Multimedia Tools and Applications (2022) 81:12599–12617
https://doi.org/10.1007/s11042-022-12381-6

Stock linkage prediction based on optimized LSTM model

Chi Ma [1] · Yan Liang [2] · Shaofan Wang [1,3] · Shengliang Lu [1,3]

Received: 15 September 2020 / Revised: 15 January 2022 / Accepted: 21 January 2022 /
Published online: 19 February 2022

Abstract
Stock linkage refers to the correlation or similar performance of two or more stocks in the stock market. The quantification of stock linkage relationship is the trend and difficulty of research in recent years. The study of stock linkage can dig out the potential relationship between stocks at a deeper level. At present, the existing research often only studies the linkage phenomenon from the perspective of the correlation or similarity of stock movement, and there is no unified and standard numerical index to effectively describe the degree of linkage phenomenon, which greatly hinders the progress of research. Aiming at the problem that it is difficult to quantify the phenomenon of stock linkage, we analyze the correlation and morphological similarity of time series, and propose the combination of correlation coefficient and time weighted distance as the numerical expression of stock linkage for the first time, so as to realize the quantification of stock linkage. In addition, the parallel network structure of LSTM model is designed, and the automatic noise reduction encoder and wavelet transform module are added as the noise reduction processing layer, which effectively improves the prediction performance of LSTM model for stock market linkage numerical time series. Three different types of comparative experiments based on 2.309 million stock market sequences show that the proposed optimized LSTM model has more accurate prediction effect, and its RMSE

✉ Yan Liang
lijingyan_2020@126.com

✉ Shaofan Wang
wsf19961230@163.com

Chi Ma
machi@hzu.edu.cn

[1] School of Computer Science and Engineering, Huizhou University, Huizhou 516007, China

[2] School of Applied Technology, University of Science and Technology Liaoning, 114051 Anshan, China

[3] School of Computer Science and Software Engineering, University of Science and Technology Liaoning, 114051 Anshan, China

error is 18.68% lower than the compared DB-LSTM model and 46.38% lower than SDAE-LSTM model.

Keywords Stock linkage · Long short-term memory · Numerical representation · Dynamic time warping · Optimized model

1 Introduction

The general definition of stock linkage [4] phenomenon is that the stock market trend has a high correlation over a period of time and a similar stock price fluctuation curve. Effective discovery of stock linkage can help investors improve portfolio efficiency and avoid certain investment risks [13]. The fluctuation of stock price is related to several factors—they not only include the operating conditions of the listed companies but also factors that are difficult to quantify such as national policies, tendencies of public opinion and investor sentiment at the given time. Therefore, the phenomenon of stock price linkage is usually difficult to fully explain, and it is even more difficult to formulate a unified analysis and evaluation criteria for it, which further increases the level of difficulty of research in this field [17].

Stock linkage is used to describe the degree of linkage effect between two stocks in the phenomenon of stock linkage, and it has a deeper meaning than correlation. Correlation mainly expresses static correlation—regardless of time order, and mainly examines the correlation in numerical values. However, linkage expresses dynamic correlation, and it is a sudden and persistent phenomenon. Even if the stock group has strong linkage, the linkage behavior will appear time-offset and duration inconsistent owing to the loss of information in the stock market and the time cost. This complexity indicates that the research on stock linkage is still limited to the study of correlation between stocks.

Based on the different stock data used, there are several ways to mine the correlation between stocks. Generally, the mining can be divided into association network based on the text information for association mining [3] or based on time series data of volume and price [11, 12].

Through the study of a large volume of financial time-series data and the analysis of the periodicity of human economic activities, scholars have realized that financial time-series data show a certain time-varying feature [14]. Relevant studies have proved that most temporal features change from original invariance to variable with the increase in time duration. In the financial domain, the traditional data mining model cannot deal with such complex data describing financial markets [18]. Moreover, the traditional model mainly relies on the artificial design of features, and it is difficult to avoid the influence of subjective, targeted, and incomplete factors. However, the development of neural network methods, especially deep learning, is useful to effectively mitigate these problems to some extent [5, 9].

Currently, in the financial domain, the application of deep learning mainly focuses on the prediction of financial market movement. To predict the continuous weekly data of different exchange rates, a deep confidence network that uses continuous restricted Boltzmann machine (CRBM) was constructed by Shen et al. Cheng S H used decision tree and neural network to solve the problem of stock classification [6]. They take financial indicators as the core, combined with decision tree technology, establish mixed classification models and prediction rules that affect stock price fluctuations. However, there are many factors affecting stock price fluctuation, and it is risky to predict only with financial indicators as the core data. Dixon M

et al. describe the application of deep neural networks for predicting financial market-movement directions. They describe the configuration and training approach and then demonstrate their application to backtest a simple trading strategy over 43 different commodity and FX future mid-prices at 5-min intervals [8]. The specific applications in the literature are very innovative, but the performance of the model in different market environments still needs more support from comparative data, and there is still some room for improvement in the accuracy of prediction. Akita et al. proposed a novel application of deep learning models: paragraph vector and long short-term memory (LSTM) for financial time-series forecasting. The performance of their approach is demonstrated on 50 companies listed on the Tokyo Stock Exchange [2]. However, the stock markets of different countries have their own characteristics. In the Chinese market, the factors affecting the stock price are complex. In the process of solving practical problems, we still need to make targeted improvements on the basis of some known models. Presently, deep learning models for various application scenarios and research tasks have been designed in various fields. This is further promoting the development of deep learning while achieving continuous progress in various fields [7, 16].

However, the accuracy of the above methods for classification or time series prediction still has a large room for improvement. In addition, due to the lack of a unified numerical index to quantify the degree of stock linkage, people cannot directly apply the deep learning method to the prediction of stock linkage, but can only be applied to the prediction of financial time series, which makes a natural barrier between the scientific research method and the actual application, and cannot be broken through.

The phenomenon of stock linkage is usually manifested as time dislocation, duration difference and linkage range difference between different stocks. The existing research often studies the performance of the stock market from the perspective of correlation or similarity, which cannot directly predict the phenomenon and degree of stock linkage. This limitation greatly hinders the development of stock linkage research. Therefore, this paper proposes a unified and standard numerical index to effectively describe the degree of stock market linkage. Based on dynamic time warping (DTW), this paper proposes a numerical criterion to describe the degree of stock linkage. Furthermore, an optimized model based on deep learning is constructed to predict the future linkage between stocks. This prediction model breaks the previous model of stock link research based on stock price time series data or stock fundamental data only. By introducing different types of features, it provides a new and optimized method for stock linkage analysis.

