Research on Stock Portfolio Construction Based on Bi-LSTM Neural Networks

Jiawei Li *

School of Statistics, Tianjin University of Finance and Economics, Tianjin, China

* Corresponding author Email: 3137704296@qq.com

Abstract: In recent years, the rapid globalisation of China's financial market has provided more opportunities for investors, but also brought about a more complex investment environment. This paper constructs stock portfolios based on Bi-LSTM neural networks, aiming to improve the accuracy of stock price prediction and the optimisation of investment portfolios using deep learning techniques. The theoretical part introduces the portfolio theory, including the mean-variance model and the capital asset pricing model, and explores the advantages of LSTM, Bi-LSTM and ATT-LSTM in processing time series data. The constituent stocks of CSI 300 index are selected in the research design part, and the stocks are screened using entropy weighted TOPSIS method and analysed based on the data from January 2018 to April 2024. The closing price and logarithmic return are predicted by constructing and using LSTM, Bi-LSTM and ATT-LSTM models, and then the trading strategies of EMA, MACD double conditions are determined, and the investment weights are determined by Monte Carlo method for the investment portfolio. The results of the empirical study show that the Bi-LSTM model has the optimal prediction performance, and based on the prediction data of the model, the trading strategy using the dual conditions of EMA and MACD achieves a higher investment return than the strategy using only MACD. In summary, this paper demonstrates the superiority of Bi-LSTM model in stock price prediction through empirical research, and proposes an effective portfolio construction method and trading strategy, which helps investors make more effective decisions in the complex market environment.

Keywords: Bidirectional long and short-term memory neural network; Entropy weight TOPSIS method; Monte Carlo method.

1. Introduction

With the acceleration of globalisation, China's financial market has become increasingly internationalised, providing more opportunities for investors but also bringing about a more complex investment environment. In this context, accurately predicting stock prices and optimising investment portfolios have become key issues. Traditional mean-variance models and capital asset pricing models have limited predictive power in the face of complex time series data. Deep learning, especially Long Short-Term Memory (LSTM) networks, excels in handling time series data. Bidirectional Long Short-Term Memory Networks (Bi-LSTM) further improves the ability to capture both forward and backward information, resulting in more accurate stock price predictions [1].

The purpose of this paper is to construct a stock portfolio model based on Bi-LSTM neural network and explore the application of deep learning techniques in stock price prediction and portfolio optimisation. We select the constituent stocks of CSI 300 index, use the entropy weight TOPSIS method for stock screening, and conduct empirical analysis based on the data from January 2018 to April 2024. By comparing the forecasting performance of LSTM, Bi-LSTM and ATT-LSTM models, and combining the trading strategy with the dual conditions of EMA and MACD, we verify the superiority of the Bi-LSTM model in enhancing investment returns.

This study enriches the portfolio theory and provides new decision-making methods for investors in complex market environments, aiming to promote the development of financial markets and the innovation of investment strategies.

2. Relevant theories

2.1. Portfolio

2.1.1. Portfolio Theory

Portfolio theory, first introduced by the American economist Markowitz in 1952, uses mean and variance to portray return and risk, two key factors that influence investment choices [2]. In a given portfolio, the mean represents the expected return, the variance is the variance of the portfolio’s return, and the standard deviation is used to measure systematic risk. This is shown in Figure 1.

2.1.2. Capital Asset Pricing (CAPM) model

The Capital Asset Pricing Model (CAPM) is used to estimate the expected return of an asset and describes the relationship between the expected return of an asset and its risk with the following expression:

\[ r_i = r_f + \beta_i (r_m - r_f) \]

(1)

Where \( r_i \) is the expected return of asset i, \( r_f \) is the risk-free rate, \( \beta_i \) is the beta coefficient of asset i, which represents the systematic risk of the asset relative to the
market portfolio, and $\bar{r}_m$ is the expected return of the market portfolio.

The CAPM, as one of the cornerstones of capital market theory, provides an important theory for the study of asset pricing, which helps investors to understand the relationship between asset return and risk, and provides a basic explanation for the operation of financial markets.

