IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 7, 2020, accepted May 12, 2020, date of publication May 20, 2020, date of current version June 4, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2995958

Characteristic Representation of Stock Time Series Based on Trend Feature Points

MENGNA ZHOU[1], JIZHENG YI[1], (Member, IEEE), JIEQIONG YANG[2], AND YI SIMA[1]
[1]College of Computer and Information Engineering, Central South University of Forestry and Technology, Changsha 410000, China
[2]College of Life Science and Technology, Central South University of Forestry and Technology, Changsha 410000, China

Corresponding authors: Jizheng Yi (kingkong148@163.com) and Jieqiong Yang (417588978@qq.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61602528, in part by the Hunan Provincial Natural Science Foundation of China under Grant 2017JJ3527, in part by the Scientific Research Fund of Hunan Provincial Education Department under Grant 16B275, in part by the Grant of China Scholarship Council (CSC) under Grant 201808430002, and in part by the Research Foundation for Advanced Talents of Central South University of Forestry and Technology under Grant 2015YJ013.

ABSTRACT Stocks are the most active part of the securities market, and the analysis of stock generally starts from the price fluctuation. Stock trading data have the characteristics of time series, which make it possible to record the transaction prices in a time-evolving order. Due to the large data and high research complexity of time series, some ideal data are difficult to obtain for analyzing and predicting stock movements. Aiming at the problem, we utilize the piecewise linear representation which combines turning points and maximum absolute deviation points in stock time series to extract sequence features. Firstly, the proposed method finds turning points that satisfy the condition given in this paper, and defines the point distance formula to calculate the fitting errors between the subsequence segment and the fitted straight line, whose average value is set as the threshold P and the subsequence length is as the threshold d. Secondly, if the fitting error or the length of the subsequence segment is greater than P or d, respectively, the maximum absolute deviation point is obtained according to the difference between the fitted value and the original data. Finally, all trend feature points are connected by linear interpolation. In this paper, different sequence lengths, different thresholds d, different methods and industrial data from different fields are discussed and compared in detail to verify the proposed method. The experimental results show that the proposed method gets satisfactory data fitting and expanding effect, and retains the characteristics and integrity of the initial time series.

INDEX TERMS Data mining, piecewise linear representation, stock time series, trend feature point.

I. INTRODUCTION

A. BACKGROUND

The securities time series refers to the data of the futures, stocks and other valuable securities in the stock exchange market as the stock price changes with time, and it is affected by various aspects [1]–[3]. All kinds of major events in the society may affect the trend of economic development, which in turn causes stock prices to fluctuate in the trading market [4]–[6]. Therefore, the stock is the most active part in the securities market. Stock trading activity is affected by the data generated for daily trading, and stock investors also buy and sell stocks according to the economic rules reflected by these data. How to effectively utilize these data to obtain valuable information and provide scientific guidance for investors has become a hot research topic [7]–[10].

Since the stock time series has the characteristics of large data volume, high data dimension and fast update, direct data mining or similarity measurement on the initial time series is not only computationally intensive, but also highly complex, and affects the reliability and effectiveness of experimental results. Researchers generally adopt the feature representation method to preprocess the stock time series. The existing feature representation methods are based on time domain representation [11], [12], feature domain representation or model representation [13]–[16]. The first method is the simplest one of representing the time series using a minimum operation, which directly represents time series data according to the processing of the time domain signal. This method preserves the initial state characteristics of the time series, such as time wave analysis and probabilistic analysis [17]. It has

The associate editor coordinating the review of this manuscript and approving it for publication was Alberto Cano.

the advantages of being easy to implement and preventing the loss of time domain information. The second one is to project the original time series into a new feature space. The new feature space can be divided into segmentation representation and global representation according to the feature classification criteria [18], [19]. The last one is the application of mathematical models to generate sequence data based on time series. The classical time series feature representation models include the auto regressive moving average model [20], [21], the first-order Markov hybrid chain [22], the hidden Markov model and dynamics Bayesian network etc., [23].

In addition to the data analysis mentioned above, we also find that some simulation methods based on entropy, Poisson mixture and agent have good effects on data analysis. Ponta and Carbone [24] proposed a piece segmented interpretation information based on the intersection of random sequences and moving averages, which was novel in that it divided the inherent informative/uninformative clusters along the sequences and better distinguish the sequence features. Teng and Shang [25] proposed a new transfer entropy coefficient to quantify the level of information flow between financial time series in view of the complex variability of transfer entropy, so as to analyze which stocks dominated the market, providing a reference for the analysis and development of stock market data. Stock series are random sequences that change according to a certain time, and the Poisson mixture itself is a random order distribution. Therefore, it is widely used in risk markets such as stock market. Ponta *et al.* [26] studied tick-by-tick financial returns of the FTSE MIB index of the Italian Stock Exchange and proposed a simple non-stationary return model based on a non-homogeneous normal compound Poisson process. Compound Poisson assigns a certain probability value to a random variable. The authors utilized the property to approximate financial data, and used three information criteria to analyze and compare its effects. It was found that the compound Poisson was most effective when the parameters were small. This study provided an alternative for the analysis of high-frequency financial time series. The essence of the agent-based simulation method is to simulate the process of market changes and provide guidance for decision-makers and managers. In [27], the authors utilized different types of stocks as characteristics and information multi-asset artificial stock market composed of heterogeneous agents to analyze the influence of the agent network on the market structure. The characteristics of agents included cash, stock, and emotion. The results showed that the agent-based simulation method is of great significance in the analysis of stock market data.

The securities industry is resource-intensive and knowledge-intensive, and the analysis of the stock market has been a major challenge for this industry, in order to overcome the problem, many experts at home and abroad have explored many methods and theories for it. The emergence of data mining technology has become an important technical means of stock analysis [12], [28], [29], providing lots of novel ideas, mainly for customer analysis, financial indicator analysis, transaction data analysis and investment analysis. In this paper, the feature representation of the stock time series is usually the mining of data information, so it belongs to the field of data mining. The mining of stock market information has certain significance not only for investors but also for social and economic development.

B. RELATED WORK

The time series refers to the sequence of corresponding changes with time. For the characteristics of the time series itself, such as high dimension and large scale, Faloutsos *et al.* [30] believe that when the sequence dimension decreases, the mining effect is more obvious. In recent years, researchers [31], [32] have proposed many new ideas and methods for the preprocessing of time series dimensionality reduction. Generally, there are two kinds of dimensionality reduction preprocessing methods for time series, which are feature extraction and feature representation.

The feature extraction of time series is to select an optimal feature subset that reflects the initial time series by some algorithms. Feature extraction of time series $T=\{t_1, t_2, \cdots, t_n\}$ is to find the subset $Q=\{q_1, q_2, \cdots, q_m\}$ that best reflects its characteristics, where Q belongs to T and m is much smaller than n.

The feature representation of the time series is based on the preservation of the initial time series features, whose goal is to reduce the sequence dimensions. A characteristic representation of a time series $T=\{t_1, t_2, \cdots, t_n\}$ is the transformation of the sequence into a low dimensional time series $Q = \{q_1, q_2, \cdots, q_m\}$.

In view of the processing methods for the time series mentioned above, domestic researchers have already done many works. Yan *et al.* [33] proposed a time series segmentation method based on local maximum and minimum points by combining some important points including the extreme points. The maximum and minimum values were obtained according to the extremum function, and the first and last endpoints of the original sequence were added to fit them. In their experiment, the authors utilized different datasets to demonstrate the effectiveness of the method. Yin *et al.* [34] studied a novel segmentation method based on the turning points. Turning points in their method were defined as the points extracted from the maximum or minimum of the time series. The segmentation level was mainly two layers. The first layer selected the maximum point and the minimum point, and the second layer eliminated some unimportant points according to four given strategies. At the same time, in order to combine a small range of trends into a large trend, some local turning points were discarded. The results showed that the segments generated by this segmentation method could retain more trends of sequences. Si and Yin [35] proposed a segmentation method based on the inflection point for the different importance of turning point to time series. Their method evaluated the importance by the tendency and shape in the sequence, and stored the turning points in Optimal Binary Search Tree

(OBST), which reduced the average retrieval cost. The results showed that their method kept the lowest average retrieval cost while retaining more trends. Luo *et al.* [36] proposed a piecewise approximation algorithm with max-error guarantees based on the given time series and the specified error. Their method firstly constructed a piecewise linear function to judge whether the error between the function and the sequence met the conditions and then selected the sequence that satisfied the conditions as an approximate representation of flow data trend. Finally, an online algorithm was designed to generate the optimal piecewise approximate representation with the maximum error guaranteed. Lin and Wang [37] researched a time series segmentation algorithm based on first-order filtering. Their method is to prevent the segmentation algorithm for edge extraction of slope falling into the local optimum under the situation of drastic slope fluctuation. In their method, firstly, the first and last points of the sequence, the relative extreme points obtained from the midpoint, the points that met the change of slope, and the rest points were divided into four levels of points. Among them, the most significant was to add the first-order filter smoothing sequence burr when extracting extreme points to reflect the basic trajectory of the sequence. Then, the priority queue was used to realize the classification storage of different important points. And finally, the final sequence was obtained according to the compression rate. The experimental results showed that the method got good quality for sequence fitting with small slope variation. For the piecewise linear representation proposed by Keogh [38], the more subsequence segments are divided, the more basic feature information is extracted. On the contrary, the extracted basic feature information of the sequence is less. The extraction of these key points is not only studied in time series analysis, but also in other aspects, such as face recognition, which is often used for positioning [39], [40].