We obtained 190,900 stock market data on a single day, and through the pairwise combination of 100 stocks, we finally got 4950 stock market data on a single day, with a total of 2,300,900 stock market difference series. Time-weighted DTW algorithm is used to mine the similar information of time series in morphology, deal with the continuity and lag of stock linkage phenomenon, and emphasize the impact of recent stock market changes on stock linkage. The time-weighted DTW distance value is calculated and converted into DTW similarity, and then combined with Pearson partial correlation coefficient to obtain the numerical expression of stock linkage considering the stock market environment. A two-layer LSTM mesh model is constructed, and wavelet transform is used to denoise the input sequence of the model. The experimental results show that our model has a great advantage in predicting the performance of stock linkage numerical time series. Compared with other methods, it can reduce the RMSE error by 46.38% at most.

The remainder of this paper is organized as follows. Section 2 introduces the numerized representation of stock linkage in detail. Section 3 describes the establishment of stock linkage

prediction model based on optimized LSTM. Section 4 presents the experiments on the proposed model and their results. Finally, Section 5 states the conclusions and comments on further research.

2 Numerized representation of stack linkage

Stock linkage is a numerical criterion used to describe the degree of stock linkage. Statistical methods have been commonly utilized for this purpose in previous studies. The degree of correlation between two time series is expressed by calculating the correlation coefficient and partial correlation coefficient between stock price time series. Although the calculated results can express stock linkage to a certain extent, there is still a discrepancy between them and the results of practical application.

In reality, the inter-stock linkage is not only reflected in the numerical correlation of stock price series but also in the degree of similarity in trends. Therefore, we propose a new numerical expression. The time-weighted DTW algorithm is used to process stock price time-series data. The calculated optimal alignment path distance is converted into similarity and combined with Pearson partial correlation coefficient. The mixed expression is considered as the numerical expression of stock linkage. This expression shows improved consideration of the continuity and lag of the stock linkage phenomenon; therefore, it is more suitable for describing stock linkage. Firstly, we select the appropriate correlation coefficient to describe the correlation of stock prices. Then, we use the algorithm based on dynamic time warping to calculate the linkage of stocks, and time weight the dynamic time warping algorithm. Finally, we linearly combine the two numerical expressions from the perspectives of numerical characteristics, morphological characteristics, time dimension and so on, the linkage index between stocks with stronger expression ability is obtained.

2.1 Stock relevance based on Pearson correlation coefficient

Pearson correlation coefficient [1] is used to quantitatively express the possible linear correlation between continuous random variables with fixed distances. The Pearson correlation coefficient of two fixed-distance continuous random variables: X and Y is equal to the quotient of the product of covariance and standard deviation between them. It is calculated by Eq. (1).

$$\rho_{x,y} = \frac{cov(X,Y)}{\sigma_x \sigma_y} = \frac{E((X-\mu_X)(Y-\mu_Y))}{\sigma_x \sigma_y} = \frac{E(XY) - E(X)E(Y)}{\sqrt{E\left(X^2\right) - E^2(X)}\sqrt{E\left(Y^2\right) - E^2(Y)}} \quad (1)$$

where,

$\mathbf{cov}(\mathbf{X}, \mathbf{Y})$—Covariance between variables X and Y,
$\sigma_{\mathbf{x}}$, $\sigma_{\mathbf{y}}$—Standard deviation of variables x and y, respectively,
$\mu_{\mathbf{X}}$,$\mu_{\mathbf{Y}}$—Mean of variables x and y, respectively,
$\mathbf{E}(\mathbf{X})$—Expectations for variable x.

The Pearson correlation coefficient is between -1 and 1. If the absolute value of the coefficient value is close to 0, the correlation degree between the random variables is low. The relationship between the range of Pearson correlation coefficient and the degree of correlation is presented in Table 1.

Table 1 Pearson correlation coefficient range and correlation degree comparison table

Pearson correlation coefficient range	Correlation degree
0.8–1.0	very high
0.6–0.8	high
0.4–0.6	moderate
0.2–0.4	low
0.0–0.2	no

Pearson correlation coefficient is usually chosen to calculate the correlation degree between two stock price time series while using the correlation coefficient to describe stock linkage. Although Pearson correlation coefficient does not consider the temporal order of the original data, it is unable to deal with time series of different lengths and cope with the lagging problem in the phenomenon of stock linkage. However, the correlation analysis of stock price volatility in a short duration can still achieve good results.

As shown in Fig. 1, during the period from 25 November, 2016 to 29 December, 2016, the downward trend of China's A-share market and the lack of investor confidence in the stock market seem to cause the downward trend of stocks 000063 and 000166 simultaneously. However, this is not indicative of a linkage effect between the two stocks.

In practice, stock linkage is not only caused by mutual influence but also by the fluctuations in the stock market as a whole. Therefore, we choose the time series data of the Shanghai-Shenzhen 300 Index as the market environment variable. They accurately express the linkage between stocks by calculating the first-order partial correlation coefficient of stock price time series by considering the market environment variables. This partial correlation coefficient is given by Eq. (2) as follows.

$$\gamma_{ij(h)} = \frac{\gamma_{ij} - \gamma_{ih}\gamma_{jh}}{\sqrt{1-\gamma_{ih}^2}\sqrt{1-\gamma_{jh}^2}} \tag{2}$$

where,

$\gamma_{ij(h)}$—Partial correlation coefficient of stocks i and j after controlling market environment variable, h,

γ_{ih}—Simple correlation coefficient between stock and environmental variables.

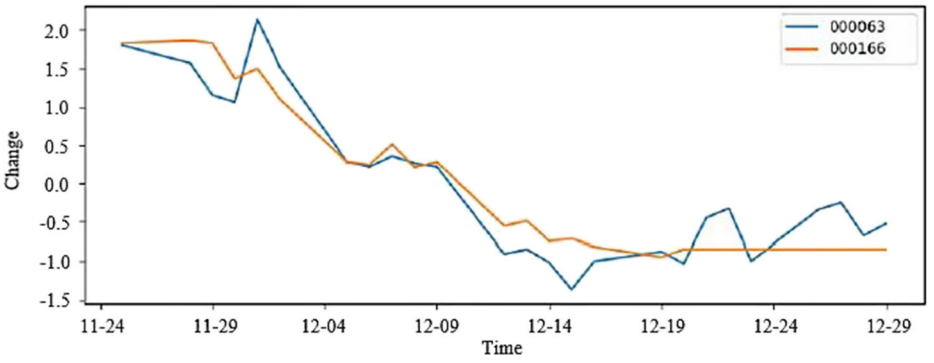

Fig. 1 Comparison of stock price trends

2.2 Stock linkage based on DTW

DTW seeks the optimal alignment path between two time series by minimizing the cumulative distance between them. DTW can not only handle time series of different lengths but also fold and twist the time so that the corresponding peaks or troughs can be aligned at different time points. This feature enables DTW to tackle the problems in stock linkage phenomenon such as the lag of time dislocation of linkage effect. An example of time series alignment for DTW is illustrated in Fig. 2 [10].