### 2.2. Bi-Directional Long Short-Term Memory

In order to deal with the problems of gradient vanishing and gradient explosion that exist in traditional RNN, Hochreiter and Schmidhuber proposed a special type of RNN in 1997, i.e., Long Short-Term Memory Neural Network (LSTM) [3], which can make up for the deficiencies of RNN due to the unique network structure that LSTM possesses. The LSTM, in comparison with RNN, has three gating units, $i_t$ and $f_t$, which solves the problem of long time sequence dependency in deep learning. Compared with RNN, LSTM adds three gating units, namely, the forgetting gate $f_t$, the input gate $i_t$ and the output gate $o_t$, which solves the problem of long time sequence dependency in deep learning, and the structure is shown in Figure 2:

![Figure 2. LSTM network structure](image)

#### 2.3. Bi-Directional Long Short-Term Memory Neural Network

An unavoidable problem in LSTM modeling is the inability to encode back-to-front information. However, bi-directional LSTMs can extract the information before and after the data at the same time, which in turn mines deeper features and better captures the dependencies of bi-directional information. Two LSTMs are combined to form a single-layer Bi-LSTM, where one LSTM can process the input sequence in the forward direction; the other LSTM can process it in the reverse direction, and the outputs of the two are spliced together to form the final Bi-LSTM output after all the time steps are completed. The forward LSTM generates a result vector after processing all time steps, and the reverse LSTM also gets another result vector. These two result vectors are combined as the result of Bi-LSTM [4], which is then input into the subsequent neural network for regression or classification to obtain the final output features, the structure of which is shown in Figure 3:

![Figure 3. Bi-LSTM structure](image)

### 2.4. The Long Short-Term Memory Neural Network based on Attention Mechanism

In daily life, human beings tend to selectively receive useful information because of their limited ability to process information. The ATT-LSTM model is similar to the process of human beings processing information [5], and it can effectively selectively receive part of the information when the model learns and receives a large amount of information, and the structure is shown in Figure 4:

![Figure 4. ATT-LSTM structure](image)

### 3. Research design

#### 3.1. Basic processing of sample data

This paper uses the data of CSI300 index constituents are historical daily data, containing three major categories of a total of 31 characteristic indicators, the time interval is from January 1, 2018 to April 1, 2024, the indicator data are from tushare financial data interface, and the use of Python software for the subsequent study. 00 index constituents are historical daily data, containing a total of 31 characteristic indicators in three categories, the time interval is from January 1, 2018 to April 1, 2024, and the indicator data are sourced from tushare financial data interface, using Python software for subsequent research.

Cumulative/distribution points are obtained based on relevant data and formulae; dividend yields are missing for some stocks, which are treated as zero for companies with excessively long dividend distributions; negative P/E ratios according to tushare are treated as missing values for removal.

#### 3.2. LSTM neural network modeling applications

Based on the relevant theoretical learning, it can be seen that LSTM solves the gradient disappearance and gradient
3.3.2. Determination of investment weights

In this paper, Monte Carlo method is used to simulate the weights and calculate the effective frontier curve of the portfolio under each weight [6], where: \( r_i \) is the return of the \( i \) stock; \( \bar{r}_t \) is the expected return of the ith stock; \( \sigma_i \) is the standard deviation of \( r_i \); \( P[r_t, \ldots, r_n] \) is the risky portfolio, which consists of \( n \) risky assets according to a certain weight ratio; The weight of each asset in \( P \) is \( w_i \) and satisfies the following requirement:

\[
\sum_{t=1}^{n} w_i = 1
\]

(8)

Assume that the market is completely free to trade without restriction, \( w_i \in R \). Based on the returns on individual assets, the return on the portfolio asset \( P \) can be calculated as:

\[
r_p = \sum_{i=1}^{n} w_i r_i
\]

(9)

The expected return and variance of the portfolio is:

\[
E(r_p) = E(\sum_{i=1}^{n} w_i r_i) = \sum_{i=1}^{n} w_i E(r_i) = \sum_{i=1}^{n} w_i \bar{r}_i \\
Var(r_p) = E[(r_p - E(r_p))^2] = \sum_{i=1}^{n} w_i \sum_{j=1}^{n} w_j Cov(r_i, r_j)
\]

(10)

Construct the MPT model: assume that the fixed expected return is \( \mu \), the weights \( w = (w_1, w_2, \ldots, w_n) \), and the portfolio of risky assets \( P \) satisfies that the variance of \( P \) is minimized among all the portfolios that can achieve the expected return \( \mu \), i.e:

\[
\min_{w} \sum_{i=1}^{n} w_i \sum_{j=1}^{n} w_j Cov(r_i, r_j) \\
\text{s.t. } \sum_{i=1}^{n} w_i = 1
\]

(13)

4. Empirical studies

4.1. Model evaluation based on entropy weight TOPSIS method

In this section, the monthly average data (March-April 2024) of CSI 300 index constituents are normalized and then principal component analysis (KOM value of 0.775, p-value of 0.000) is used to downsize 31 indicators into 10 indicators (explaining more than 90% of the information of the overall indicators), and then entropy-weighted TOPSIS is used to screen out five stocks for LSTM prediction afterwards [7].

