Predicting Stock Price Using Convolutional Neural Network

Kaiye Li
Software Engineering, Jilin University, Changchun, Jilin, 130015, China
liky5514@mails.jlu.edu.cn

2022 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA) | 978-1-6654-9991-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICAICA54878.2022.9844607

***Abstract*—For academics, predicting stock price based on raw data attributes has been a major yet difficult job. Most studies of stock price prediction concentrate on individual stocks, overlooking the link between similar stocks throughout the whole stock market. This study presents a clustering approach for mining related stocks. C-HTM is an online learning algorithm that uses Hierarchical temporal memory (HTM) to learn trends from comparable stocks and then forecast them. C-HTM has superior forecast accuracy than HTM, which has not learnt similar stock patterns, and its performance in terms of short-term prediction is better than the baseline models.**

Keywords—Stock Prediction, Machine Learning, Convolutional Neural Network

I. INTRODUCTION

To achieve a better accuracy, various methods be used by many scholars because of the stock data's real-time, high-noise and nonlinear. Traditional statistical approaches, such as the Autoregressive Integrated Moving Average (ARIMA) [1], a time-series prediction model that uses differences, show less promise than machine learning techniques [2]. Another difficulty with certain past studies is that their models are prone to overfitting or underfitting, requiring regular parameter adjustments.

Because of its nonlinear modeling capability, artificial neural networks (ANN) have become a useful tool for evaluating and predicting time series, which could assist anticipate stock values reliably and effectively. Traditional artificial neural networks (ANNs) lack the ability to simulate long-term time series dependencies, prompting the development of the Long Short-Term Memory (LSTM) network [3], a gated memory cell. Through its recurrent nature, LSTM collects time dependencies of the before-after associated data, making it an effective tool for time series prediction.

Empirical mode decomposition (EMD) is a method for processing nonlinear and non-stationary time series that decomposes the original data into a finite number of intrinsic mode functions (IMFs) and a residue that contains frequency information.

Convolutional neural networks (CNNs) [4] are one of the most successful deep learning approaches for extracting features from input data using convolutional kernels. CNNs can obtain features under multiple frequencies by using kernels of different lengths after the original series has been divided into frequency components.

This research provides a clustering approach that combines morphological similarity distance (MSD) and k-means clustering for mining comparable stocks. The approach entails using a clustering algorithm based on k-means to locate related companies in the stock market, with morphological similarity distance (MSD) being used as a metric of similarity, denoted as K-MSD. The MSD has been shown to be better suited for assessing time series similarity. In addition, we discover patterns across comparable stocks using the Hierarchical Temporal Memory (HTM) model [5], a physiologically restricted theory of intelligence first reported in. The findings of the experiment reveal that, when compared to HTM that hasn't learnt comparable stock patterns, the HTM after clustering, which we name C-HTM, has a higher prediction accuracy. Although researchers have successfully applied deep learning methods such as CNN, RNN, and GNN, etc., to fulfill the forecasting tasks, there is still a lack of a review which provides a comprehensive summary of these latest works done in stock price prediction. Therefore, in this article, first a detailed introduction to the task-related background was given, including the subtasks of the task, and the datasets and variables it uses. Then, we introduce some of the latest neural network models adopted in each task, including the principles of the model. In addition, we review some of the latest work and demonstrate which problems and methods the task focuses on solving [6]. Finally, based on the analysis of the previous work, we put forward some possible challenges in the task. Our review can provide a good guide for beginners.

II. BACKGROUND

As we all know, the stock price in the financial market is a huge dynamic field, which is difficult to model and predict. It is precisely because of these characteristics that the financial field is called a particularly attractive field for analysis using deep learning method. For the complex and changeable financial market, practitioners' grasp of risk is fundamentally a prediction problem, and the in-depth learning method is mainly used for prediction tasks, so it is very suitable for financial market prediction.

In the conventional stock market, stock price and trading volume are the most common parameters. We can easily get the corresponding mathematical model and change law. However, in the process of stock forecasting, many information cannot be quantified, such as the number of employees, work enthusiasm, or company reputation. Due to these obvious or hidden factors, stock prediction and selection is a very subjective process. In addition, the predictive variables are highly correlated. When the number of prediction variables is close to the observation count or the prediction variables are highly correlated, the traditional prediction methods will collapse.

Insightful investors can take advantage of financial markets to acquire inexpensive assets and sell overpriced assets. Observing the market and determining the steps that

must be made to maximize earnings while minimizing risks is one strategy to take advantage of this circumstance. Quantitative analysis has mirrored the concept of utilizing mathematical models to anticipate all elements of financial markets. The fundamental idea of this area is to examine market time series using mathematical and statistical models. This allows us to make important forecasts about the market in general, such as asset volatility, trends, and true worth [7]. These mathematical models, on the other hand, rely on manual parameter tweaking, which reduces the accuracy of their predictions. Furthermore, asset price fluctuations in financial markets frequently exhibit irrational behavior because they are heavily influenced by human actions that mathematical models cannot represent.