In the DTW algorithm, a distance mapping table, D, is created between the time series by calculating distance di, j between the different elements in the time series with lengths m and n, and then the minimum cumulative distance, Di, j, between different elements of the two time series is calculated by dynamic programming. Subsequently, a cumulative distance map between the time series is constructed. Di, j represents the minimum cumulative distance required to reach point (i, j) from origin (0,0), and it is given by Eq. (3) as follows.

$$D_{i,j} = d_{i,j} + \min\left\{\mathrm{D_{i,j-1}}, \mathrm{D_{i-1,j}}, \mathrm{D_{i-1,j-1}}\right\} = 2, \ , \ \mathrm{m} = 2, \ , \ \mathrm{n} \tag{3}$$

The initial conditions in the equation are set as follows:

$$D_{1,1} = d_{1,1} \tag{4}$$

$$D_{1,j} = \sum_{p-1}^{j} d_{1,p} j = 1, , n \tag{5}$$

$$D_{i,1} = \sum_{q-1}^{i} d_{q,1} i = 1, , m \tag{6}$$

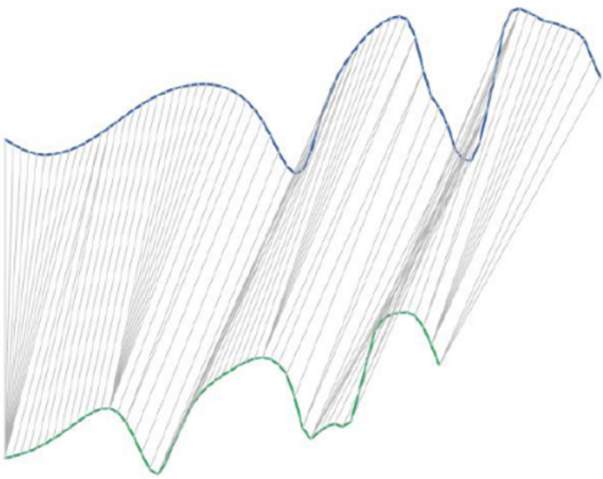

Fig. 2 Non-linear alignment of time series by DTW

DTW algorithm can not only find the shortest alignment path of two time series, but also calculate the distance of the path. Therefore, DTW algorithm is also a common time series similarity calculation method. Therefore, DTW algorithm is also a common time series similarity calculation method. In order to complete different time series analysis tasks, DTW algorithm has developed a variety of step patterns in calculating the distance between different time series elements, which are " symmetric1", " symmetric2", " asymmetric", " rabinerJuang" and " symmetric5". A visual example of them is shown in Fig. 3.

In Fig. 3, each hollow and solid node is the corresponding pairing point of each element of two time series, and it is also the node in the distance mapping graph between time series. The line segment represents the path that can calculate the cumulative distance, and the number on the line segment represents the corresponding weight of the path. By designing different available paths and assigning different path weights, different optimal alignment paths and different time series similarity can be obtained. "symmetric1" and "symmetric2" modes can get a continuous alignment path, while "asymmetric", "rabinerJuang" and "symmetric5" modes can get a discontinuous alignment path. This "discontinuity" means that elements in one time series have no aligned elements in another time series. This discontinuous alignment path actually increases the tolerance of time series non stationarity.

It is incorrect to pursue only morphological similarity between two stock price time series while calculating stock linkage. Owing to the strong timeliness of the stock market, there is no obvious rule for the emergence of stock linkage under the influence of complex factors in the market. There can be several variations in the time of appearance and disappearance, duration, and strength of the linkage effect. The closer the time between the two stock price time series, the more similar is the shape of the price curves. Therefore, in this study, the first exponential smoothing is added to the part of the DTW algorithm that calculates the minimum cumulative distance to emphasize the influence of recent morphology on the computational similarity for predicting stock linkage in the future. This operation is shown in Eq. (7).

$$D_{i,j} = ad_{i,j} + (1-a)\min\left\{\mathrm{D_{i,j-1}}, \mathrm{D_{i-1,j}}, \mathrm{D_{i-1,j-1}}\right\} \quad = \quad 2, \ , \ \mathrm{m} \quad = \quad 2, \ , \ \mathrm{n} \tag{7}$$

According to the equation, the larger the weight coefficient, the greater the influence of recent shape on the calculation of similarity between stock price time series. In this study, a = 0.98.

In the constructed stock time series data set, the price of different stocks differs at multiple levels. To eliminate the influence of dimension difference, this study uses the Z-score method to standardize the price of each stock and improves the effect of the time-weighted DTW algorithm in finding the optimal alignment path and calculating time series similarity, as shown in Fig. 4.

Compared to Pearson correlation coefficient, the time-weighted DTW algorithm can not only mine the similar information of time series in morphology but also tackle the continuity

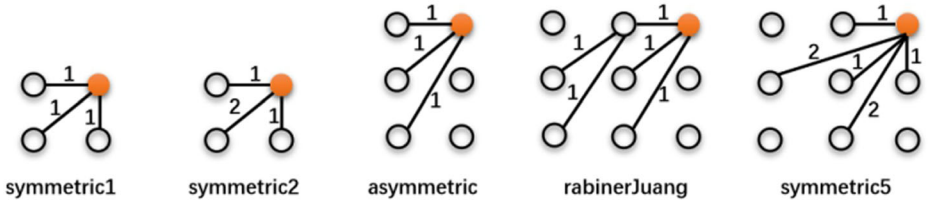

Fig. 3 Comparison of stock price trends [19]