C. OUTLINE OF PAPER AND APPROACH

As can be seen from the research about the piecewise linear representation of the stock time series in the previous section, the basic features of the initial time series should be retained as much as possible. Therefore, how to achieve a better modeling effect of time series under the condition of obtaining a higher compression rate is the key to the current time series feature representation. In this paper, we mainly focus on the research of the stock time series. The existing piecewise linear representation cannot retain the characteristics of the initial time series well, so we propose an improved piecewise linear representation method based on searching the trend feature points. By the proposed method, we could also obtain the trend fluctuation feature points in the local range. On the basis of the piecewise linear representation, the improved scheme of trend feature points searching can effectively extract and discriminate the initial time series trend fluctuation information, thus providing the necessary data for the securities investment.

In this paper, we first introduce the judgment method of time series trend fluctuation point, which is robust to find the segmentation point on the initial time series. Then, we connect the found segment points by the straight-line interpolation, that is, the fitting of the preliminary feature points is realized. Finally, according to the search principle of the fitting error and the trend fluctuation feature points, the feature points are searched again for the subsequence segments that satisfy the defined requirements.

The structure of this paper is arranged as follows: Section II reviews the theoretical basis of time series. Section III analyzes stock time series data. Section IV introduces the algorithm flow and the improved method of piecewise linear representation. Section V describes the experimental results analysis, and Section VI summarizes the conclusions.

II. THEORETICAL BASIS OF TIME SERIES

With the rapid development of science and technology, the ability of computer is becoming more and more powerful, and the data stored in computers are pretty huge, such as the basic information of students stored in the school management system, the large amount of stock data stored in the securities company system and so on. Many of the data in these systems are listed in a chronological order which is the time series [41]–[43]. The researches on time series mainly focus on how to find out the intrinsic connection of things based on the found sequences, and draw more valuable information from these sequences [44]–[46]. This section briefly introduces the basic concepts of time series, time series feature representation and some models of time series analysis, providing a theoretical basis for the next work.

A. TIME SERIES ANALYSIS

The definition of time series analysis has been presented in much literature, although it is expressed in different ways, the basic theory is same. With the passage of time, the creative methods and technologies about time series are constantly emerging, so we can discover potential information more effectively for solving problems related to time series analysis. Time series analysis is based on the time series data from relevant systems, and then establishes the mathematical model by parameter estimation and curve fitting. It includes the general statistical analysis, the creation and identification of statistical analysis models, and the prediction and utilization of relevant time series. Generally, time series analysis includes two contents: frequency domain analysis and time-domain analysis. For the first one, frequency, phase and amplitude are usually used to extract some meaningful features of time series. For the second one, it mainly covers time-domain analysis methods currently available linear Auto-Regressive (AR) models [47], [48], Moving Average (MA) models [49], [50], and Auto-Regressive Moving Average (ARMA) models [51].

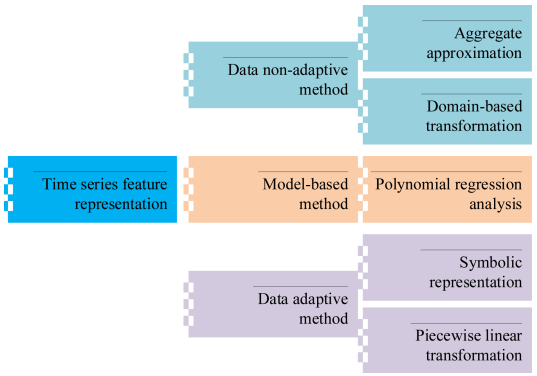

FIGURE 1. Time series feature representation.

B. CLASSIFICATION OF TIME SERIES FEATURE REPRESENTATION

Feature representation of time series refers to data dimensionality reduction, which is to convert the initial time series data from high-dimensional space to low-dimensional space, and reflects the characteristics of the initial time series as much as possible. The time series feature representation is shown in Fig. 1.

1) The data non-adaptive method refers to transforming the time series from the current dimensional domain to another dimensional domain, and the transformation process is related to the setting of the characteristic coefficients, and has nothing to do with the original data.

2) The data adaptive method is affected by the local sequence data value, and its experimental effect is affected by the overall sequence.

3) The model-based representation method, such as regression model, hidden Markov model and neural network, is applied to the time-series data for searching the optimal model parameters, and then extracts some meaningful time series features.

III. STOCK TIME SERIES DATA ANALYSIS

A. RECORDING METHOD OF STOCK DATA

The stock data is from the records of stock trading in the stock exchange market, and each transaction price and volume constitute the basis of stock data. There is no fixed time interval between each trade. There may be only a few times in an hour for the low-frequency stock trading and dozens of times in one second for the high-frequency stock trading. In order to record these data, they are generally recorded at a fixed time interval in the securities field. As shown in Table 1, this kind of recording in the securities market utilizes four prices including the opening price, the closing price, the lowest price and the highest price to represent the stock trend in a fixed time interval.

B. RAW DATA AND ACQUISITION TO STOCK TRADING

The stock trading data is generally derived from the stock exchange. There are four stock exchanges in China, which

TABLE 1. Price classification used in the securities market recording method.

Price categories	Representative meaning
Opening price	The price of the first transaction in a fixed time
Closing price	The price of the last transaction in a fixed time
Lowest price	The lowest price in a fixed time
Highest price	The highest price in a fixed time

are the Shanghai Stock Exchange, the Shenzhen Stock Exchange, the Hong Kong Stock Exchange and the Taiwan Stock Exchange. For ordinary individuals, we can obtain the data through website data such as Sina Finance and Hexun Finance's stock interface, or export data in trading software such as communication software, or by using the financial data interface package Tushare package in Python language to obtain financial data. In addition, the data can be purchased in some interfaces, such as Nezip stock interface.

C. DATA TYPES OF STOCK TRADING

Through data mining technology, we can acquire new ideas and knowledge which could be utilized to the real-time stock trading, picking or stock analysis. In order to achieve these goals, it becomes important to know how to push the data. There are two types of pushing data in stock trading software: on-demand data and push data. On-demand data refers to that stock is selected according to requirements specified by users, and then downloads the data of the stock. The advantage of this pushing mode is that it occupies fewer resources, avoids network congestion, and is conducive to the stability of the server while the disadvantage is that it is impossible to timely realize the stock picking. The full push data means that the data resources are updated in time, all the stock data are sent synchronously, and there will be no stagnation when viewed. Generally, it is sent every three seconds, which is beneficial to real-time stock picking and intraday warning. Due to the large network occupied by the push data, if the server is interrupted or the network is blocked, it is necessary to manually supplement the missing data.

