საქართველოს მეცნიერებათა ეროვნული აკადემიის მოამბე, ტ. 18, №3, 2024
BULLETIN OF THE GEORGIAN NATIONAL ACADEMY OF SCIENCES, vol. 18, no. 3, 2024

Informatics

Machine Learning in Financial Data Analysis and Forecasting

Vladislav Dashtu*, Nick Inassaridze**

* *School of Science and Technology, University of Georgia, Tbilisi, Georgia*
** *A. Razmadze Mathematical Institute, Ivane Javakhishvili Tbilisi State University; Georgian Technical University; Tbilisi Centre for Mathematical Sciences, Tbilisi, Georgia*

(Presented by Academy Member Hvedri Inassaridze)

In the modern world, data analysis and prediction have become extremely relevant due to the large amount of information humanity has to process every day. Machine learning algorithms are identified as one of the most promising approaches. This paper presents a neural network-based approach that provides the possibility of predicting the future price dynamics based on the analysis of historical data of the company's share price, as well as the published opinions about the company. This work considers only the processing of historical data of stock price of companies, but the presented approach can potentially be generalized to any kind of time series. *© 2024 Bull. Georg. Natl. Acad. Sci.*

time series analysis, neural networks, LSTM, stock market

The purpose of this research is to demonstrate an algorithm based on neural networks for predicting the likely direction of a company's stock price. Financial data are mostly presented in the form of time series, which refers to observations made at equally spaced time intervals (e.g. daily), arranged chronologically [1, 2].

Neural networks are a modern machine learning approach used to solve a wide range of problems, including time series classification and prediction. This paper utilizes a special case of recurrent neural networks – Long Short-Term Memory (LSTM), which, unlike conventional recurrent networks, have the ability to store and process contextual information for a longer period of time [3]. In [4], neural networks were compared to the Box-Jenkins method and it was concluded that NNs outperform the Box-Jenkins model for series with short memory. For series with long memory, both methods produced similar results. Zhang et al. [4] gave a detailed review of neural networks for forecasting. They trialed an auto-regressive integrated moving average (ARIMA) model with an artificial neural network to predict time series. The results showed that the ANN was more advantageous in analyzing and processing nonlinear data. A recent review [5] discusses the applications of deep learning to time series data. The stock price prediction using LSTM [6] and other forms of deep learning [7, 8] continues to be a relevant research topic.

The rest of the research is structured as follows: Section 2 describes the essence, structure and sources of the data used in the research, and defines transformations carried out on the data. Section 3 provides a brief description of neural networks and proposes a strategy for tuning hyperparameters for the network and selecting the best model. Section 4 presents the evaluation of neural network prediction results – with the help of statistical metrics, and also by comparison with real data. In the conclusion section, the results of the research are summarised and the perspective of future work is discussed.

Data Structure

A time series is a set of observations made on a time-varying event, arranged chronologically [1]. In case of having more than one variable that depends on time, the time series is called multivariate. In stock market price time series, the following five variables are most commonly found:

1. Opening price (Open) – price value at the beginning of the time interval.
2. Maximum (High) – the maximum value of the price during the interval.
3. Minimum (Low) – the minimum value of the price during the interval.
4. Closing price (Close) –price value at the end of the interval.
5. Volume – the number of shares traded during the interval.

According to the book “Technical Analysis and Stock Market Profits” [9] by Richard Schabacker, the most important and informative of these is the closing price. In addition, studies have shown that the press read on the Internet really influences the decision made by investors [10, 11]. Accordingly, the news headlines published during each day were divided into three categories for the research – “positive”, “neutral” and “negative”.

The scope of the research is constrained to the daily price analysis of three American transnational corporations – Apple Inc. (AAPL), Adobe Inc. (ADBE) and Alphabet Inc. (GOOGL). The data was downloaded from the online platform Kaggle. Specifically, we downloaded a dataset of historical price records of companies listed in the S&P500 index, as well as two datasets of news (investing.com, 2019; benzinga.com, 2020). Subsequently, the times in the data sets were rounded down to the nearest day. The data excludes the days when the American stock exchange does not operate – these are weekends and American public holidays. In total, the range of the analyzed financial data was limited to the period from May 13, 2016 to May 14, 2020 (four years). The last 10 records of the time series (May 15-29) were plotted separately and used to estimate the accuracy of the forecast.