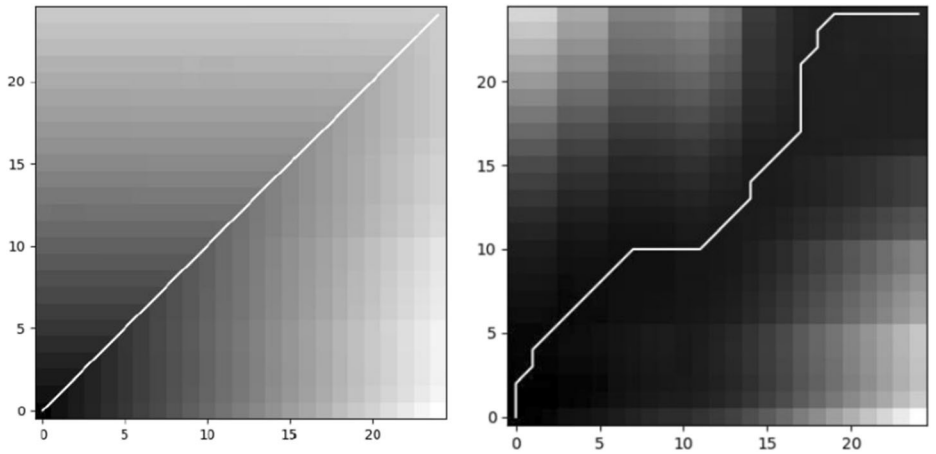

Fig. 4 DTW path planning results before (left) and after (right) Z-score

and lag of the stock linkage phenomenon and emphasize the impact of recent stock market changes on the stock linkage.

2.3 Numerization algorithm of stock linkage

The Pearson partial correlation coefficient considers the influence of market environment factors on the numerical correlation of stock price series. It includes the numerical correlation characteristics of stock linkage. Time-weighted DTW similarity can be used to represent morphological similarity between sequences with different time offsets and lengths.

Specifically, the Pearson partial correlation coefficient is more sensitive to the range of volatility in stock price time series. It is more suitable to measure the correlation of the gentle trend part of the sequence. Time-weighted DTW similarity measures the similarity between sequences directly from the morphological point of view. It emphasizes the consistency of wave directions; however, the difference contribution of relative fluctuation range is insufficient.

We use the method of time weighted dynamic time warping algorithm combined with Pearson partial correlation coefficient to analyze the numerical correlation and morphological similarity of time series, so as to transform the problem of linkage relationship mining among stocks into the problem of linkage value prediction among stocks. The time weighted dynamic time warping algorithm is used to process the stock price time series data. The optimal alignment path distance is converted into similarity and combined with Pearson partial correlation coefficient. The hybrid expression is used as the numerical expression of stock linkage. Therefore, considering the continuity and lag of stock linkage phenomenon more comprehensively, it is more suitable for the quantification of stock linkage than a single method.

For stocks i and j, the time-weighted DTW distance of the time series can be transformed into DTW similarity, $s_{i,j}$, using Eq. (8), and it can be linearly combined with Pearson partial correlation coefficient, $\gamma_{\mathbf{i,j(h)}}$. Then, the numerical expression of linkage considering the stock market environment factor, h, can be obtained as shown by Eq. (9).

$$s_{i,j} = \frac{1}{1 + d_{i,j}} \tag{8}$$

$$c_{i,j} = \alpha_1 \cdot s_{i,j} + \alpha_2 \cdot \gamma_{i,j(h)} \tag{9}$$

3 Establishment of stack linkage prediction model based on optimized LSTM

Based on LSTM model, this paper takes the characteristics of stock price and transaction scale as the description attributes of stocks, and as the characteristics of constructing input time series, so as to predict the linkage changes between stocks in the future. The modeling process of stock linkage prediction based on LSTM model includes four parts: the construction of model training samples, the detailed design of neural network model, model training optimization and avoiding over fitting, model effect improvement and scalability.

3.1 Construction of training samples

To merge the time series data of two stocks to construct input that is acceptable to the model, the difference construction method is used. In other words, while predicting the stock linkage between stocks A and B for a future time period, the input samples are constructed by using the difference between the two series. The time series structure of the sample is illustrated in Fig. 5.

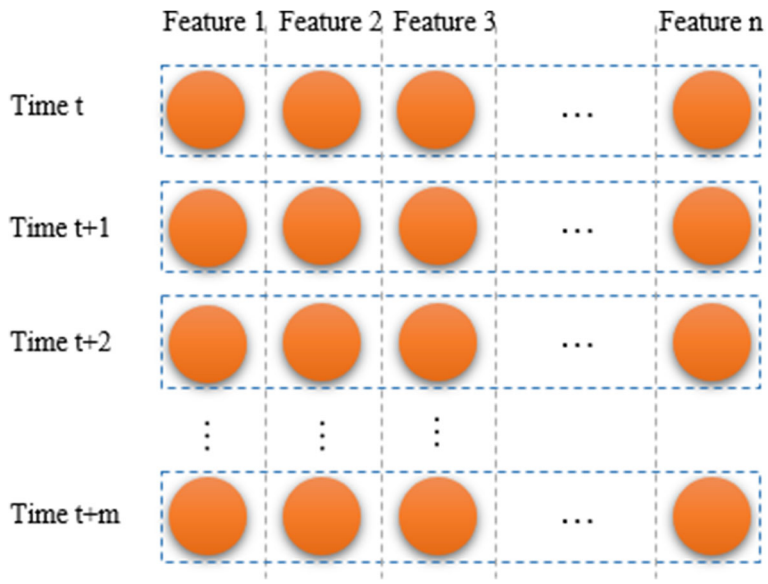

Fig. 5 Structural diagram of input sample time series

In this figure, each column corresponds to a feature time series of stocks, e.g., stock closing price difference time series, stock trading volume difference time series. Input samples are listed as rows for a given time. Each sample comprises multiple feature dimensions. Each input sample corresponds to a linkage value at the same time. This value is a linkage index calculated by using the historical data of the sample from the current time as the starting point and a certain period of time ahead as the prediction object of the model.

The training input to the model is a fixed time-length sample sequence. It comprises n features with n dimensions. The training output is the linkage value at the next moment after the end of the period, and its dimension is one. By learning the variation of stock market data in a period of time, the model extracts the variation rules to predict the possible degree of linkage in the future.

3.2 Structural design of optimized LSTM model

In the prediction task of stock linkage, the structural design of the LSTM model needs to obtain one-dimensional similarity output from multi-dimensional input samples. Input samples comprise multiple parallel sequences composed of attribute values; therefore, the sequences are independent of each other. Considering that the model may need to add more attributes in the actual environment to fully describe the relative changes between stocks, a model structure reflecting independence and association is designed.

Moreover, the financial time series contains a considerable amount of noise. Therefore, to improve the prediction performance of the model, the corresponding denoising automatic encoder module or wavelet transform module is configured as the denoising processing layer after the input layer according to the characteristics of the corresponding attributes. The structural design of the prediction model is illustrated in Fig. 6.