IV. THE PROPOSED METHOD

This section mainly describes the algorithm for finding trend feature points and the piecewise linear representation based on trend feature points aiming at the shortcomings of current trend feature points search schemes. During finding the trend feature points, firstly, standardizing the data based on the theoretical knowledge. Then, presenting the trend feature points of the paper. Finally, analyzing experimental results. The basic frame diagram of the proposed method is shown in Fig. 2. In the extraction of the trend turning point, the slope and the set threshold are used to derive point B as the turning point. The data is segmented according to the turning points, where points O, P, Q, A, B and C are turning points. L_{O-P} in Fig. 2 represents the number of points between OP, E_{OP} represents the fitting error of line segment OP, and ME_{OQ} represents the average fitting error of line segment OQ. In

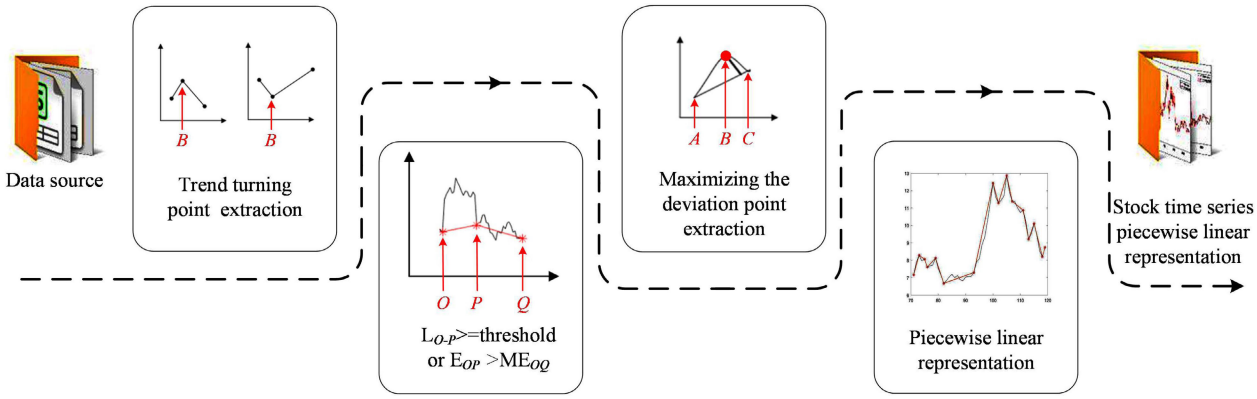

FIGURE 2. Basic frame diagram.

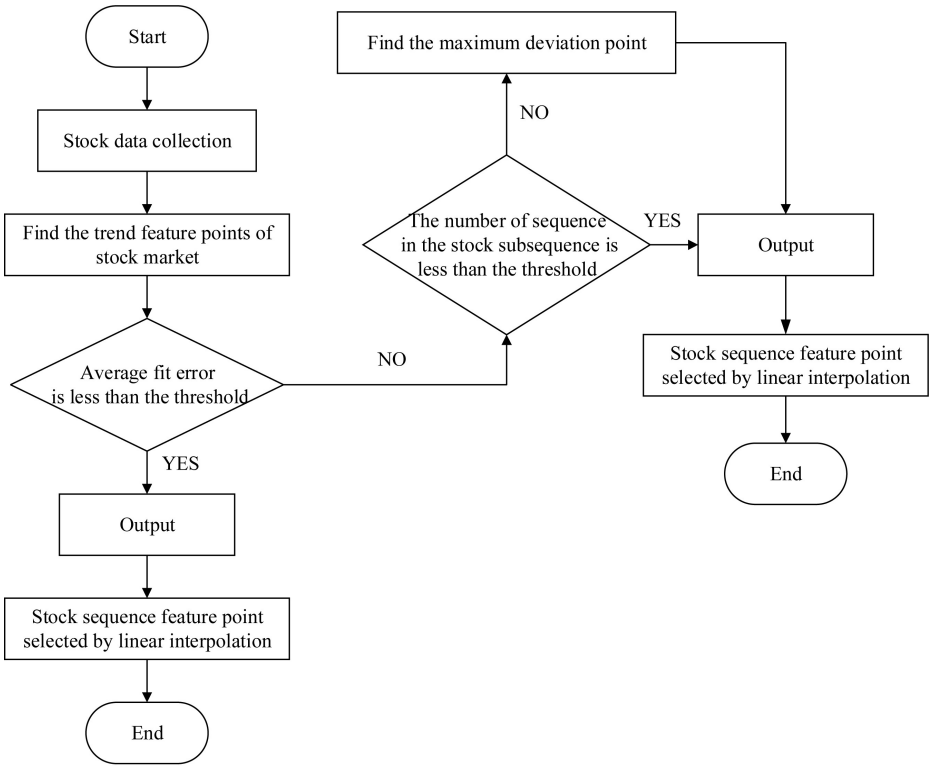

FIGURE 3. Algorithm flow char. The diagram describes the operation of the entire method, the most important of which is the search and fitting of feature points during the first and second step.

the maximization deviation point extraction, the straight line AC is the fitted line segment, the curve AC is the initial data line segment, and the point B in the curve is the maximum deviation point. Finally, the found turning points and the deviation points are combined into feature points and linearly represented.

In this paper, the piecewise linear representation of stock time series is defined as in the stock time series $Q = \{q_1, q_2, \cdots, q_N\}$, which is segmented in the form:

$$Q(t) = \begin{cases} f_1(t, w_1) + e_1(t), & t \in [1, t_1] \\ f_1(t, w_2) + e_2(t), & t \in [t_1, t_2] \\ \cdots \\ f_k(t, w_k) + e_k(t), & t \in [t_{k-1}, t_k] \end{cases} \quad (1)$$

where w_i represents the coordinates of the two endpoints of the time interval $[w_{i-1}, w_i]$, and $f_i(t, w_i)$ represents the linear function of the two endpoints in the connected mode w_i, and $e_k(t)$ is the error between the stock time series and its segmentation mode.

A. THE ALGORITHM FLOW

The algorithm flow is shown in Fig. 3. This paper realizes the feature representation of stock time series data based on trend feature points through two parts of feature points search. Get data from the JoinQuant trading platform and standardize the data. The most critical part of the process is the determination of feature points. This paper combines the slope and distance difference to fit the data.

TABLE 2. Description of the trend feature point lookup algorithm.

Based on the trend feature point search process:
Input: initial time series $A[N]$, threshold d.
Output: Trend feature point $D[m]$.
Using the trend feature point judgment method to find the initial time series $A[N]$, trend feature point set $D[m]$;
Define the average sequence point fitting error E, the first i segment fitting error E_i;
Define the length $length(i)$ of the subsegment of the segment;
While ($length(i) >= d \;\&\&\; E_i >= E \times length(i)$)
Continue to find trend feature points in the i paragraph using the principle of maximizing deviation points;
End

B. THE ALGORITHM DESCRIPTION FOR FINDING TREND FEATURE POINTS

The algorithm for feature points lookup is shown in Table 2. In this paper, the key of piecewise linear representation algorithm based on trend feature points is to find the piecewise points in subsequence. According to the definition of the trend feature point and the judgment method of the fitting error, it is determined whether the subsequence still needs to continue the process of finding the feature point of the trend.

Firstly, the algorithm traverses the original time series, calculates whether the slope changes between the adjacent points meet the given threshold, and finds the initial feature points. In this process, $(x_{i+1} - x_i)/(t_{i+1} - t_i)$ operation needs to be performed *n-1* times, and the time complexity for T_1 equals $O((n-1) \times f(n))$, $O((n-1) \times f(n))$ equals $O(f(n))$ and $O(f(n))$ equals $O(n)$. Thus, finally the time complexity for T_1 equals $O(n)$ in this step. Then, the fitting error of sequence points is introduced to calculate whether the fitting of subsequence segment meets the given threshold value, so as to determine whether the second step of feature point search needs to be carried out. At this step, the judgment needs to be made for n times, so $T_2 = T_1 = O(n)$. Finally, a single scan was performed to obtain all sequence points meeting the condition, $T_3 = T_2 = O(n)$. As described in this review, the time complexity of the trend feature points is $O(n)$.

When judging whether the current point is the initial feature point, two constrains need to be met. On one hand, this point is an extreme point. On the other hand, the slope values of adjacent points are different from each other and larger than the set threshold. This method can be used for online segmentation to facilitate the application of time series analysis.

C. THE MANIFESTATION OF TREND FEATURE POINTS

The piecewise linear representation replaces the initial time series with interconnected straight-line segments. The purpose is to remove noise points and some unimportant points, and to better detect the trend of the sequence change. When people observe data, they always pay attention to those special points. As shown in Fig. 4, the evolutionary trend between the two points is divided into ascending, descending and stationary. The development trend of some continuous points is evolved from the basic development model between the two points. As shown in Fig. 5, the evolutionary trend for three adjacent points is divided into nine situations, such as ''rise-up'', ''rise-stable'', ''rise- down'', ''stable-stable'', ''stable-rise'', ''stable-down'', ''down-down'', ''down-up'' and ''down-stable''.

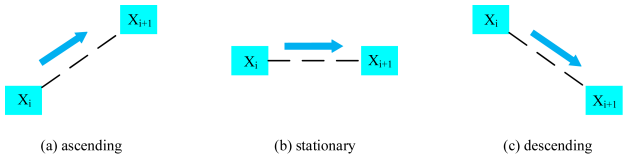

FIGURE 4. The evolutionary trend for two points in the time series.

