Annals of Mathematics and Computer Science
Vol 2 (2021) 10-14

MACHINE LEARNING METHODS IN TIME SERIES FORECASTING: A REVIEW

SAIDJON KAMOLOV[1*], ISKHAKOV DILSHOD[2], BAKHROM ZIYAEV[3]

Abstract. The improvements in machine learning algorithms have spurred their application in various fields. In this paper, we consider the use of machine learning in financial forecasting. Our review consists of two parts: data driven approaches and model-based solutions. We discuss state-of-the-art literature in this field and analyze future trends.

1. Introduction

The recent successes of machine learning applications in fields such as computer vision, speech recognition, and recommender systems have galvanized interest from researchers in other fields [1, 4, 18, 21]. There has been an enormous effort across a range of scientific fields to apply machine learning method. In particular, machine learning has been applied in time series forecasting. There exist different fields of time series forecasting such financial, meteorological, astronomical, biological, and others. Financial forecasting is arguably the most popular subfield of time series analysis. A great number of research exists on applying machine learning to forecast the future price of an asset. In this paper, our goal is to present the current trends in research related to the use of machine learning for financial forecasting.

The application of machine learning to financial forecasting can be grouped into two parts: data preprocessing and learning algorithms. The data preprocessing addresses issues such identifying and selecting important feature in the data, imbalanced and missing data. The learning algorithms are used to analyze the data to build predictive models. There exists a plethora of machine learning algorithms that can be used to build predictive models such neural networks, support vector machines, random forests and other. Arguably the most popular method is deep learning which is based on multilayer neural networks. Furthermore, there are different types of deep learning methods that allow to learn from sequential data such as recurrent neural networks (RNNs) and long short-term memory (LSTMs). A correct combination of data-based and algorithm-based approaches can determine the success of machine learning in financial forecasting.

Date: Received: May 27, 2021; Accepted: Jul 12, 2021.

* Corresponding author.

Key words and phrases. machine learning, forecasting, feature selection, imbalanced data, deep learning.

Our paper is structured as follows. In Section 2, we discuss the data driven methods in financial forecasting such as feature selection, imbalanced data, and missing values. In Section 3, we present an overview of machine learning algorithms for constructing forecasting models. And Section 4 concludes with brief summary remarks.

2. Data driven methods

Data preprocessing is an important step in many machine learning pipelines. In order to achieve the optimal results the data must be appropriately prepared [16].

2.1. Feature selection.

One of the common issues in machine learning applications including forecasting is high number of features. For instance, in predicting the next day SP500 index one can consider the values of the 500 constituents. Given that many share prices in SP500 are correlated, having all of them in the input may lead to undesirable outcomes. Another issue with too many features in the computational complexity. A high dimensional dataset takes a long time to train. And finally, high dimensional data is prone for overfitting. To ameliorate the issues related to high dimensional data the it is often advised to reduce the number of features [23].

There exist a number of approaches to feature selection. The most popular feature selection method is called filtering where the features are ranked and selected based on some metric. Information gain and chi squared are two of the commonly used metric for feature evaluation. For continuous features, F-measure and correlation can also be used. However, filter methods from a major drawback which is feature interactions. Filter methods do not take into account interactions between features. Feature interactions can have a drastic effect of feature importance. To address the issue of feature interactions, a variance decomposition-based feature selection method was recently proposed in [7, 15]. By decomposing the target variable variance into its sources based on input variables in an orthogonal fashion allows taking into account feature interactions. Another simple, yet powerful feature selection method was introduced in [20]. The features who are closest to the target variable with respect to the L2 norm are selected. The geometric nature of the feature selection approach means that it can be applied in a variety of situations.