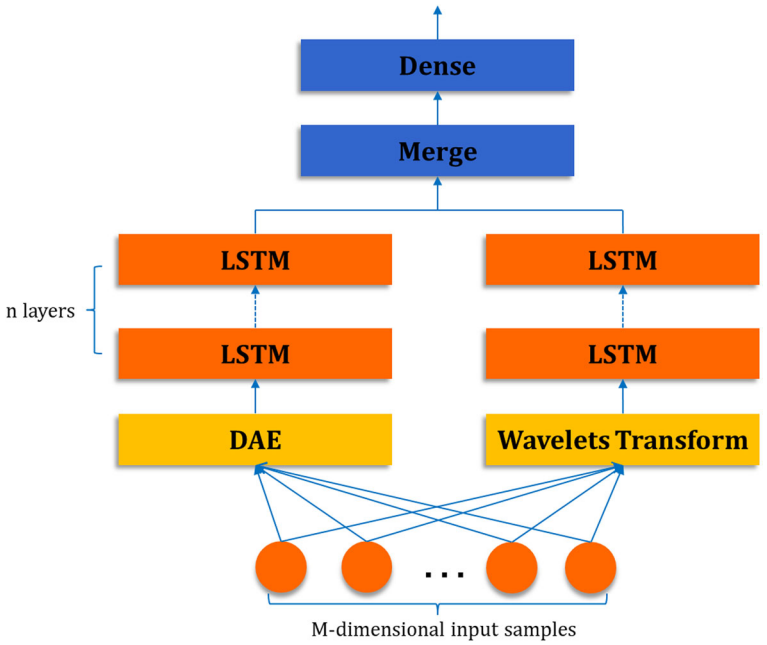

Fig. 6 Prediction model network structure diagram

In this figure, the input to the optimized LSTM model is the time series of attribute difference. Two sets of parallel feature sequences are obtained through the denoising layers of the automatic encoder and wavelet transform. These two sets of parallel feature sequences are spliced at the merge layer present after the LSTM modules with different number of layers. Then the full connection operation is completed at the dense layer. The final one-dimensional output is the predicted value of stock linkage corresponding to the M-dimensional input sample.

3.3 Model training optimization

The prediction model based on LSTM network structure is vulnerable to poor results and over-fitting in the training process. It is optimized by the gradient-updating algorithm and the regularization improvement over-fitting algorithm.

- Adam algorithm

Adam algorithm is usually used to update the gradient while training the neural network-based model. It is defined as follows.

$$m_t = \mu m_{t-1} + (1-\mu) g_t \tag{10}$$

$$n_t = \nu n_{t-1} + (1-\nu) g_t^2 \tag{11}$$

$$\theta_t = \frac{m_t^{corrected}}{\sqrt{n_t^{corrected}} + \epsilon} \eta \tag{12}$$

where,

m_t—First moment estimation of gradient,
n_t—Second moment estimation of gradient.
$m_t^{corrected}$—Corrected value of m_t is approximately the unbiased estimate of expectation.
$n_t^{corrected}$—Corrected value of n_t is approximately the unbiased estimate of expectation.
η—Learning rate of the model.

- Dropout method

The dropout method of the cyclic neural network model is slightly different from other neural networks. It does not cause random inactivation of neurons in the circulatory structure to artificially damage the data because with LSTM, it is easy to magnify error data. As illustrated in Fig. 7, to achieve the regularization effect, the output of each module can be randomly dropped at the full connection layer of the LSTM model in a certain proportion. The structure can also be improved so that when data flows through the LSTM modules at different layers, it maintains its original state in the forward direction of the time series; however, the dropout operation is performed randomly in the direction of the LSTM modules at different layers.

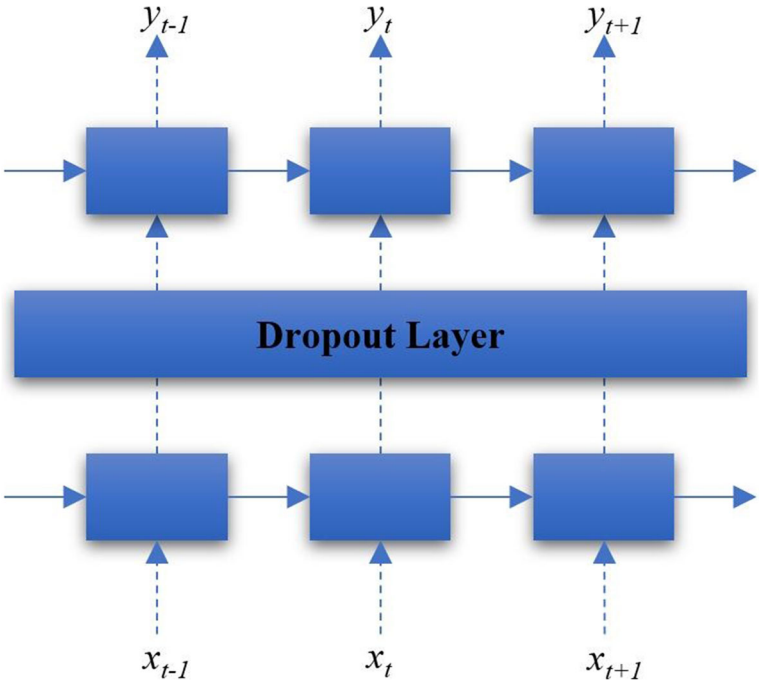

Fig. 7 Dropout regularization of multilayer LSTM modules

4 Experimental results and analysis

4.1 Data acquisition and preprocessing

According to the weight of each component of the Shanghai and Shenzhen 300 index, we selected the first 100 stocks as the research object of this paper. Among the 100 sample stocks, there were 44 sub sectors, accounting for 17% of the banking industry, accounting for 17% of the total, followed by the securities, insurance and Baijiu industries. At the same time, there are 20 regions, of which Beijing accounts for the highest proportion, accounting for 27%, followed by Shanghai, Shenzhen and Jiangsu respectively. 60% of the regions contain less than 3 sample stocks. These differences in industry and geographical distribution also reflect the current situation of China's economic environment to a certain extent. As the stock market situation changes over time, too long stock data are not obviously helpful to the study of the current and future stock market. Therefore, the nearly two years from July 1, 2016 to June 30, 2018 are selected as the time span of the study.