D. TREND FEATURE POINT OF TIME SERIES

1) JUDGMENT OF TREND FEATURE POINT

a: JUDGMENT OF PRELIMINARY FEATURE POINT

As shown in Fig. 6, the feature point is not only the extreme point, but also the fluctuation range between adjacent points reaches a certain extent. In Fig. 6, a, b and c are adjacent three points, and point b is the desired trend feature point, namely X_i. The time series could be seen as an ordered set, and the elements contain the time records t_i and the value v_i. In the time series:

$$\{X = (v_1, t_1), (v_2, t_2), \cdots, (v_N, t_N)\} \tag{2}$$

the foundations for judging X_i whether it is the extreme point with a relatively sharp fluctuation range is listed as follows:

1) X_i must be an extreme point of time series, except the first endpoint X_1 and the tail point X_N;

2) Set the threshold value K, and the slopes of adjacent segments are respectively k_{ab} and k_{bc}, if $k_{ab} \times k_{bc} < 0$, and $|k_{ab} - k_{bc}| > K$ is established.

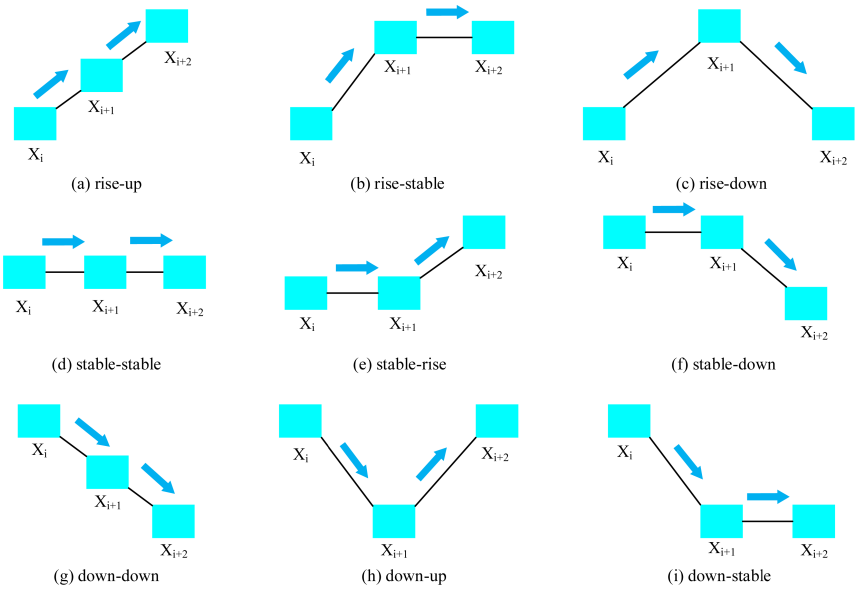

FIGURE 5. The evolutionary trend for three adjacent points in the time series. X_i, X_{i+1} and X_{i+2} are neighbors.

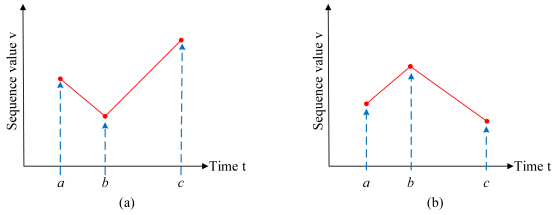

FIGURE 6. Extreme data fluctuations. Point b is the preliminary feature point.

b: TIME SERIES FEATURE POINTS AGAIN TO FIND THE BASIS FOR JUDGMENT

For the subsequence that satisfies the search condition again, calculating the difference between the fitted value of each subsequence and its original value to obtain the maximum absolute deviation point, which is the point to be found. If $y = f(x)$ is a function of the experimental fit in this paper, and the subsequence has n data points, the absolute deviation is maximized:

$$M = \max\{|y_i - f(x_i)|\} \quad (i = 1, 2, \cdots, n) \tag{3}$$

where $\{y_i\}$ are the feature points searched again in the original time series.

The sequence points determined by the above judgment method represent the trend of the stock price at this point. In this paper, the points obtained by the judgment methods (a) and (b) are defined as the trend feature points of the stock. According to this method, this paper can quickly judge the changing characteristics of any point in the stock and effectively describe the changing trend of the stock market.

2) PIECEWISE LINEAR REPRESENTATION BASED ON TREND FEATURE POINTS

The key to the segmentation of the stock sequence is the determination of the feature points. According to the definition of trend feature points, the stock time series feature representation method based on trend feature points is divided into two parts. The first part is to find the feature points according to the slope and the set threshold, and the second part is to find the feature points according to the distance difference. And then the stock time series is represented in a piecewise linear based on the feature points found.

There are two kinds of time series representation after linear segmentation: the first one is linear regression: It shows that each segment of the segment fits all the original data in the segment by least squares method, and the adjacent segments can be discontinuous. The least squares method is:

$$a = y - bx \tag{4}$$

$$b = \frac{\sum_{i=1}^{n} x_i y_i - nxy}{\sum_{i=1}^{n} x_i^2 - nx^2} \tag{5}$$

The second type is linear interpolation: It means that each line segment is a simple connection between the beginning and the end, and the adjacent segments are connected end to end. The method used in this paper is linear interpolation, and the segmentation of linear interpolation indicates that the model is continuous.

(1) Using the preliminary search principle of time series trend feature points in Section IV-D-1, determine all the trend feature points in the time series V as:

$$U = (U_1, U_2, \cdots, U_m) \tag{6}$$

m represents the number of trend feature points found in the original stock market sequence, and time series V as:

$$V = (V_1, V_2, \cdots, V_N) \tag{7}$$

N represents the number of points in the original stock time series. The subsequence containing the first and last endpoints, and the trend feature points of the initial time series is linearly represented as:

$$X = (f_1(V_1, U_1), (f_2(U_1, U_2), \cdots, f_m(U_m, V_N))) \quad (8)$$

where f represents the linear function fitted for each sequence, and the subsequence linear function values $(V_1', V_2', \cdots, V_N')$ of all the sequence points in the initial time series are obtained based on the function. By fitting the calculation formula to the fitting error, the fitting error of the sequence value after the above processing and the initial time series value is calculated as E_N.

The fitting error reflects the difference between the recorded value and the fitted value. In this paper, the residual between the original time series value and the fitting value of time series is measured by the Euclidean distances:

$$E_N = \sqrt{\sum_{i=1}^{N} (X_i - X_i')^2} \quad (9)$$

where X_i is the original sequence value, and X_i' is the fitting value.

(2) Define the average fitting error of each sequence point as E_N/N, the fitting error is E_i at the *i-th* subsequence and the length of *i-th* subsequence is $length(i)$, and $length(i) = W$, W is the number of data points contained in the *i-th* segment, and then compare and find out the subsequence segment fit value of the stock market trend feature point. Set the threshold as $P = (E_N/N) \times length(i)$:

1) If $E_i \geq P$, the *i-th* subsequence needs to perform the process of finding the trend feature point and the fitting error. In this process, it can be seen from (b) in Section IV-D-1 that the method of obtaining the initial feature points sequence segment is combined with the method of searching for the feature points again to obtain large deviation points, that is, the desired feature points are obtained;

2) If $E_i < P$, then the segment subsequence need not continue to be divided.

(3) Repeat the above steps (1) and (2) until the fitting error $E_k < P$ at the *k-th* subsequence or the subsequence length $length(k)$ is less than the set threshold d ($d \leq 7$).

(4) Through steps (1), (2) and (3), find all the trend feature points of the stock market time series, and linearly interpolate these feature points, and then use the improved linear representation method based on the trend feature point to perform these feature points in piecewise linear representation.

V. EXPERIMENTAL RESULTS ANALYSIS

A. EXPERIMENTAL DATA

All experiments selected the shares of Ping An Bank, CITIC Securities, Hualian Holdings, China Southern Airlines and Zhejiang Energy Power and the index of five different industries as the research objects. Stock sequence data are captured by JoinQuant trading platform and python language. The data is the daily closing price of five stocks from January 5, 2015 to December 31, 2015, with 244 data per stock. In order to highlight whether the experimental results are influenced by the sequence length, this paper also adds closing price of stock in one year to each stock on the basis of the original data. The specific time period is from January 2, 2014 to December 31, 2015, with 489 closing prices per stock. Five different industries are selected from the Eastmoney, and the five industries are the real estate and construction industry, machinery and equipment industry, energy industry, petrochemical industry and jewelry industry. The selected indices are the total sales of commodity housing, the price index of hardware and machinery, US crude oil index, chemical index and gold index. Each industry index selects 200 values. In the field of financial metrology, daily, weekly, monthly, quarterly or annual type of financial data belongs to low frequency data, and the time series data studied in this paper are all based on daily data. The experimental algorithm implementation uses matlabR2016a to obtain segmentation points and draw images. The experimental environment is i7-8700 CPU, memory 32GB, operating system is Windows 10. The first step after getting the data is to standardize the data and transform them into between 0 and 1, which is convenient for calculating the fitting error. The standard formula is defined as following:

$$norm(x_i) = \frac{x_i - min(X)}{max(X) - min(X)} + 0.000001 \quad (10)$$

where x_i is the original data, $\min(X)$ and $\max(X)$ are the minimum and maximum values in the original data, respectively. 0.000001 is to avoid the data equal to 0.