4.1.1. Entropy weighting

In the first step, the normalized data for each principal component \((X'_{ij})\) is used to calculate the weight of the value of the \( i \) stock under the \( j \) indicator in the sum of all values \((P_{ij})\):

\[
P_{ij} = \frac{X'_{ij}}{\sum_{i=1}^{n} X'_{ij}}
\]

(14)

In the second step, the information entropy value of the \( j \) indicator \((e_j)\) is calculated.

\[
e_j = -k \times \sum_{i=1}^{n} (P_{ij} - \ln P_{ij})
\]

(15)

In the third step, the information entropy redundancy \((d_j)\) is calculated.

\[
d_j = 1 - e_j
\]

(16)

In the fourth step, the weights of the evaluation indicators are calculated \((W_j)\):

\[
W_j = \frac{d_j}{\sum_{j=1}^{m} d_j}
\]

(17)

The specific weights of each principal component are shown in Table 1 below.
Table 1. Summary of the results of weight calculation by entropy method

<table>
<thead>
<tr>
<th>Term</th>
<th>information entropy</th>
<th>information utility value</th>
<th>weighting factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pca10</td>
<td>0.9955</td>
<td>0.0045</td>
<td>4.85%</td>
</tr>
<tr>
<td>Pca9</td>
<td>0.9939</td>
<td>0.0061</td>
<td>6.60%</td>
</tr>
<tr>
<td>Pca8</td>
<td>0.9876</td>
<td>0.0124</td>
<td>13.46%</td>
</tr>
<tr>
<td>Pca7</td>
<td>0.9933</td>
<td>0.0067</td>
<td>7.27%</td>
</tr>
<tr>
<td>Pca6</td>
<td>0.9922</td>
<td>0.0078</td>
<td>8.49%</td>
</tr>
<tr>
<td>Pca5</td>
<td>0.9951</td>
<td>0.0049</td>
<td>5.28%</td>
</tr>
<tr>
<td>Pca4</td>
<td>0.9944</td>
<td>0.0056</td>
<td>6.04%</td>
</tr>
<tr>
<td>Pca3</td>
<td>0.9901</td>
<td>0.0099</td>
<td>10.74%</td>
</tr>
<tr>
<td>Pca2</td>
<td>0.9987</td>
<td>0.0013</td>
<td>1.38%</td>
</tr>
<tr>
<td>Pca1</td>
<td>0.9670</td>
<td>0.0330</td>
<td>35.90%</td>
</tr>
</tbody>
</table>

4.1.2. TOPSIS

In the first step, the weighted normalized decision matrix \((Z_{ij})\) is determined.

\[
Z_{ij} = W_j X_{ij}
\]  

In the second step, determine the positive ideal solution \((Z_j^+)\) and the negative ideal solution \((Z_j^-)\).

\[
\begin{align*}
Z_j^+ &= \max \{Z_{ij}\} \quad (j = 1, 2, \ldots, 10) \\
Z_j^- &= \min \{Z_{ij}\} \quad (j = 1, 2, \ldots, 10)
\end{align*}
\]  

In the third step, the distance of the evaluation object from the positive and negative ideal solutions is calculated \((D_j^+), (D_j^-)\).

\[
\begin{align*}
D_j^+ &= \frac{\sum_{j=1}^{m} (Z_{ij} - Z_j^+)^2}{\sum_{j=1}^{m}} \\
D_j^- &= \frac{\sum_{j=1}^{m} (Z_{ij} - Z_j^-)^2}{\sum_{j=1}^{m}}
\end{align*}
\]  

In the fourth step, the closeness of the evaluation object to the positive and negative ideal solutions is calculated \((C_i)\).

\[
C_i = \frac{D_j^+}{D_j^+ + D_j^-}
\]

The results of the TOPSIS evaluation calculations are shown in Table 2.