This paper uses five different data sources to collect the stock price time series data and the basic information of the stock. Data acquisition from tushare platform (http://www.tushare.org). The data interface is mainly obtained from Sina Financial platform (https://finance.sina.com.cn), TongHuashun platform (http://www.10jqka.com.cn), Shenzhen Stock Exchange (http://www.szse.cn) and Shanghai Stock Exchange (http://www.sse.com.cn). The crawled data are compared in many ways. When there are differences, Voting and manual review are adopted to ensure the correctness and consistency of stock time series data set and stock basic information. To sum up, the start and end dates of the stock price time series data collected in

this paper are July 1, 2016 and June 30, 2018 respectively, including 487 trading days, covering 100 sample stocks, and a total of 190,881 trading market data.

Finally, the stock data are preprocessed by unified format, post complex weight processing and interpolation. The processed stock price time series data ensure the time continuity of the data and the consistency of the time scale.

4.2 Experimental setup

The environment configuration for the contrastive experiments on the stock linkage prediction model based on LSTM is as follows.

1) Ubuntu 16.04 LTS(64 bit).
2) Python 2.7.15.
3) TensorFlow, TensorBoard, Keras, Sklearn, NumPy.

In this study, Pearson partial correlation coefficient and time-weighted DTW distance between stock sequences are calculated by three time windows of 3, 5, and 20 days. The time-weighted DTW distance with the highest co-direction fluctuation ratio and the time window size of 20 days is chosen as the numerical expression of stock linkage. One-day market data comprising 190,900 stocks, after the two groups of 100 stocks made a difference, finally obtained a one-day market difference sequence of 4,950 stocks, a total of 23,009,000 stocks, and time-weighted DTW distance labeling for each market difference. Stocks 000001 and 000002 are used as examples, and the time-weighted DTW distance between them from July 01, 2016 and June 30, 2018 are presented in Fig. 8.

As shown in Fig. 8, there have been four obvious linkage phenomena between the two stocks in the entire time span. The duration of the phenomenon is 10.3% of the total, i.e., approximately 48 trading days, with an average duration of 10 trading days. It can be seen that there may be linkage phenomenon in the short run even between unrelated stocks in general.

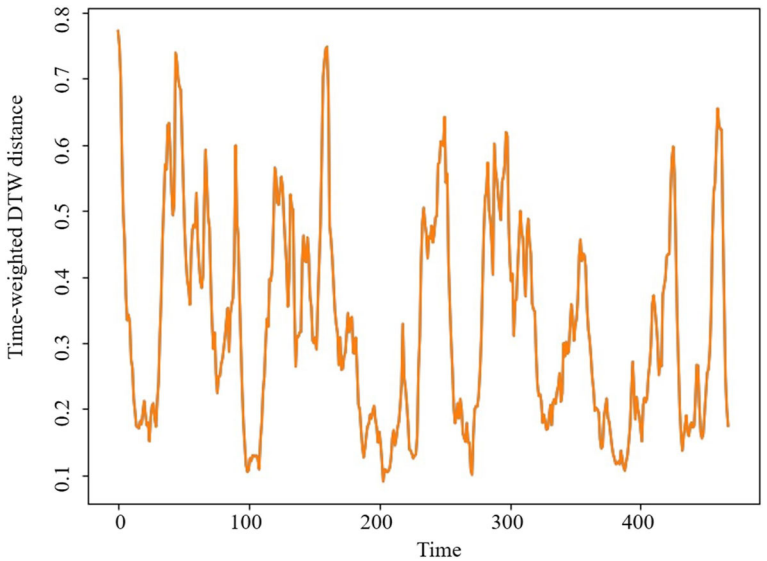

Fig. 8 Time-weighted DTW distance between Stocks 000001 and 000002

Timely prediction of such linkage changes can be useful to avoid possible risks and risk concentration.

To determine the number of layers and the size of hidden units in LSTM module, prediction models of LSTM module with 1, 2, and 3 layers are implemented. Rectified linear activation unit (ReLU) is adopted as the activation function in all models, and the dropout ratio is set to 0.1. Data of 100 trading days are used to train the model and predict the linkage of 20 trading days in the future.

Stock linkage prediction is a regression problem; therefore, the experimental results can be evaluated by root mean square error (RMSE), mean square error (MSE), and mean absolute error (MAE). These three evaluation criteria can be calculated as follows.

$$\mathrm{RMSE} = \sqrt{\frac{1}{\mathrm{N}} \sum_{\mathrm{t}=1}^{\mathrm{N}} (\mathrm{o_t} - \mathrm{y_t})^2} \tag{13}$$

$$MSE = \frac{1}{N} \sum_{t=1}^{N} (o_t - y_t)^2 \tag{14}$$

$$MAE = \frac{1}{N} \sum_{t=1}^{N} |o_t - y_t| \tag{15}$$

where,

N—Size of test set.

o_t—Observation value of t-time.

y_t—The predicted value of t-time.

4.3 Experimental results

The experimental results of the optimized model proposed with different number of layers of LSTM modules are presented in Table 2.

Table 2 shows the comparison experimental results of prediction models of different LSTM module layers. The increase of LSTM module layers cannot steadily improve the prediction performance of the model. Through the comparison of the prediction results of single-layer, double-layer and three-layer models, at present, the double-layer LSTM model has achieved the best results. Compared with the single-layer LSTM model and three-layer LSTM model with the same characteristic number, the double-layer LSTM model has higher prediction

Table 2 Error comparison with different number of layers of LSTM models

Number of layers	RMSE	MSE	MAE
1	0.207	0.045	0.164
2	0.183	0.031	0.142
3	0.219	0.047	0.178

performance. Therefore, in the stock linkage prediction experiment, the prediction model uses two-layer LSTM module. The reason why the double-layer LSTM model achieves better results is that it contains more parameters than the single-layer LSTM model in the current data scale, has stronger nonlinear expression ability, and learns stronger sequence features through dropout and other methods. However, the network complexity of the three-layer LSTM model is higher, but it has the opposite effect. It can be seen that the more the network layers, the better.

Furthermore, quality of the input samples has a great impact on the prediction performance of the model. To mitigate the problem of large amount of noise in the input samples, this study adds a module to denoise input samples on the basis of the LSTM basic model. The final optimized LSTM model is constructed by combining the wavelet transform with DB4 wavelet basis with the denoising automatic encoder.

As shown in Fig. 9, the high-frequency noise in the input time series of attribute difference is significantly reduced by using the wavelet transform with DB4 wavelet basis. The fluctuation curve becomes smoother, which is beneficial for the LSTM model to make robust predictions.

As shown in Fig. 10, the noise reduction automatic encoder designed in this study uses a three-layer network structure that comprises 20, 16, and 20 hidden units in the ratio of 5:4:5. Here, ReLU is used as the activation function, and adding 5% Gaussian noise to the input sequence creates information loss to re-learn and construct more expressive input features.