In order to better understand the characteristics of the experimental data in this paper, we make descriptive statistics on the research data, as shown in Tables 3, 4 and 5, in which the indicators include mean, standard deviation, maximum, minimum, kurtosis, and skewness. Kurtosis Coefficient (hereinafter abbreviated as *KC*) is an index to describe the sharpness of the peak of the symmetrical distribution curve. There are two forms of expression:

(1) If $KC > 0$, the data show a sharp peak distribution;

(2) If $KC < 0$, the data reflect a flat peak distribution.

Skewness Coefficient (hereinafter abbreviated as *SC*) is an index describing the symmetry of data based on the standard of normal distribution. *SC* is expressed as follows:

(1) If $SC = 0$, the data embody a symmetric form;

(2) If $SC > 0$, the data show a negative skewed distribution;

(3) If $SC < 0$, the data reflect a positive distribution;

(4) If $SC > 1$ or $SC < -1$, the data give the performance to a highly skewed distribution;

(5) If $SC \in [0.5, 1]$ or $SC \in [-0.5, -1]$, the distribution of the data is medium skewed distribution.

In this paper, we mainly analyze the distribution of all the data from the skewness and kurtosis coefficients. From Table 3, it can be concluded that the closing of Ping An Bank shows a medium skewness distribution with flat peak; Hualian Holdings and China Southern Airlines show a

TABLE 3. Descriptive statistics of stock sample data with sequence length 244.

Data sequences	Samples	Mean	Standard deviation	Maximum	Minimum	Skewness	Kurtosis
Ping An Bank	244	10.335	1.509	13.780	7.540	0.777	-0.600
CITIC Securities	244	23.054	6.477	34.670	12.340	-0.084	-1.370
Hualian Holdings	244	6.010	1.641	10.560	3.360	0.284	-0.311
China Southern Airline	244	8.577	2.425	15.350	4.300	0.272	-0.209
Zhejiang Energy Power	244	7.217	1.382	12.870	5.340	1.706	3.55

TABLE 4. Descriptive statistics of stock sample data with sequence length 489.

Data sequences	Samples	Mean	Standard deviation	Maximum	Minimum	Skewness	Kurtosis
Ping An Bank	489	10.725	1.442	15.840	7.540	0.917	0.894
CITIC Securities	489	18.336	7.422	34.670	10.050	0.693	-1.051
Hualian Holdings	489	4.644	1.858	10.560	2.630	0.858	-0.135
China Southern Airline	489	5.725	3.374	15.350	2.260	0.647	-0.760
Zhejiang Energy Power	489	6.539	1.286	12.870	4.470	1.679	5.048

TABLE 5. Descriptive statistics of industry sample data.

Data sequences	Samples	Mean	Standard deviation	Maximum	Minimum	Skewness	Kurtosis
Commodity housing sales	200	5.0E+07	4.62E+05	2.50E+08	1.51E+04	0.981	0.293
Hardware and electrical price index	200	98.039	1.101	100.620	96.430	0.314	-0.956
US crude oil index	200	56.465	4.592	66.300	42.530	-0.287	0.175
Chemical index	200	760.755	22.883	818	728	0.477	-0.850
Gold index	200	82.100	6.910	95.26	73.89	0.464	-1.262

flat peak negative skewness distribution; CITIC Securities presents a flat peak positive skewness distribution and the closing price of Zhejiang Energy Power presents a sharp peak-type highly skewness distribution. Table 4 is the data sample descriptive statistics added one year on the basis of Table 3, in which the closing price of Ping An Bank shows a sharp peak-type medium skewness distribution; CITIC Securities, Hualian Holdings and China Southern Airlines present a medium skewness distribution with flat peak; Zhejiang Energy Power embodies a sharp peak-type highly skewness distribution.

Since Table 5 is the sales index of some industries, the indicators are relatively large. From the analysis of skewness and kurtosis, Commodity housing sales shows a medium skewness distribution with sharp peak; Hardware and electrical price index, Chemical index and Gold index embody a flat peak negative skewness distribution; US crude oil index presents a positive skewness distribution with sharp peak.

B. ANALYSIS OF EXPERIMENTAL RESULTS

Considering the length of the paper, only the figure of one stock and one industry index is given here, and we give the specific evaluation indexes of all stocks and industries indexes in the following tables. The initial time series data before data modeling are shown in the black curve in all figures, and the sequences represent the daily closing price of the stock after standardization. From the initial data, we can see that the stock closing price changes irregularly and the trend fluctuates greatly. Therefore, in order to reduce the risk as much as possible, we will process the initial data to obtain important information.

In time series analysis, compression ratio and fitting error are generally adopted as evaluation indexes for the performance of piecewise linear representation algorithm of time series. Data compression refers to reducing the amount of data to reduce space and improve transmission efficiency without losing information. The compression ratio is an evaluation of the effectiveness of the data compression. The fitting error is to evaluate how close the fitted data is to the original data. If the compression ratio of time series is higher and the fitting error is smaller, then the performance of the piecewise linear representation method is better and the initial time series features can be depicted. Otherwise, the performance is lower and the initial time series features cannot be depicted.

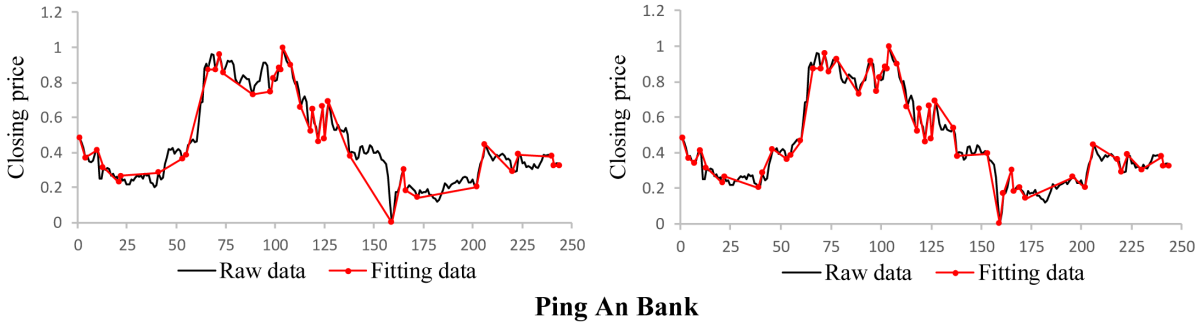

FIGURE 7. Segmentation points extraction for stock time series with data length 244. Ping An Bank: Initial fitting (left) and re-fitting (right). The closing price is standardized by the formula (10).

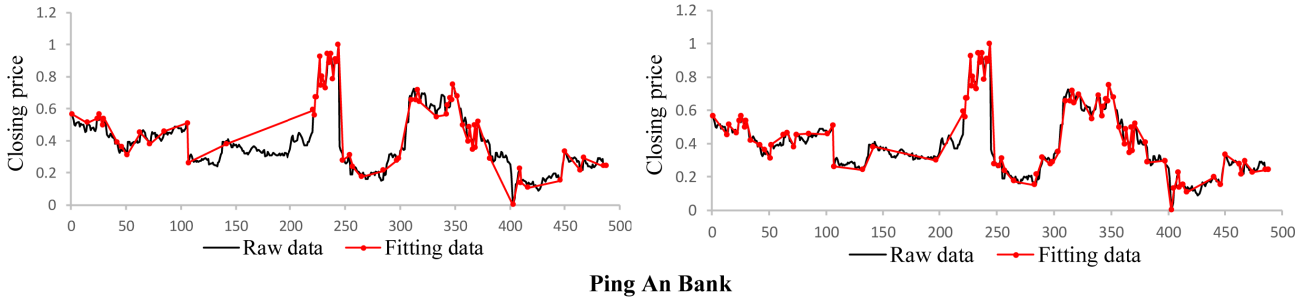

FIGURE 8. Segmentation points extraction for stock time series with two year data length. Ping An Bank: Initial fitting (left) and re-fitting (right). The closing price is standardized by the formula (10). The figure shows the fitting trends change after adding one year's closing price to the original data and the total sequence length is 489.