Predicting the Price Direction Using Neural Networks

The structure of neural networks can be summarized as a set of interconnected neurons. A single neuron has one or more inputs and one or more outputs. The structure of a neuron can be simple or complex (i.e. it implies a combination of many mathematical functions). Neurons are arranged and connected differently to each other. A specific structural configuration of the network is called a model, and the characteristic parameters of this configuration, which makes it unique – the number of layers, the number of neurons in one layer, the activation function, the optimizer, and others – are called hyperparameters. [12]

Recurrent neural networks are one of the best for time series analysis, because some of their neurons can connect to themselves or to neurons in the previous layer; accordingly, information moves in both directions and the network acquires the ability of “memory” [13]. The main problem of recurrent networks is the gradual “forgetting” of events that happened a long time ago, which is solved by a special sub-type - Long Short-Term Memory (LSTM). A single neuron in an LSTM network consists of four parts that, at each learning iteration, evaluate the relevance of the data at the current point in time. A transformer-type neural network called FinBERT [14] was used to evaluate the sentiment of news headlines. Transformer-type networks are one type of neural network that is a combination of an autoencoder-type network and an "attention" mechanism [14, 15].

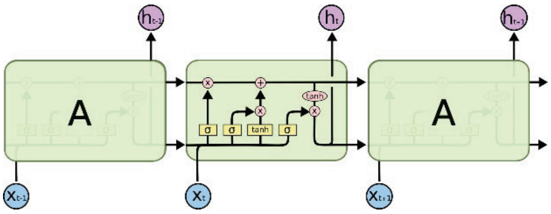

Fig. 1. The four layers of an LSTM cell. Source: [16].

In this paper, one input data item to the neural network is the sequence of the current day's OHLCV and the relative shares of positive, neutral, and negative news over the previous 14 days. In total, the input vector consists of 52 elements. The output is the delta of the current day's closing price compared to the previous day. The data points were randomly divided into training and test sets in an 80-20 ratio. Before entering the network, the data was scaled using the MinMaxScaler tool of the scikit-learn library. The mean squared error (MSE) was used as the loss function. The Adaptive Moment Estimation (Adam) algorithm proposed by Kingma and Ba in 2014 is used as an optimizer. [17] In contrast to gradient descent, Adam excels at handling "incomplete" data, where missing or zero values are often encountered [18]. The Adam optimizer parameters are fixed in the study as $\beta_1 = 0.9,\ \beta_2 = 0.999,\ \varepsilon = 10^{-8}$ respectively.

The rest of the network parameters were determined by trying several possible combinations – this approach is known as "grad student descent" [19], and its name derives from the name of the gradient descent algorithm. We tested the following parameter values:

1. Number of layers: [1,2,3,4].
2. Number of neurons: [16,32,64,128].
3. Batch size: [16,32,64,128].
4. Seed value: [1,2,3,4,5,6,7,8,9,10].

Accordingly, a total of 640 possible combinations of parameters were formed for each company. The accuracy of direction prediction on the test set was used as the network accuracy metric. The direction of the price movement is defined as the mathematical sign of the difference between the corresponding price values at two time points. Predicted directions are therefore signs of differences in the predicted values of two adjacent days. Prediction accuracy is the ratio between the number of correctly guessed directions and the total amount of guesses, which is one less than the amount of data points. By fixing the random generation seed, the difference between the accuracies obtained in the case of training twice with the same hyperparameters was minimized. After each training epoch, the prediction accuracy of the test set was calculated, and finally only the state of the network giving the best accuracy was kept. The maximum number of epochs is 50, although each network is trained to the best condition for its parameters. One combination of parameters was selected for each company. Other parameters being equal, the variant with smaller packet size and number of epochs is better [20]. Regarding the number of layers and the number of neurons per layer, a 2020 study by Yadav, Jha, and Sharan [21] on Indian companies compared the accuracy of an LSTM network with one to seven hidden layers using the RMSE metric. The conclusion of the study is as follows: LSTM with the number of layers equal to 1 has the best overall accuracy, although increasing the number of layers reduces the standard deviation of network predictions (makes it more stable). Considering the above circumstances, the optimal combinations of parameters were selected for each company participating in this study, presented in Table 1.