2.2. Imbalanced data.

Imbalanced data refers to skewed distribution of the target variable. It occurs in a range of machine learning application including medicine, intrusion detection, finance, and others. In financial forecasting, imbalanced data occurs in the context of direction prediction. Financial assets that are experiencing a period of growth or a period of decline will have a skewed distribution of the directional variable. For instance, the price of bitcoin has been consistently increasing over the last 2 years. Hence if we consider forecasting the directional change of bitcoin price there would be substantially more positive changes than negative changes. Another case where imbalanced data occurs is in forecasting significant changes in stock price. The price of the stock in most cases is relatively stable on a daily basis with significant fluctuations happening in rare cases. Thus we obtain imbalanced distribution between significant and nonsignificant price changes.

Imbalanced data has a negative impact on the performance of machine learning models [19]. To combat imbalanced data there are number of approaches: cost-sensitive approach, sampling methods, and one class learning. In cost-sensitive approach the the weight the minority class points is made higher than the weight on the majority class points. Consequently, the cost of misclassifying the minority point increases forcing the classifier to learn the minority patterns. Sampling methods achieve a balanced dataset by either reducing the number of majority point or increasing the number of minority points. In one-class learning the minority points are learned independently of the majority points. Sampling techniques are the most popular approaches [3, 22].

In oversampling, the number of minority points is increased to balanced with the majority points. One of the earliest attempts in this direction is the SMOTE algorithm, where a new minority point is synthetically generated between neighboring points [2]. There are also other methods that mimic the idea of SMOTE including ADASYN and Borderline-SMOTE [6, 5]. More recently kernel density estimation has also been used to approximate the underlying distribution of the minority points and use it to generate [14]. Gamma distribution has been used to generate new minority points between the existing neighboring points [12].

3. Deep learning for financial forecasting

Deep learning has been widely used in many applications including financial forecasting. Since financial forecasting is based on time series data, special types of neural networks such as RNNs and LSTMs have been used in such cases. RNNs are designed specifically to handle sequential data. In RNNs, the predicted value from the previous time step is used as an input for predicting the value in the next time step. One of the issues with RNNs is the exploding gradient problem where the value of the gradient during the backpropagation procedure become extremely large. In order to address the exploding gradients issue the LSTM model was proposed. LSTMs use special gates to control the flow of information inside the network. LSTMs are currently the preferred type of neural network for financial forecasting.

In [9], the authors investigated the use deep learning to forecast the SP500 index. The input for the forecasting model are the past index values and the output is the next-day index value. A similar work was carried out in [11]. In [10], the authors looked at forecasting the significant fluctuations in stock prices using deep learning. In particular, LSTM was used to predict changes in stock price above 1 standard deviation from the historical mean. This approach allows to forecast the stock volatility.

One of the key question in financial forecasting is the choice between return and price values. Although price and return are closely related, they represent slightly different quantities. Thus, the choice between the share price and return can potentially play an important role in the model performance. This issue was addressed in [8], where the authors analyzed the effects of price vs return on the performance of deep learning models. The results showed that price is a better predictor than return in deep learning model. However, the difference become negligible when other technical inputs are used.

4. Conclusion

Machine learning has become a popular tool in financial forecasting. It has been used in various subfields including stock price forecasting, index forecasting, exchange rate forecasting, and many others. In this note, we reviewed the current literature and issues surrounding the application of deep learning for financial forecasting. Our review consisted of two parts: data-driven approaches and model-based approaches. In data-drive approaches various preprocessing strategies such as feature selection and imbalanced data have been discussed. For model-based approaches, the use of deep learning algorithms such as LSTM were discussed in the context of stock price prediction.