Giving the definition that the stock time series is:

$$X = (X_1, X_2, \cdots, X_N) \tag{11}$$

The endpoint of each subsequence is $(X_{t1}, X_{t2}, \cdots X_{tN})$, and the piecewise linear representation of the subsequence is

$$X_L = (f_1(X_1, X_{t1}), f_2(X_{t1+1}, X_{t2}), \cdots, f_n(X_{tN+1}, X_N)) \tag{12}$$

Thus, the compression ratio is defined as:

$$C = (1 - n/N) \times 100\% \tag{13}$$

where n is the number of segment points, and N is the number of points in the original time series.

In this paper, a piecewise linear representation algorithm based on trend feature points is utilized to extract segmentation points for stock time series with data length 244, as shown in Fig. 7, which are the sequence diagrams of the fitting of the stock data (closing price). To illustrate the reliability of this method, this paper not only adds one year of data as shown in Fig. 8 for the stock which is presented in Fig. 7, but also gives index data for five industries and different segmentation threshold of stock data. Besides, other methods in related work are added to further illustrate the performance of the method. Figure 9 presents the sequence diagrams of the fitting of the industry index, the change figures of segmentation threshold are shown in Fig. 10 and others methods are presented in Fig. 11. Due to the large variation of stock data, the way of changing the threshold (the segmentation length which can better reflect the change of sequence points) is selected in this paper to evaluate the performance of the method. Set thresholds d ($d \leq 5$) and d ($d \leq 9$) to compare with the original threshold d ($d \leq 7$). Considering that the threshold cannot be too large or too small, two other thresholds d ($d \leq 5$) and d ($d \leq 9$) are set based on the threshold d ($d \leq 7$) for experimental comparative analysis.

The effect diagrams of the initial fitting and re-fitting of the stock time series can be seen from Figs. 7 and 8, where the black curve is the initial stock data and the red segment line is the fitted stock data trend. The red dot on the left indicates the feature point for the first search, and the red dot on the right indicates the feature point determined by combining the feature point of the first search with the maximum absolute deviation point. During extracting the feature points of the initial fitting of the stock data, the principle of extreme points of the data needs to be followed and a certain trend change is reached between adjacent points. The feature points that are fitted again are the maximum absolute deviation points obtained according to the previous formula (3). As shown in Fig. 7, although the initial fitting curve of the stock data is close to the trend of the original data, it is found that many important points are ignored and a large deviation has been generated compared to the original data. Combining the

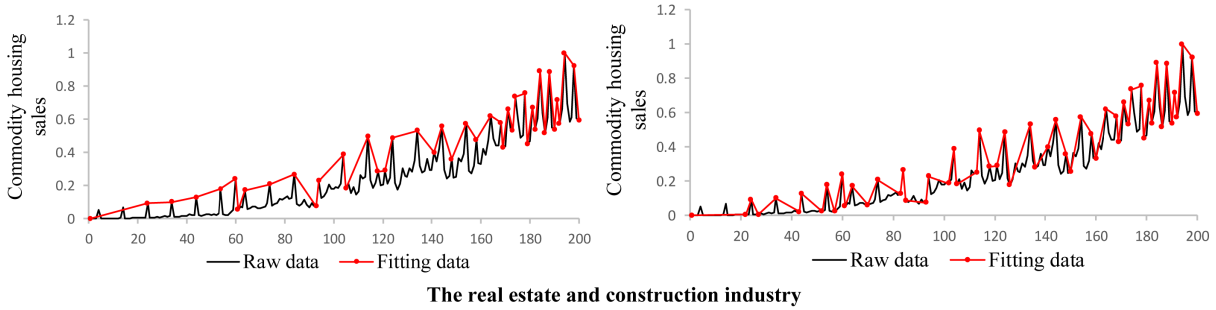

FIGURE 9. The sequence diagrams of the fitting of the industry index data. The real estate and construction industry: Initial fitting (left) and re-fitting (right).

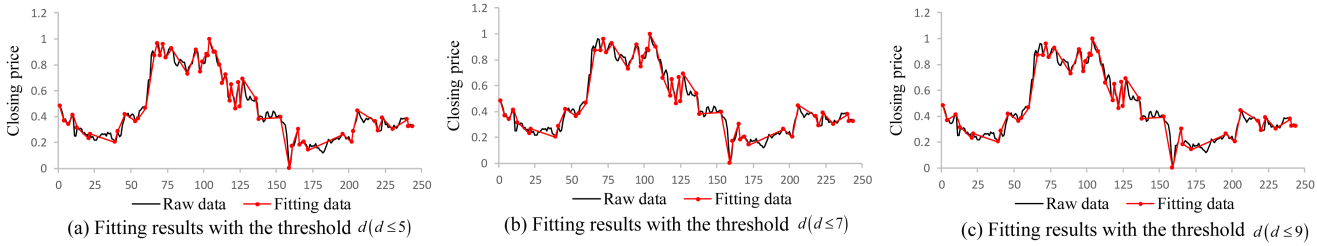

FIGURE 10. Ping An Bank: The black curve shows the trend of the original stock sequence and red is the fitting curve. The red curves in (a), (b) and (c) indicate the fitting trends when the thresholds are d ($d \le 5$), d ($d \le 7$) and d ($d \le 9$), respectively. From a visual point of view, the difference of trend change between (a) and (b) is less, while the difference of trend change between (b) and (c) is obvious. To illustrate their differences, a detailed numerical analysis is performed as shown in Table 7.

principle of finding points again to fit the stock price makes the result closer to the initial price data, and more consistent with the change trend of the initial data, and the fitting effect is better.

It can be seen from Fig. 8 that the fitting effect is comparable or better than that in Fig. 7, which indicates that the results are still reliable and the fitting effect is better even if the length of the data time series is different. As shown in Fig. 9, the effect maps of the initial fitting and re-fitting of the industry indices time series could be seen, where the black curve corresponds to the initial index, and the red segmentation line is for the fitted data trend. The specific evaluation criteria for the data fitting effect of Figs. 7-9 are derived from Table 6. Figure 10 presents the effect of fitting the different segmentation thresholds to the original data, in which the black curve is the initial data and the red segmentation line is the fitted stock data trend. The red dots in (a) are the fitting feature points with the threshold d ($d \le 5$), (b) is the fitting graph of the feature points with the threshold d ($d \le 7$), and (c) is the fitting graph of the feature points with the threshold d ($d \le 9$). Visually, it is found that the differences of the fitted curves given by the threshold d ($d \le 5$) and threshold d ($d \le 7$) are less, while the difference between the threshold d ($d \le 7$) and threshold d ($d \le 9$) is obvious. Therefore, their numerical values are analyzed in this paper, and the specific performance is shown in Table 7, from which we can see their differences. Figure 11 gives the effects of fitting other methods in related work and the proposed method in this paper, in which (a), (b) and (c) mean different segmentation methods in [30], [31] and [34] respectively, and (d) represents the proposed method of this paper. In addition to the visual changes, the specific evaluation indicators are obtained from Table 8, in order to find more significant differences between different methods.

Table 6 gives the analysis of the fitting of the stock data with different sequence length and the fitting analysis of five different industries indices. Table 7 gives a comparative analysis of fitting error and compression ratio represented by different thresholds, in which d ($d \le 7$) is the threshold set in this paper. Table 8 shows the comparative analysis of compression ratio and fitting error of different methods. We mainly use the evaluation index of compression ratio and fitting error, and the compression ratio (the calculation of compression ratio is shown in formula (13): $C = (1 - n/N) \times 100\%$, and n is the number of segment points, and N is the number of points in the original time series) is related to the number of feature points to be searched. It can be seen from Table 6 that the number of feature points found for the first time is smaller than the feature points found again. Although the compression ratio is relatively high, the fitting error of the initial search is also high, and the fitting effect is relatively poor, while the proposed method gives the much lower fitting error when the difference of compression ratio is less. As shown in Table 6, the difference between the initial fitting and re-fitting compression ratio of Hualian Holdings with data length 244 is only 4.508%, while the fitting error is

TABLE 6. Fit data compression rate and fitting error of five different stocks and five different industry indices.