Table 1. The optimal parameters of LSTM neural networks

Stock	Layers	Neurons	Batch	Seed	Best Epoch	Accuracy
AAPL	1	16	32	9	13	60.804%
ADBE	4	32	16	4	34	64.322%
GOOGL	1	128	16	3	45	60.804%

Table 2. Evaluation of LSTM accuracy using different criteria

Stock	Test Acc.	Real Acc.	MAE	RMSE	R^2
AAPL (S)	60.804%	60%	0.807849	1.539519	-1.706376
ADBE (S)	63.819%	60%	3.538750	6.965875	-0.260702
GOOGL (S)	60.804%	40%	11.684246	16.326593	-0.023063

Results

One common way to evaluate neural network accuracy and compare the models is to calculate statistical metrics, including MAE (Mean Absolute Error), RMSE (Root-Square MSE), and the R^2 metric, whose formula is given by:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{\sum_{i=1}^{n}\left(Y_i - \overline{Y}_i\right)^2}.$$

The R^2 metric is used in regression problems and shows how close the predicted direction line is to the movement trend of the actual data. If the results predicted by a model have worse accuracy than the function that simply calculates the mean of the data, the R^2 value is negative. In addition to statistical metrics, the accuracy of “predicting” price direction over the next 10 days is evaluated, which is an example of real-world use of the network.

It can be seen from Table 2 that the real-world accuracy of the trained model is close to the accuracy of the test set in two out of three cases. Noteworthy are the negative values of the R^2 metric in the evaluation, which theoretically means that the movement predicted by the network misses the correct one more than the horizontal line drawn in the middle of the movement graph.

Conclusion

In this research, a neural network-based approach for predicting financial data is discussed. Sentiment estimates of three American technology giants' stock prices and recent news headlines are analysed. Price direction prediction accuracy for all three companies was 60% both on the test set and on real-world data. The mentioned percentage is considered enough to evaluate the price trends on the trade landscape and to choose the next action. Therefore, future work will consider developing a more efficient strategy for selecting network parameters, as well as trying other types of “supporting data” in network training besides news is a matter of future work.

ინფორმატიკა

მანქანური სწავლება ფინანსური მონაცემების ანალიზისა და წინასწარმეტყველებისთვის

ვ. დაშტუ*, ნ. ინასარიძე**

** საქართველოს უნივერსიტეტი, მეცნიერებისა და ტექნოლოგიების სკოლა, თბილისი, საქართველო*
*** ივანე ჯავახიშვილის სახ. თბილისის სახელმწიფო უნივერსიტეტი, ა. რაზმაძის მათემატიკის ინსტიტუტი; საქართველოს ტექნიკური უნივერსიტეტი; თბილისის მათემატიკის მეცნიერების ცენტრი, თბილისი, საქართველო*

(წარმოდგენილია აკადემიის წევრის ხ. ინასარიძის მიერ)

თანამედროვე სამყაროში მონაცემთა ანალიზი და პროგნოზირება ძალზე აქტუალური გახდა იმის გამო, რომ კაცობრიობამ ყოველდღიურად უნდა დაამუშაოს დიდი მოცულობის ინფორმაცია. ამ საკითხის დასაძლევად მანქანური სწავლების ალგორითმები გამოვლინდა, როგორც ერთ-ერთი ყველაზე პერსპექტიული მიდგომა. მოცემულ ნაშრომში წარმოდგენილია ნეირონულ ქსელზე დაფუძნებული მეთოდი, რომელიც იძლევა კომპანიის აქციების ფასის ისტორიული მონაცემების ანალიზისა და, ასევე, კომპანიის შესახებ გამოქვეყნებული მოსაზრებების საფუძველზე, მომავალი ფასების დინამიკის პროგნოზირების შესაძლებლობას. ნაშრომში განვიხილავთ მხოლოდ კომპანიების აქციების ფასის ისტორიული მონაცემების დამუშავებას, მაგრამ წარმოდგენილი მიდგომა პოტენციურად შეიძლება განზოგადდეს ნებისმიერი სახის დროით მწკრივზე.