To verify the predictive performance of the optimized LSTM model for the numerical time series of inter-stock linkage, three different types of comparative experiments were conducted in this study.

The first group of experiments is with the basic reference model, including the traditional LSTM model and the autoregressive integrated moving average (ARIMA) statistical model.

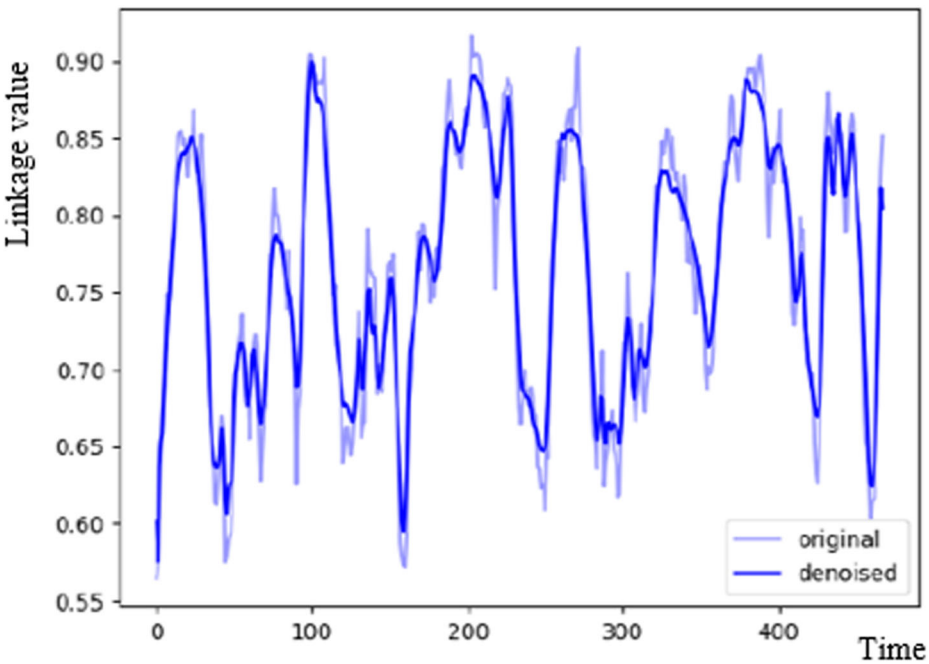

Fig. 9 Comparison of denoising results of linkage value series

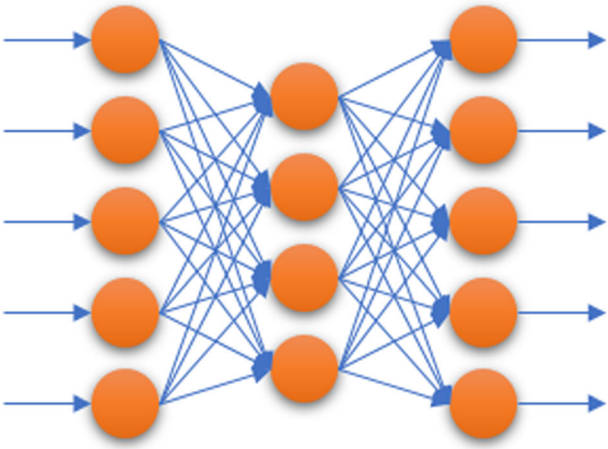

Fig. 10 Network structure of noise reduction automatic encoder

The ARIMA model is trained by using the first-order differential sequence with stationarity. According to the calculations of autocorrelation function (ACF) and partial autocorrelation function (PACF), the Akaike information criterion (AIC) is used as the primary evaluation criterion. Finally, the first-order autoregressive (AR) and 0-order moving-average (MA) models with the best selection effect are determined for the experiments.

The second group of experiments involves a comparative model that uses the wavelet transform technology to denoise the input sequence. It includes the deep bidirectional ARIMA (DB-ARIMA) and the deep bidirectional LSTM (DB-LSTM) models that have achieved good results in stock-price trend prediction. Both models use the DB4 wavelet basis.

The third group of experiments involves the reconstruction of the contrast model of input sequence by using automatic encoder. Furthermore, the stacked denoising autoencoder (SDAE)-LSTM model for reconstructing input samples is included [15]. In the task of stock index trend prediction, the four-layer network structure with 16, 8, 8 and 8 hidden units respectively achieves the best prediction performance. It is verified that the structure also achieves the optimal results under the model in the stock linkage forecasting task.

In this study, the data set was divided according year, and 20% of the transaction-day data at the end of each year were used as test set to verify the performance of the prediction models. RMSE, MSE, and MAE were used as the performance criteria of the prediction models. The final experimental results are the mean values of five repeated experiments for each year. The comparative experimental results of the prediction model are shown in Table 3.

Table 3 Comparison of errors in prediction models

Models	RMSE	MSE	MAE
LSTM	0.156	0.028	0.123
ARIMA	0.555	0.308	0.493
DB-ARIMA	0.356	0.127	0.317
DB-LSTM	0.091	0.009	0.075
SDAE-LSTM	0.138	0.019	0.113
Optimized LSTM	0.074	0.006	0.059