Computational intelligence models that can learn the nonlinear connection between input and output values have steadily been deployed in recent decades. Researchers have been attempting to incorporate these models into financial forecasting in order to improve the accuracy of price predictions. Using large-scale and high-frequency time series derived from stock trading orders as input, a deep learning approach based on convolutional neural network CNN is utilized to forecast the stock price trend in this study.

A. Datasets

The experimental text selects the data set of the bidding document. The data set is manually constructed from the bidding document information crawled from 30 government procurement websites such as China government procurement network. The main data types are: previous closing price, opening price, highest price, lowest price, closing price, trading volume, transaction amount, rise and fall, and average price. After cleaning and screening the team data, 25435 data with complete structure were selected for experiment.

B. Metrics

The experiment uses the accuracy rate to evaluate the classification results. The accuracy rate is defined as:

$$acc = \frac{num_{cor}}{num_{all}} \tag{1}$$

num_{cor} represents the total number of correctly predicted texts in the sample classification results, and num_{all} represents the total number of samples in the data set. Evaluation index acc represents the percentage of correct sample categories predicted by the model in the data set. The higher the accuracy acc , the better the performance of the model.

III. Neural Network models

Ordinary neural networks and convolutional neural networks are extremely similar. They are made up of neurons with weights and bias constants that may be learned. Each neuron takes in information and does dot product computations. The score for each category is the output. Some calculating approaches used in traditional neural networks can still be used [8].

A. Input Layer

The number of units in the input layer is consistent with the size of the predictor variables. All index data of the stock market need to be preprocessed to ensure that the neural network can be trained normally.

B. Convolutional Layer

Each Convolution Layer of a Convolutional Neural Network is made up of numerous convolution units, each of which has its parameters adjusted using a back-propagation technique. The convolution process is used to extract various input information. Only low-level characteristics such as edges, lines, and angles may be extracted by the first Convolution Layer. Increasingly layers in a network allow iterative extraction of more complicated characteristics from low-level information.

The Convolutional Layer incorporates more flexible predictor correlation items by adding convolution operations between the input and output. Each neuron applies a non-linear activation function to activate the neuron. The features extracted through the convolution operation need to be converted into a two-dimensional or three-dimensional structure and input into the training model. Each feature represents a different information dimension of the sample data. Therefore, the Convolutional Neural Network can enhance the structural performance and extract different information when increasing the number of convolution kernels. As shown in figure 1.

1	1	1	1	1
-1	0	-3	0	1
2	1	1	-1	0
0	-1	1	2	1
1	2	1	1	1

$\otimes$

1	0	0
0	0	0
0	0	-1

$=$

0	-2	-1
2	2	4
-1	0	0

Figure 1 Schematic diagram of convolution operation process

The basic two-dimensional convolution operation is shown in the equation. Assuming that the sample matrix of a signal is x and the size of filter w is U×V, V output y is the convolution of signal sequence x and filter w. The same filter w can extract different information of signal samples. Here, it is assumed that the subscript (i,j) of the output y of the convolution starts from (U,V)

$$y_{i,j} = \sum_{u=1}^{U}\sum_{v=1}^{V} w_{u,v} x_{i-u+1,j-v+1} \tag{2}$$

Its general continuous form is:

$$y(n) = (f \otimes w)(n) = \int_{-\infty}^{\infty} f(x) w(n-x) dx \tag{3}$$

The discrete form is:

$$y(n) = (f \otimes w)(n) = \sum_{x=-\infty}^{\infty} f(x) w(n-x) \tag{4}$$

C. Pooling Layer

It is, in fact, a type of downsampling. Nonlinear pooling functions come in a variety of shapes and sizes, with maximum pooling and average sampling being the most popular. Typically, a feature with a big dimension is acquired after the convolutional layer, the feature is sliced into many parts, and the maximum or average value is used to create a new feature with a lower dimension.

The number of features in the Convolutional Layer is lowered, while the number of neurons is almost maintained. To lower the feature dimension and avoid over-fitting, the pooling procedure must be conducted at the pooling layer.

D. Fully-connected Layer

The Fully-connected Layer aggregates all weighted signals into the final prediction, maps the multi-dimensional vector to a one-dimensional vector, integrates all the learned features, and maps the integrated features to the output space to give the classification result. Generally speaking, Convolutional Neural Network bureaus will have two Fully-connected Layers. The function of the first layer is to map the weighted signals obtained by the previous Convolutional Layer and Pooling Layer to the sample label space. The second Fully-connected Layer uses a certain activation function to get the final probability of classification prediction.

IV. Related Researches

For a long time, computational intelligence has been applied in financial applications. Fundamentals and technical analysis are the two primary divisions. These technologies are intended to forecast the stock market's trend or future price. Although stock price data appears to exhibit random movements, there are certain principles. It is possible to forecast future stock price rises and falls by evaluating stock price trends. Previous studies have tended to integrate stock price fluctuations with fundamental analytical indicators to forecast future stock prices. It may be used as a reference for investors by analyzing historical data, such as opening and closing prices, stock market volatility, and other data. As a result, this analytical method may be used to forecast data. Despite the drawbacks of classic linear regression and other basic analytical mathematical models, they are nonetheless widely used due to their simplicity. Because the model produced by this technique is easy to explain, more investors and academics embrace it. Linear regression and other traditional models, on the other hand, are unsuitable for complex and dynamic financial areas. As a consequence, it is unsuitable for financial market forecasting.