Stock type	Data length	Evaluation index	First look	Improved lookup
Ping An Bank	244	Feature points	39	**52**
		Compression ratio	84.016%	**78.689%**
		Fitting error	1.175	**0.587**
	489	Feature points	66	**88**
		Compression ratio	86.475%	**81.967%**
		Fitting error	1.700	**0.891**
CITIC Securities	244	Feature points	51	**66**
		Compression ratio	79.098%	**72.951%**
		Fitting error	0.700	**0.483**
	489	Feature points	85	**103**
		Compression ratio	82.618%	**78.937%**
		Fitting error	0.913	**0.407**
Hualian Holdings	244	Feature points	44	**55**
		Compression ratio	81.967%	**77.459%**
		Fitting error	1.179	**0.387**
	489	Feature points	76	**91**
		Compression ratio	84.298%	**81.198%**
		Fitting error	1.148	**0.396**
China Southern Airline	244	Feature points	42	**53**
		Compression ratio	82.787%	**78.279%**
		Fitting error	0.982	**0.479**
	489	Feature points	59	**71**
		Compression ratio	87.935%	**85.481%**
		Fitting error	1.986	**0.476**
Zhejiang Power	244	Feature points	42	**52**
		Compression ratio	82.787%	**78.689%**
		Fitting error	1.289	**0.452**
	489	Feature points	77	**94**
		Compression ratio	84.254%	**80.777%**
		Fitting error	1.101	**1.009**
Industry index	**Data length**	**Evaluation index**	**First look**	**Improved lookup**
Commodity housing sales	244	Feature points	43	**57**
		Compression ratio	78.5%	**71.5%**
		Fitting error	1.815	**1.212**
Hardware and electrical price index	244	Feature points	20	**24**
		Compression ratio	90%	**88%**
		Fitting error	1.4	**0.487**
US crude oil index	244	Feature points	25	**35**
		Compression ratio	87.5%	**82.5%**
		Fitting error	2.044	**0.807**
Chemical index	244	Feature points	11	**18**
		Compression ratio	94.5%	**91%**
		Fitting error	1.119	**0.66**
Gold index	244	Feature points	18	**27**
		Compression ratio	91%	**86.5%**
		Fitting error	1.012	**0.518**

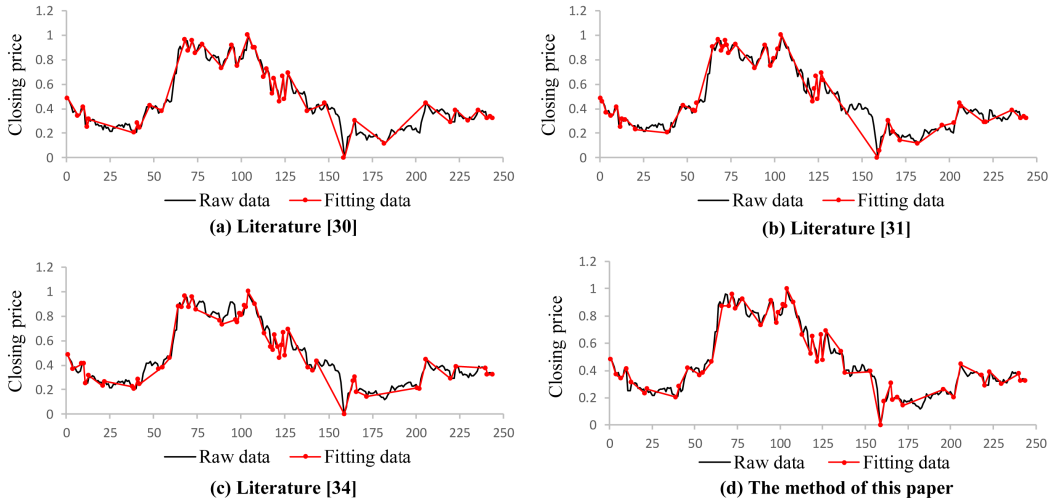

FIGURE 11. Ping An Bank: (a), (b) and (c) are for the fitting methods in [30], [31] and [34], respectively, (d) is for the method fitting of this paper.

TABLE 7. Fit data compression rate and fitting error of five stocks with different thresholds.

Stock type	Evaluation index	$d(d \le 5)$	$d(d \le 7)$	$d(d \le 9)$
Ping An Bank	Feature points	56	52	49
	Compression ratio	77.049%	78.689%	79.912%
	Fitting error	0.548	0.587	0.572
CITIC Securities	Feature points	74	66	59
	Compression ratio	69.672%	72.951%	75.820%
	Fitting error	0.375	0.483	0.564
Hualian Holdings	Feature points	66	55	52
	Compression ratio	72.951%	77.459%	78.689%
	Fitting error	0.326	0.387	0.513
China Southern Airline	Feature points	55	53	49
	Compression ratio	77.459%	78.279%	79.918%
	Fitting error	0.479	0.479	0.514
Zhejiang Power	Feature points	58	52	47
	Compression ratio	76.230%	78.689%	80.738%
	Fitting error	0.410	0.452	0.554

0.792. From Table 7, we can see that when the threshold d is less than or equal to 5, although the fitting error is relatively low, the compression ratio is also low; when the threshold d is less than or equal to 9, the compression ratio is high, while the fitting error is also high. When fitting the data, the fitting effect is better only when the compression ratio is high and the fitting error is low. This means that the threshold should be set moderately according to the data, too high or too low to meet the requirements. In this paper, the threshold is more close to the data changes and more suitable for data fitting. As shown in Table 7, the compression ratio and fitting error of Hualian Holdings are 72.951% and 0.326 when threshold d is less than or equal to 5; the compression ratio and fitting error of Hualian Holdings are 78.689% and 0.513, respectively, when threshold d is less than or equal to 9. In this paper, the compression ratio is 77.459% and the fitting error is 0.387.

TABLE 8. Fit data compression rate and fitting error of different methods.

Stock type	Evaluation index	This paper	Literature [30]	Literature [31]	Literature [34]
Ping An Bank	Feature points	**52**	42	50	57
	Compression ratio	**78.689%**	82.787%	79.508%	76.639%
	Fitting error	**0.587**	0.779	1.012	0.945
CITIC Securities	Feature points	**66**	61	39	66
	Compression ratio	**72.951%**	75%	84.016%	72.951%
	Fitting error	**0.483**	0.414	0.970	0.552
Hualian Holdings	Feature points	**55**	43	49	56
	Compression ratio	**77.459%**	82.377%	79.918%	77.049%
	Fitting error	**0.387**	0.825	0.542	1.133
China Southern Airline	Feature points	**53**	38	50	55
	Compression ratio	**78.279%**	84.426%	79.508%	77.459%
	Fitting error	**0.479**	0.635	0.754	0.969
Zhejiang Power	Feature points	**52**	40	50	56
	Compression ratio	**78.689%**	83.607%	79.508%	77.049%
	Fitting error	**0.452**	1.183	0.715	1.190
Mean	Feature points	**57**	45	48	58
	Compression ratio	**77.213%**	81.639%	80.492%	76.223%
	Fitting error	**0.478**	0.767	0.599	0.958

In conclusion, the proposed method provides a lower fitting error under keeping a higher compression ratio.

Table 8 provides the results of different methods. It is found that the mean fitting error generally is lower in the proposed method than that in the methods from [30], [31] and [34]. As shown in Table 8, the mean compression ratio and the mean fitting error are 81.639% and 0.767 in [30], 80.492% and 0.599 in [31] and 76.223% and 0.958 in [34], respectively. Moreover, it can also be found that the proposed method of this paper got the mean compression ratio of 77.213% and the mean fitting error of 0.478. Compared with the method in [30], the proposed method got a lower fitting error under keeping compression ratio with less difference; compared with the method in [31], although the compression ratio in this paper is 1% to 2% lower, the fitting error is basically half of it, for example, Ping An Bank; compared with the method in [34], two indicators are better from the perspective of single stock or overall mean, and even about a third of it, for example, Hualian Holdings and Zhejiang Energy Power. In a word, the proposed method in this paper has more advantages than those in [30], [31] and [34].