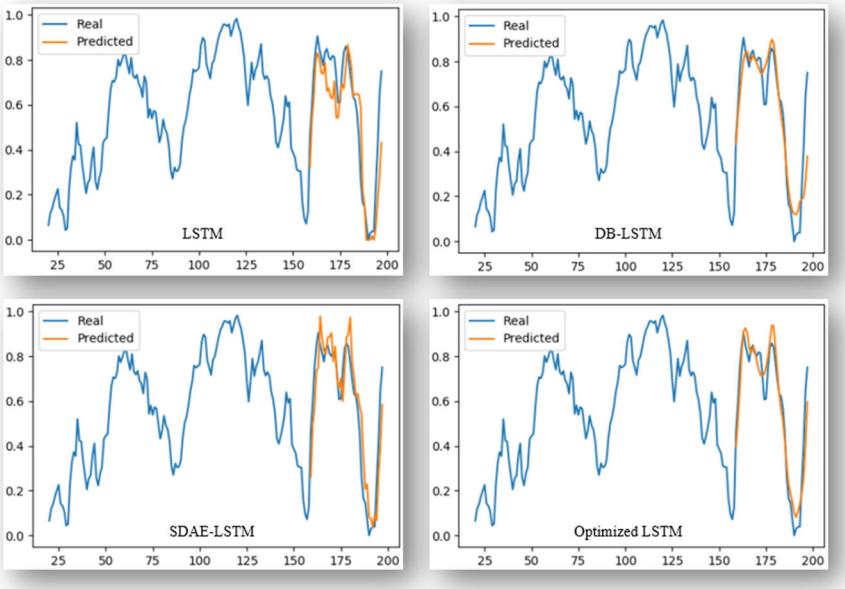

Fig. 11 Contrast charts of prediction results of models

Table 3 shows that noise reduction in input samples can effectively improve the performance of the prediction models. The prediction error, RMSE, of the ARIMA model is reduced by 35.86% through DB noise reduction. The RMSE error of the LSTM model is reduced by 11.54% after SDAE reconstruction of the input samples and by 41.67% after DB denoising. Wavelet transform with DB4 wavelet basis can effectively smooth the linkage numerical curve, maintain the trend of linkage while removing most of the noise, and effectively improve the prediction performance of the model. Quality of the reconstructed samples based on SDAE input improved after greedy layer-by-layer training, increasing Gaussian noise, and random zeroing operation; however, the generalization ability reduced owing to the increased complexity of the network. Compared with the SDAE-LSTM model, the optimized LSTM model uses simplified structure to reconstruct input samples. The reconstructed new feature sequence and the smoothing trend feature obtained by the wavelet transform are used as training materials of the model, and a relatively improved prediction performance is obtained. The RMSE error of the optimized LSTM model is 18.68% lower than that of the DB-LSTM model and 46.38% lower than that of the SDAE-LSTM model. The predictive effect of some models on the test set is presented in Fig. 11.

5 Conclusions

This paper presents a detailed research on the prediction of numerical stock linkage using deep learning model. Starting with the numerical correlation and morphological similarity of time series, we used a time-weighted dynamic time-regularization distance with emphasis on time-effect as a numerical expression of stock linkage. Consequently, the problem of discovering

the linkage relationship between stocks can be transformed into the problem of numerical prediction of the linkage between stocks.

Combining the noise reduction automatic encoder and the wavelet transform modules to act as the noise reduction processing layer, we proposed an optimized LSTM model to predict the stock linkage based on the model and the numerical sequence of stock linkage. The performance of the prediction model was verified by a number of comparative experiments, and the factors affecting the performance of the model were analyzed in detail.

Authors' contributions Not applicable.

Funding This paper is supported by the Foundation of Guangdong Educational Committee under the Grant No. 2018KTSCX218, No. 2021ZDJS082 and the Professorial and Doctoral Scientific Research Foundation of Huizhou University under the Grant No. 2018JB020.

Data availability Not applicable.

Code availability Not applicable.

Declarations

Conflicts of interest/Competing interests The authors declare that they have no conflicts of interest.

References

1. Ahlgren P, Jarneving B, Rousseau R (2014) Requirements for a cocitation similarity measure, with special reference to Pearson's correlation coefficient[J]. J Assoc Inform Sci Technol 54(6):550–560
2. Akita R, Yoshihara A, Matsubara T et al (2016) Deep learning for stock prediction using numerical and textual information[C]. IEEE/ACIS International Conference on Computer & Information Science. IEEE. https://doi.org/10.1109/ICIS.2016.7550882
3. Al-Augby S, Majewski S, Nermend K et al (2016) Proposed investment decision support system for stock exchange using text mining method[C]. Al-Sadeq International Conference on Multidisciplinary in IT and Communication Science & Applications. https://doi.org/10.1109/AIC-MITCSA.2016.7759917
4. Arshanapalli B (1993) International stock market linkages: Evidence from the pre- and post-October 1987 period[J]. J Bank Financ 17(1):193–208. https://doi.org/10.1016/0378-4266(93)90088-U
5. Chen JF, Chen WL, Huang CP et al (2016) Financial time-series data analysis using deep convolutional neural networks[C]. 2016 7th International Conference on Cloud Computing and Big Data (CCBD). https://doi.org/10.1109/CCBD.2016.027
6. Cheng SH (2014) Predicting stock returns by decision tree combining neural network[M]. Intelligent Information and Database Systems. https://doi.org/10.1007/978-3-319-05458-2_37
7. Day MY, Lee C (2016) Deep learning for financial sentiment analysis on finance news providers[C]. 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). ACM. https://doi.org/10.1109/ASONAM.2016.7752381
8. Dixon M, Klabjan D, Bang JH (2016) Classiffication-based Financial markets prediction using deep neural networks[J]. Algorithmic Finance. https://ssrn.com/abstract=2756331
9. Heaton JB, Polson NG, Witte JH (2017) Deep learning for finance: deep portfolios[J]. Applied Stochastic Models in Business and Industry 33(1):3–12
10. Morel M, Achard C, Kulpa R et al (2018) Time-series averaging using constrained dynamic time warping with tolerance[J]. Pattern Recognit 74:77–89. https://doi.org/10.1016/j.patcog.2017.08.015
11. Nieto B, Rodríguez R. Correlation between Individual Stock and Corporate Bond Returns[J]. Social Science Electronic Publishing. https://doi.org/10.2139/ssrn.2386043

12. Okamoto T, Iida D, Toge K et al (2016) Optical correlation domain reflectometry based on coherence synchronization: theoretical analysis and proof-of-concept[J]. J Lightwave Technol 34(18):4259–4265. https://doi.org/10.1109/JLT.2016.2590507
13. Pan MS, Fok CW, Liu YA (2007) Dynamic linkages between exchange rates and stock prices: Evidence from East Asian markets[J]. Int Rev Econ Finance 16(4):503–520. https://doi.org/10.1016/j.iref.2005.09.003
14. Patton AJ (2012) A review of copula models for economic time series[J]. J Multivar Anal 110(none):4–18
15. Shao L, Cai Z, Liu L et al (2017) Performance evaluation of deep feature learning for RGB-D image/video classification[J]. Inf Sci 385-386:266–283. https://doi.org/10.1016/j.ins.2017.01.01
16. Sohangir S, Wang D, Pomeranets A et al (2018) Big Data: Deep Learning for financial sentiment analysis[J]. J Big Data 5(1):3. https://doi.org/10.1186/s40537-017-0111-6
17. Thomas S (2011) Dynamic linkages between fundamental drivers and stock market: an empirical analysis from India[J]. Social Science Electronic Publishing.https://ssrn.com/abstract=1853003
18. Wei B, Jun Y, Yulei R et al (2017) A deep learning framework for financial time series using stacked autoencoders and long-short term memory[J]. PLoS ONE 12(7):e0180944. https://doi.org/10.1371/journal.pone.0180944
19. Zhao J, Itti L (2018) shapeDTW: Shape dynamic time warping[J]. Pattern Recognit 74:171–184