ACADEMIC
World Academic Union

ISSN 1746-7659, England, UK
Journal of Information and Computing Science
Vol. 16, No. 2, 2021, *pp*.108-117

Analysis of feature selection for stock price prediction with LSTM: A case study on China's new energy leading stocks

Wanbao Zhou

Nanjing University of Information Science and Technology, Nanjing, China
(Received June 21, 2021, accepted September 07, 2021)

Abstract: Stock price prediction has always been the focus of investors' attention in the stock market. In recent years, deep learning technology has been widely used in this field. In the era of big data, feature selection is a necessary part of data preprocessing. Feature selection is a data dimensionality reduction technology, and its main purpose is to select the relevant features that are most beneficial to the algorithm from the original data, reduce the dimensionality of the data and the difficulty of learning tasks, and improve the efficiency of the model. This paper has performed analysis of input feature selection with three feature selection methods: Multiple linear regression analysis, Correlation matrix heatmap, Feature importance. Plus the original features set, four different input features sets were provided for predicting stock price of ten China's new energy leading stocks with LSTM. From the conducted experiments, it is found that after using the feature selection method, the prediction results of all ten stocks are performed better than the prediction results under the original features.

Keywords: Stock Price Prediction, LSTM, Feature Selection, Multiple linear regression analysis, Correlation matrix heatmap, Feature importance

1. Introduction

With the rapid development of China's economy, finance has become the core force of the modern economy and an important core competitiveness of the country. The effective development of the financial market plays a vital role in promoting the development of China's overall economy. The stock market is an important part of the financial market. The stable development of the stock market has become an important prerequisite for the sustained and stable development of the Chinese economy.

Environmental protection is a basic national policy in China. With the deepening of economic reform and development, especially after joining the World Trade Organization, environmental protection has attracted more and more attention. Coal prices have risen sharply this year, and many provinces and cities have issued power rationing orders on companies. From the perspective of environmental protection, this aspect has been affected by the 30-year carbon peak and 60-year carbon neutrality of the "14th Five-Year Plan". On the one hand, it is also affected by energy scarcity. Environmental issues once again sounded the alarm. The transformation of the energy structure is the general trend, and the replacement of traditional energy by new energy is an irreversible process.

In recent years, the source and scale of stock market data have increased rapidly. Deep learning models have been introduced into many research scenarios in stock market forecast due to their excellent large-scale data processing capabilities [1]. Recurrent Neural Networks (RNNs) is a mainstream deep neural network used to sequence modeling, which solves the problem that traditional feedforward neural networks cannot handle variable-length sequences. It has received a great amount of attention due to their flexibility in capturing nonlinear relationships [2-4]. However, traditional RNNs suffer from the problem of vanishing gradients and thus have difficulty in capturing long-term dependencies [5]. Long Short-term memory units (LSTM) have overcome this limitation and achieved great success in many applications, such as sequence prediction [6].

For a particular learning algorithm, which feature is effective is unknown. Therefore, it is necessary to select relevant features that are beneficial to the learning algorithm from all the features. And in practical applications, the problem of dimensional disasters often occurs. If only some of the features are selected to construct the model, the running time of the learning algorithm can be greatly reduced, and the interpretability of the model can also be increased. First of all, the definition of feature selection was looking for the smallest feature subset for the model to be effective under ideal conditions [7]. Subsequently, the definition of feature

selection evolved to try to find the smaller feature subset between the feature and the original data in the case of similar set distributions [8]. The ensemble learning idea was first proposed by Breiman from the statistics Bootstrap idea. This idea extended the classification advantages of the base learner and greatly promoted the research and development of classification models. After that, Breiman and Cutler first proposed the Bagging idea and random subspace combination method as an effective integrated learning classification prediction tool [9]. This method is continuously improved by combining the decision tree algorithm to form Random Forest, laying the foundation for the integrated learning algorithm. Random survival forest was proposed to effectively process high-dimensional data [10]. Extremely randomized trees were proposed to improve the overall analysis of variance and bias for Random Forest which does not include attribute selection [11].

This paper has performed an analysis of look-back period used with Long Short-term Memory (LSTM) for predicting stock prices. First, using the data of a stock to determine the parameters for model LSTM. Then, three feature selection methods as follows were used for obtaining new input features subset: Multiple linear regression analysis, Correlation matrix heatmap, Feature importance. Last, plus the original features set, four different input features sets were provided for predicting stock price. The main concern is the difference in prediction results under different input features. Ten China's new energy leading stocks are analyzed in this research work.

2. Long Short-term Memory Network

Long Short-term Memory Network (LSTM) is a type of Recurrent Neural Network (RNN). RNN has huge difficulties in dealing with long-term dependencies, because the calculation of connections between distant nodes will involve multiple multiplications of the Jacobian matrix, which will cause the vanishing gradient problem or exploding gradient problems. In order to solve this problem, researchers have proposed many solutions. Among them, the most successful and widely used is the gated RNN , and LSTM is the most famous kind of the gated RNN. Leaky units allow RNN to accumulate long-term connections between distant nodes by designing the weight coefficients between connections. And gated RNN generalizes this idea, allowing the coefficient to be changed at different times, and allowing the network to forget the current accumulation Information. LSTM is such a gated RNN, and its single node structure is shown in the Fig.1. The ingenuity of LSTM is that by increasing the input gate, forget gate and output gate, the weight of the self-loop is changed. In this way, when the model parameters are fixed, the integration scale at different moments can be dynamically changed, combining short-term memory with long-term memory which avoids the problem of the vanishing gradient problem or exploding gradient problems to a certain extent.

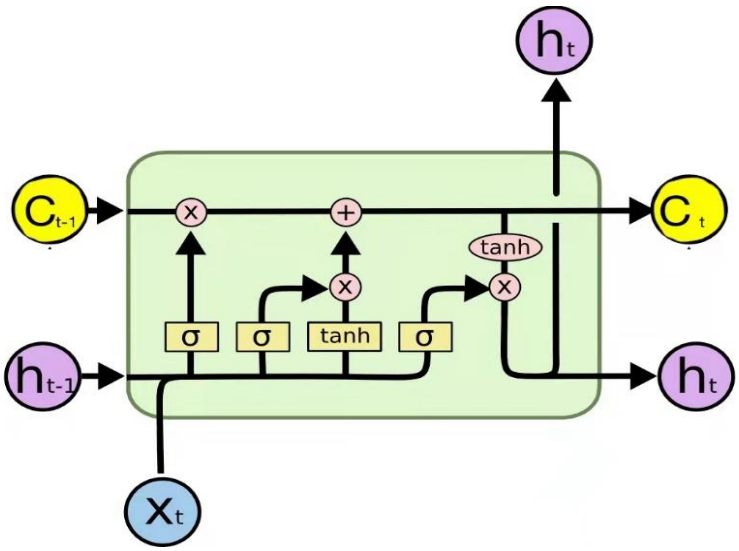

Fig.1. LSTM Architecture