REFERENCES

1. Profillidis V. A. & Botzoris G. N. (2019a) Trend projection and time series methods. Modeling of transport demand: 225–270. Elsevier.
2. Lazrieva N., Mania M., Mari G., Mosidze A., Toronjadze A., Toronjadze T. et al. (2000) Probability theory and mathematical statistics for economists. Tbilisi (in Georgian).
3. Petneházi G. (2019) Recurrent neural networks for time series forecasting. In arXiv [q-fin.ST]. https://doi.org/10.48550/ARXIV.1901.00069.
4. Zhang G. P. (2003) Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*, **50**: 159–175. https://doi.org/10.1016/s0925-2312(01)00702-0.
5. Gamboa J. C. B. (2017) Deep learning for time-series analysis. https://doi.org/10.48550/ARXIV.1701.01887.
6. Istiake Sunny M. A., Maswood M. M. S., & Alharbi A. G. (2020) Deep learning-based stock price prediction using LSTM and bi-directional LSTM model. (2020), 2nd Novel Intelligent and Leading Emerging Sciences Conference (NILES), 87–92.
7. Mehtab S. & Sen J. (2020) Stock price prediction using convolutional neural networks on a multivariate timeseries. In arXiv [q-fin.ST]. http://arxiv.org/abs/2001.09769.
8. Song Y., Lee J. W. & Lee J. (2019) A study on novel filtering and relationship between input-features and target-vectors in a deep learning model for stock price prediction. *Applied Intelligence*, **49**(3): 897–911. https://doi.org/10.1007/s10489-018-1308-x.
9. Schabacker R. (2021) Technical analysis and stock market profits, Harriman definitive edition. Harriman House Publishing.
10. Mao H., Counts S. & Bollen J. (2011) Predicting financial markets: Comparing survey, news, Twitter and search engine data. In arXiv [q-fin.ST]. http://arxiv.org/abs/1112.1051.
11. Kaminski J. (2014) Nowcasting the bitcoin market with Twitter signals. In arXiv [cs.SI]. http://arxiv.org/abs/1406.7577.
12. Nielsen M. A. (2019) Neural networks and deep learning. http://neuralnetworksanddeeplearning.com
13. Bengio Y., Simard P. & Frasconi P. (1994) Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, **5**(2): 157–166. https://doi.org/10.1109/72.279181.
14. Araci D. (2019) FinBERT: financial sentiment analysis with pre-trained language models. https://doi.org/10.48550/ARXIV.1908.10063.
15. Devlin J., Chang M.-W., Lee K. & Toutanova, K. (2018) BERT: Pre-training of deep bidirectional Transformers for language understanding. https://doi.org/10.48550/ARXIV.1810.04805
16. Olah C. (2015) Understanding LSTM networks. Colah's Blog. http://colah.github.io/posts/2015-08-Understanding-LSTMs/
17. Kingma D. P. & Ba J. (2014) Adam: a method for stochastic optimization. https://doi.org/10.48550/ARXIV.1412.6980.
18. Ruder S. (2016) An overview of gradient descent optimization algorithms. https://doi.org/10.48550/ARXIV.1609.04747.
19. Gencoglu O., van Gils M., Guldogan E., Morikawa C., Süzen M., Gruber M., Leinonen J. & Huttunen H. (2019) HARK side of deep learning -- from grad student descent to automated machine learning. In arXiv [cs.LG]. http://arxiv.org/abs/1904.07633.
20. Keskar N. S., Mudigere D., Nocedal J., Smelyanskiy M. & Tang, P. T. P. (2016) On large-batch training for deep learning: generalization gap and sharp minima. https://doi.org/10.48550/ARXIV.1609.04836.
21. Yadav A., Jha C. K. & Sharan A. (2020) Optimizing LSTM for time series prediction in Indian stock market. *Procedia Computer Science*, 167: 2091–2100. https://doi.org/10.1016/j.procs.2020.03.257.

Received June, 2024