Table 2. TOPSIS evaluation calculations

<table>
<thead>
<tr>
<th>Term</th>
<th>positive ideal solution distance (D^+)</th>
<th>Negative ideal solution distance (D^-)</th>
<th>relative proximity (C)</th>
<th>Sorting results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject of evaluation 55</td>
<td>1.498</td>
<td>13.628</td>
<td>0.901</td>
<td>1</td>
</tr>
<tr>
<td>Subject of evaluation 104</td>
<td>8.068</td>
<td>5.880</td>
<td>0.422</td>
<td>2</td>
</tr>
<tr>
<td>Subject of evaluation 12</td>
<td>8.210</td>
<td>5.646</td>
<td>0.407</td>
<td>3</td>
</tr>
<tr>
<td>Subject of evaluation 49</td>
<td>8.637</td>
<td>5.234</td>
<td>0.377</td>
<td>4</td>
</tr>
<tr>
<td>Subject of evaluation 141</td>
<td>9.060</td>
<td>4.779</td>
<td>0.345</td>
<td>5</td>
</tr>
</tbody>
</table>

Find out the stock names of the corresponding rows based on the serial numbers of the evaluation objects, i.e., evaluation object 55 is Guizhou Moutai, evaluation object 104 is Kingsoft Office, evaluation object 12 is Beifang Huachuang, evaluation object 49 is Gujing Gongjiu, and evaluation object 141 is Ningde Times.

4.2. LSTM models and comparisons

According to the feature importance ranking graph, open, high, low, pre_close, change, pct_chg, vol, amount, Ln_return, are used as input features to the LSTM model (where pct_chg is the raw data of yield). Because there is a Dropout layer to prevent overfitting, the program is run more often to find more accurate model results, and the time is in reverse order because the data is pulled directly from tushare. Here, the Guizhou Moutai stock is used as an example to compare the three models. The results are shown in Figure 5, Figure 6 and Figure 7.
results are the best among the three models, so using Bi-LSTM model prediction data for portfolio construction.

<table>
<thead>
<tr>
<th>Model</th>
<th>Target audience</th>
<th>Average absolute error MAE</th>
<th>Rms error RMSE</th>
<th>Mean absolute percentage error MAPE</th>
<th>Coefficient of determination $R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSTM</td>
<td>closing price</td>
<td>5.9449</td>
<td>9.0114</td>
<td>0.34%</td>
<td>0.9839</td>
</tr>
<tr>
<td>Bi-LSTM</td>
<td>closing price</td>
<td>4.2523</td>
<td>7.1815</td>
<td>0.25%</td>
<td>0.9898</td>
</tr>
<tr>
<td>ATT-LSTM</td>
<td>closing price</td>
<td>5.1381</td>
<td>7.4302</td>
<td>0.30%</td>
<td>0.9891</td>
</tr>
<tr>
<td>LSTM</td>
<td>logarithmic yield</td>
<td>0.0015</td>
<td>0.0024</td>
<td>55.26%</td>
<td>0.9653</td>
</tr>
<tr>
<td>Bi-LSTM</td>
<td>logarithmic yield</td>
<td>0.0008</td>
<td>0.0014</td>
<td>22.14%</td>
<td>0.9891</td>
</tr>
<tr>
<td>ATT-LSTM</td>
<td>logarithmic yield</td>
<td>0.0011</td>
<td>0.0017</td>
<td>21.82%</td>
<td>0.9828</td>
</tr>
</tbody>
</table>

5. Conclusions

Through this study, we validate the superiority of stock portfolio models based on Bi-LSTM neural networks in stock price prediction and portfolio optimisation. Empirical analyses show that the Bi-LSTM model significantly outperforms the traditional LSTM and ATT-LSTM models in terms of prediction accuracy when dealing with complex time series data. Combined with the dual-conditional trading strategy of EMA and MACD, the predicted data based on the Bi-LSTM model achieves higher investment returns with fewer trades, which further proves the effectiveness of the method in practical applications.

This study not only enriches the content of portfolio theory, but also provides investors with new tools to make scientific investment decisions in a complex market environment. By introducing deep learning technology, we effectively improve the accuracy of stock price prediction and optimise the portfolio construction method. This provides financial market participants with new ideas to help cope with the increasingly complex investment environment.

Future research can validate the robustness of the model over a wider dataset and longer time horizon, while exploring more applications of deep learning models in the financial market. Through continuous improvement and innovation, it is believed that deep learning technology will play a greater role in the financial field.

References