References

1. Ahmed, N. K., Atiya, A. F., Gayar, N. E., & El-Shishiny, H. (2010). An empirical comparison of machine learning models for time series forecasting. Econometric Reviews, 29(5-6), 594-621.
2. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357.
3. Dai, F., Song, Y., Si, W., Yang, G., Hu, J., & Wang, X. (2021). Improved CBSO: A distributed fuzzy-based adaptive synthetic oversampling algorithm for imbalanced judicial data. Information Sciences, 569, 70-89.
4. Fan, Z., Li, J., Zhang, L., Zhu, G., Li, P., Lu, X., ... & Wei, W. (2021). U-net based analysis of MRI for Alzheimer's disease diagnosis. Neural Computing and Applications, 1-13.
5. Han, H., Wang, W. Y., & Mao, B. H. (2005, August). Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In International conference on intelligent computing (pp. 878-887). Springer, Berlin, Heidelberg.
6. He, H., Bai, Y., Garcia, E. A., & Li, S. (2008, June). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence) (pp. 1322-1328). IEEE.
7. Kamalov, F. (2021). Orthogonal variance decomposition based feature selection. Expert Systems with Applications, 182, 115191.
8. Kamalov, F., Gurrib, I., & Rajab, K. (2021). Financial Forecasting with Machine Learning: Price Vs Return. Journal of Computer Science, 17(3), 251-264.
9. Kamalov, F., Smail, L., & Gurrib, I. (2020, December). Forecasting with Deep Learning: S&P 500 index. In 2020 13th International Symposium on Computational Intelligence and Design (ISCID) (pp. 422-425). IEEE.
10. Kamalov, F. (2020). Forecasting significant stock price changes using neural networks. Neural Computing and Applications, 32(23), 17655-17667.
11. Kamalov, F., Smail, L., & Gurrib, I. (2020, November). Stock price forecast with deep learning. In 2020 International Conference on Decision Aid Sciences and Application (DASA) (pp. 1098-1102). IEEE.
12. Kamalov, F., & Denisov, D. (2020). Gamma distribution-based sampling for imbalanced data. Knowledge-Based Systems, 207, 106368.
13. Kamalov, F., & Leung, H. H. (2020). Outlier detection in high dimensional data. Journal of Information & Knowledge Management, 19(01), 2040013.

14. Kamalov, F. (2020). Kernel density estimation based sampling for imbalanced class distribution. Information Sciences, 512, 1192-1201.
15. Kamalov, F. (2018, December). Sensitivity analysis for feature selection. In 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA) (pp. 1466-1470). IEEE.
16. Kulczycki, P., & Franus, K. (2021). Methodically unified procedures for a conditional approach to outlier detection, clustering, and classification. Information Sciences, 560, 504-527.
17. Nguyen, H. M., Cooper, E. W., & Kamei, K. (2011). Borderline over-sampling for imbalanced data classification. International Journal of Knowledge Engineering and Soft Data Paradigms, 3(1), 4-21.
18. Prathap, B. R., Krishna, A. V., & Balachandran, K. Crime Analysis and Forecasting on Spatio Temporal News Feed Data—An Indian Context. Artificial Intelligence and Blockchain for Future Cybersecurity Applications, 307.
19. Thabtah, F., Hammoud, S., Kamalov, F., & Gonsalves, A. (2020). Data imbalance in classification: Experimental evaluation. Information Sciences, 513, 429-441.
20. Thabtah, F., Kamalov, F., Hammoud, S., & Shahamiri, S. R. (2020). Least Loss: A simplified filter method for feature selection. Information Sciences, 534, 1-15.
21. Wang, R., Gao, X., Gao, J., Gao, Z., & Xie, J. (2021). Uncertain texture features fusion based method for performance condition evaluation of complex electromechanical systems. ISA transactions, 112, 108-121.
22. Xiao, J., Wang, Y., Chen, J., Xie, L., & Huang, J. (2021). Impact of resampling methods and classification models on the imbalanced credit scoring problems. Information Sciences, 569, 508-526.
23. Zhou, H., Yu, K. M., Chen, Y. C., & Hsu, H. P. (2021). A Hybrid Feature Selection Method RFSTL for Manufacturing Quality Prediction Based on a High Dimensional Imbalanced Dataset. IEEE Access, 9, 29719-29735.

[1]Faculty of Engineering, Tajik Technical University, Dushanbe, Tajikistan.
Email address: said.kamolov@ttu.tj

[2]Bank Card Department, Spitamen Bank, Dushanbe, Tajikistan.
Email address: d.iskhakov@spitamen.com

[3]Macroeconomics, Trade and Investment Global Practice Department, World Bank, Dushanbe, Tajikistan.
Email address: bziyaev@worldbank.org