The advancement of machine learning and deep learning technologies has gotten a lot of attention in this area. Machine learning and deep learning have three properties that make them ideal for dealing with ambiguous functional forms. The first is that it is diverse. It may put a wide variety of networks in conventional search as a collection of distinct ways. Second, machine learning and deep learning can approximate complicated nonlinear relationships by deliberate design. Third, to minimize overfitting bias and error discovery, parameter penalty and conservative model selection criteria complement the variety of functional forms provided by these approaches.

Company disclosures, on the other hand, are more trustworthy resources since they usually include the latest officially confirmed information. They usually report information regarding quarterly earnings, adjustment of the board of the directors and challenges of the business. Particularly, textual data from company announcements was extracted in literature [9] to understand their impact on both short-term and long-term stock index predictions. Moreover, based on instant announcement from the website of Thomson and Reuters, literature [10] construct a decision supporting model. They also set up sentiment analysis that focuses on negation scope detection.

V. Challenges

The classification approach based on deep learning offers enough benefits in feature extraction over the classic machine learning classification algorithm, but it also has certain disadvantages. RNN, CNN, transformer, and other popular deep learning algorithms are only a few examples. RNN represented by LSTM / Gru and its deformation, for example, can not only successfully capture long-distance text characteristics, but also handle linear sequence text data with varying length. However, while computing the output result of the current time, RNN must consider the hidden state of the previous time, which makes parallel operation and slow operation speed problematic for RNN. As presented above, previous literature extensively studies the machine learning methods on financial market prediction and propose variety of innovative models. Unfortunately, there are currently two main limitations existing in predicting stock price using machine learning algorithms. The first issue involves with the fact that although the textual features are employed in current models to better interpret the mainstream attitude towards a certain stock in social media, they usually collect information based on previous data mining tools (text mining technologies). The conventional text mining strategies usually fail to account for the semantic and other helpful resources that also improve the performance of machine learning models. Additionally, when determining the percentage of text or financial features to be adopted during data construction, the action to reduce dimensionality in feature sets is a of great necessity. Previous prediction models were mostly based on principal component analysis (PCA) and latent Dirichlet allocation (LDA) to reduce the feature dimensionality. Admittedly, they are both powerful tools for reducing dimensionally. However, both methods are somewhat problematic -- PCA method is prone to lose information and fails to process nonlinear data, while the LDA method is unable to evaluate semantic information in social media. Since both mechanisms will damage some of the information during dimensionality reduction, these two methods are, in general, not tailored for the stock price prediction. As a result, the challenge of employing machine learning models in real life financial market prediction is that better text mining tools and dimensionality reduction methods need to be invented. As a result, RNN research and

implementation in text feature extraction is restricted. CNN collects features automatically by combining one or more convolution layers, nonlinear layers, and pooling layers. Cross layer connection mechanisms may generally be employed to improve the network depth in the construction of CNN network structures in order to extract more effective features, although the performance of CNN in long-distance feature extraction is lower than that of RNN.

VI. Conclusion

The semi-supervised bidding document classification method based on graph convolution neural network proposed in this paper creates a knowledge map, connects multiple data, clarifies the structure of unstructured text, and uses external information to supplement the features of classification nodes to compensate for the knowledge map's lack of semantic information. Furthermore, the model's node feature extraction criteria are insufficiently stringent. A graph convolution neural network may be used to aggregate neighbor node information for the central node, allowing for greater characterization and feature information for the nodes to be categorized. Simultaneously, just a few labeled samples are required to provide a huge quantity of data into the model for training. The model effectively increases classification accuracy, according to the testing data. This publication will continue to research graph neural networks and incorporate attention mechanisms and relational weight to increase model classification accuracy in the future.

References

[1] Jung-Hua W and Jia-Yann L 1996 Stock market trend prediction using ARIMA-based neural networks Proceedings of International Conference on Neural Networks (ICNN'96) 4 Washington, DC, USA pp. 2160-65.

[2] Alwadi S, Almasarweh M and Alsaraireh A 2018 Predicting Closed Price Time Series Data Using ARIMA Model. Modern Applied Science. 12. Simposium Kebangsaan Sains Matematik ke-28 (SKSM28) Journal of Physics: Conference Series 1988 (2021) 012041 IOP Publishing doi:10.1088/1742-6596/1988/1/012041 11

[3] Nan Jing, Zhao Wu, Hefei Wang, A hybrid model integrating deep learning with investor sentiment analysis for stock price prediction, Expert Systems with Applications, Volume 178, 2021, 115019, ISSN 0957-4174.

[4] K. Cho, B. Van Merriënboer, C. Gulcehre et al., "Learning phrase representations using RNN encoder-decoder for statistical machine translation," 2014, arXiv preprint arXiv:1406.1078.

[5] Chung J, Gulcehre C, Cho K, Bengio Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv:14123555. 2014.