To sum up, by analyzing the experimental results above, the piecewise linear representation of the time series using the trend feature points initially found in the time series has a higher compression ratio and larger fitting error. On this basis, based on the judgment of subsequence fitting error and global fitting error mean, the operation of finding feature points is performed in the subsequence segment that meets the threshold. From the Table 6, we can see the change of the compression ratio and the fitting error of the data in the different industries and the different stocks with different sequence length. In the initial fitting, the fitting error and the compression ratio are high, but the fitting error is significantly reduced on the re-fitting and the compression ratio retains relatively high. From the compression ratio and fitting error shown by different thresholds in Table 7, it can be seen that the threshold in this paper has low fitting error when the compression ratio is guaranteed to be high, which meets the requirements of data compression. The analysis and comparison of different methods in Table 8 give same conclusions that the proposed method obtains low fitting error and maintains high compression ratio. According to the experimental results, we can conclude that the method in this paper retains the characteristics and integrity of stock time series, and gets a good fitting effect in the feature representation of the stock time series, which provide a choice for data fitting. At the same time, the stock data fitting of different time series lengths shows that the method in this paper is not affected by the

sequence length. Last but not least, the research in this paper also brings benefits to managers and traders. For managers, the research on changes in stock market trend can prompt them to improve the relevant regulatory system and make the regulation more targeted, so as to further promote the healthy and stable development of the stock market. A stable stock market plays a very important role for the national economy, and drives the economic lifeline ahead of development and benefit many people. For traders, the research of stock market turbulence provides a certain theoretical basis and guiding suggestions when they make decisions, which can reduce investment risks for them to a certain extent.

VI. CONCLUDING REMARKS

The feature representation of the security time series is to convert the original data from higher dimensions to lower dimensions on the premise of preserving the initial time series features as much as possible, which is beneficial to improve the efficiency of data mining, similarity measurement and other research work. Piecewise linear representation of stock time series is an effective method to reduce the difficulty of stock time series data processing. According to the prior knowledge, firstly this paper selects the feature points that meet the conditions, and then using the piecewise linear representation method to process the stock time series. Finally, analyzing the experimental results according to the evaluation indicators.

The experimental results show that the proposed algorithm achieves the purpose of effectively compressing time series data while reflecting the trend of time series, and work well on time series with obvious periodicity and drastic mode fluctuations. At the same time, the research is also adapted to online segmentation. However, there are still some weak points that are open for debate. Firstly, the threshold of this paper is manually set, which is time-consuming to manually set for experiments with a large amount of data. Secondly, historical data is used to analyze the stock trends, which the indicator is single. In order to better predict the movements of the stock market, in future research work, we will search for some ways that enable intelligent selection of thresholds based on data, and also start from the behavioral finance in social media to predict the stock trend by interpreting the emotions in the text information and combining stock historical data. The textual information related to social media and online news have been proven to be effective in predicting the future trend of the stock market. By increasing textual feature information in the media, we can make the results more reliable and greater benefits market regulators and traders.

REFERENCES

[1] M. E. Bildirici and M. M. Badur, "The effects of oil prices on confidence and stock return in China, india and russia," *Quant. Finance Econ.*, vol. 2, no. 4, pp. 884–903, 2018.
[2] R. J. Cebula and D. Capener, "The impact of federal income tax rate cuts on the municipal bond market in the U.S.: A brief exploratory empirical note," *Quant. Finance Econ.*, vol. 2, no. 2, pp. 407–412, 2018.
[3] Y. Liu, Y. Zheng, and B. M. Drakeford, "Reconstruction and dynamic dependence analysis of global economic policy uncertainty," *Quant. Finance Econ.*, vol. 3, no. 3, pp. 550–561, Sep. 2019.
[4] J. Zhang, Y.-H. Shao, L.-W. Huang, J.-Y. Teng, Y.-T. Zhao, Z.-K. Yang, and X.-Y. Li, "Can the exchange rate be used to predict the shanghai composite index?" *IEEE Access*, vol. 8, pp. 2188–2199, 2020.
[5] S. Zhou and J. Zhang, "Empirical test of size effect in China stock market," in *Proc. IEEE Int. Conf. Bus. Intell. Financial Eng.*, Beijing, China, Jul. 2009, pp. 691–694.
[6] Y. Xia, "The effects of economic and political events on the behaviour of stock market index in China," in *Proc. 2nd Int. Conf. Artif. Intell., Manage. Sci. Electron. Commerce (AIMSEC)*, Aug. 2011, pp. 2524–2527.
[7] J. Engel, M. Wahl, and R. Zagst, "Forecasting turbulence in the asian and European stock market using regime-switching models," *Quant. Finance Econ.*, vol. 2, no. 2, pp. 388–406, 2018.
[8] X. Pang, Y. Zhou, P. Wang, W. Lin, and V. Chang, "An innovative neural network approach for stock market prediction," *J. Supercomput.*, vol. 76, no. 3, pp. 2098–2118, Mar. 2020.
[9] Y. Zhang, P. Shang, and H. Xiong, "Multivariate generalized information entropy of financial time series," *Phys. A, Stat. Mech. Appl.*, vol. 525, pp. 1212–1223, Jul. 2019.
[10] J. R. Thompson and J. R. Wilson, "Multifractal detrended fluctuation analysis: Practical applications to financial time series," *Math. Comput. Simul.*, vol. 126, pp. 63–88, Aug. 2016.
[11] E. Şafak, "Time-domain representation of frequency-dependent foundation impedance functions," *Soil Dyn. Earthq. Eng.*, vol. 26, no. 1, pp. 65–70, Jan. 2006.
[12] S. J. Wilson, "Data representation for time series data mining: Time domain approaches," *Wiley Interdiscipl. Rev., Comput. Statist.*, vol. 9, no. 1, Jan. 2017, Art. no. e1392.
[13] G. Taskin, H. Kaya, and L. Bruzzone, "Feature selection based on high dimensional model representation for hyperspectral images," *IEEE Trans. Image Process.*, vol. 26, no. 6, pp. 2918–2928, Jun. 2017.
[14] M. G. Baydogan and G. Runger, "Time series representation and similarity based on local autopatterns," *Data Mining Knowl. Discovery*, vol. 30, no. 2, pp. 476–509, Mar. 2016.
[15] Z. Wu, W. Lin, P. Liu, J. Chen, and L. Mao, "Predicting long-term scientific impact based on multi-field feature extraction," *IEEE Access*, vol. 7, pp. 51759–51770, 2019.
[16] P. Montero-Manso, G. Athanasopoulos, R. J. Hyndman, and T. S. Talagala, "FFORMA: Feature-based forecast model averaging," *Int. J. Forecasting*, vol. 36, no. 1, pp. 86–92, Jan. 2020.
[17] S. Gaur and M. C. Deo, "Real-time wave forecasting using genetic programming," *Ocean Eng.*, vol. 35, nos. 11–12, pp. 1166–1172, Aug. 2008.
[18] L. Sun and G. Liu, "Visual object tracking based on combination of local description and global representation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 4, pp. 408–420, Apr. 2011.
[19] S.-H. Park, J.-H. Lee, S.-J. Chun, and J.-W. Song, "Representation and clustering of time series by means of segmentation based on PIPs detection," in *Proc. 2nd Int. Conf. Comput. Autom. Eng. (ICCAE)*, Feb. 2010, pp. 17–21.
[20] D. Faranda, F. M. E. Pons, E. Giachino, S. Vaienti, and B. Dubrulle, "Early warnings indicators of financial crises via auto regressive moving average models," *Commun. Nonlinear Sci. Numer. Simul.*, vol. 29, nos. 1–3, pp. 233–239, Dec. 2015.
[21] H. Shatkay and S. B. Zdonik, "Approximate queries and representations for large data sequences," in *Proc. 12th Int. Conf. Data Eng.*, Feb. 1996, pp. 536–545.
[22] S. A. P. Kani, G. H. Riahy, and D. Mazhari, "An innovative hybrid algorithm for very short-term wind speed prediction using linear prediction and Markov chain approach," *Int. J. Green Energy*, vol. 8, no. 2, pp. 147–162, Mar. 2011.
[23] S. Rebello, H. Yu, and L. Ma, "An integrated approach for system functional reliability assessment using dynamic Bayesian network and hidden Markov model," *Rel. Eng. Syst. Saf.*, vol. 180, pp. 124–135, Dec. 2018.
[24] L. Ponta and A. Carbone, "Information measure for financial time series: Quantifying short-term market heterogeneity," *Phys. A, Stat. Mech. Appl.*, vol. 510, pp. 132–144, Nov. 2018.
[25] Y. Teng and P. Shang, "Transfer entropy coefficient: Quantifying level of information flow between financial time series," *Phys. A, Stat. Mech. Appl.*, vol. 469, pp. 60–70, Mar. 2017.
[26] L. Ponta, M. Trinh, M. Raberto, E. Scalas, and S. Cincotti, "Modeling non-stationarities in high-frequency financial time series," *Phys. A, Stat. Mech. Appl.*, vol. 521, pp. 173–196, May 2019.
[27] L. