figure 19. Feature Attention Weights

5.4.2. Time Attention Weights Explained

Figure 20 shows a sample of the temporal attention weight heat map. From the figure, it can be seen that the color blocks near the prediction time point are darker in color and have higher weights. It indicates that the model mainly focuses on the time steps closer to the prediction time. The analysis results in the above figure are consistent with the domain knowledge, which verifies the reasonableness of the prediction model in this paper from the perspective of temporal importance.

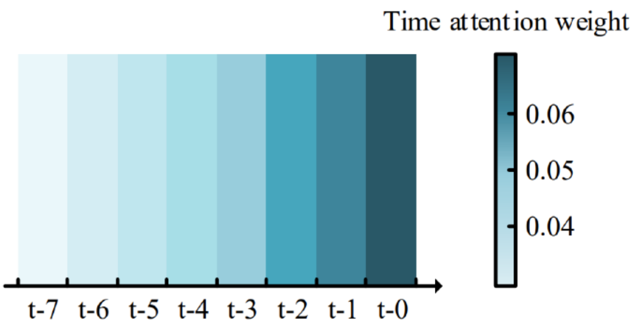

Figure 20. Time Attention Weights

6. CONCLUSION

In this work, we propose a TCN that introduces feature attention and a GRU framework that introduces time series attention. FATCN is responsible for processing time series data to extract deep features, and a feature attention mechanism is introduced to focus on important features, and then TAGRU is responsible for predicting frame. Finally, by comparing with other models and testing multiple stock data sets, better results are achieved. The results obtained by the model in this paper can provide a certain reference value for stock market investors on the microscopic level. On the macroscopic level, if risks can be predicted in advance, economic losses can be avoided in advance. In the future, we will try to introduce more sentiment indicators and test the effect of the experiment using a two-way structure in the prediction model.

ACKNOWLEDGEMENTS

This work was partially funded by the Science and Technology Project of Jiangxi Provincial Education Department, grant number GJJ200824.

REFERENCES

[1] Y. Lin, H. Guo, and J. Hu, "An svm-based approach for stock market trend pre-diction," in The 2013 international joint conference on neural networks (IJCNN), pp. 1–7, IEEE, 2013

[2] W. Fenghua, X. Jihong, H. Zhifang, and G. Xu, "Stock price prediction based on ssa and svm," Procedia Computer Science, vol. 31, pp. 625–631, 2014.

[3] K. A. Althelaya, E.-S. M. El-Alfy, and S. Mohammed, "Evaluation of bidirectional lstm for short-and long-term stock market prediction," in 2018 9th International Conference on Information and Communication Systems (ICICS), pp. 151–156, 2018.

[4] A. Sethia and P. Raut, "Application of lstm, gru and ica for stock price prediction," in Information and communication technology for intelligent systems, pp. 479–487, Springer, 2019.

[5] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," Journal of computational science, vol. 2, no. 1, pp. 1–8, 2011.

[6] T. H. Nguyen, K. Shirai, and J. Velcin, "Sentiment analysis on social media for stock movement prediction," Expert Systems with Applications, vol. 42, no. 24, pp. 9603–9611, 2015.

[7] J. Wang and J. Kim, "Predicting stock price trend using macd optimized by historical volatility," Mathematical Problems in Engineering, vol. 2018, 2018.

[8] Y. Xu and S. B. Cohen, "Stock movement prediction from tweets and historical prices," in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1970–1979, 2018.

[9] H. Kutlu and E. Avcı, "A novel method for classifying liver and brain tumors using convolutional neural networks, discrete wavelet transform and long short-term memory networks," Sensors, vol. 19, no. 9, p. 1992, 2019.

[10] H. Liu and Z. Long, "An improved deep learning model for predicting stock market price time series," Digital Signal Processing, vol. 102, p. 102741, 2020.

[11] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," arXiv preprint arXiv:1803.01271, 2018.

[12] L. Gong, M. Yu, S. Jiang, V. Cutsuridis, and S. Pearson, "Deep learning based prediction on greenhouse crop yield combined tcn and rnn," Sensors, vol. 21, no. 13, p. 4537, 2021.

[13] C. Yang, J. Zhai, and G. Tao, "Deep learning for price movement prediction using convolutional neural network and long short-term memory," Mathematical Problems in Engineering, vol. 2020, 2020.

[14] H. Wang, J. Wang, L. Cao, Y. Li, Q. Sun, and J. Wang, "A stock closing price prediction model based on cnn-bislstm," Complexity, vol. 2021, 2021.

[15] F. Xu and S. Tan, "Deep learning with multiple scale attention and direction regularization for asset price prediction," Expert Systems with Applications, vol. 186, p. 115796, 2021.

[16] J. Fan, K. Zhang, Y. Huang, Y. Zhu, and B. Chen, "Parallel spatio-temporal attention-based tcn for multivariate time series prediction," Neural Computing and Applications, pp. 1–10, 2021.

[17] J. Qiu, B. Wang, and C. Zhou, "Forecasting stock prices with long-short term memory neural network based on attention mechanism," PloS one, vol. 15, no. 1, p. e0227222, 2020.

[18] P. Hewage, A. Behera, M. Trovati, E. Pereira, M. Ghahremani, F. Palmieri, and Y. Liu, "Temporal convolutional neural (tcn) network for an effective weather forecasting using time-series data from the local weather station," Soft Computing, vol. 24, no. 21, pp. 16453–16482, 2020.

[19] D. Li, C. Lin, W. Gao, Z. Chen, Z. Wang, and G. Liu, "Capsules tcn network for urban computing and intelligence in urban traffic prediction," Wireless Communications and Mobile Computing, vol. 2020, 2020.

[20] Y. Wang, J. Chen, X. Chen, X. Zeng, Y. Kong, S. Sun, Y. Guo, and Y. Liu, "Short-term load forecasting for industrial customers based on tcn-lightgbm," IEEE Transactions on Power Systems, vol. 36, no. 3, pp. 1984–1997, 2020.
[21] H. Li, Y. Shen, and Y. Zhu, "Stock price prediction using attention-based multi input lstm," in Asian conference on machine learning, pp. 454–469, PMLR, 2018.
[22] L.-C. Cheng, Y.-H. Huang, and M.-E. Wu, "Applied attention-based lstm neural networks in stock prediction," in 2018 IEEE International Conference on Big Data (Big Data), pp. 4716–4718, IEEE, 2018.

AUTHOR

Fu Yifeng (1997 -), male, from Nanchang City, Jiangxi Province, China, is studying for a master's degree in computer technology at Jiangxi University of Technology. His main research fields are in-depth learning and quantitative trading.

© 2022 By AIRCC Publishing Corporation. This article is published under the Creative Commons Attribution (CC BY) license.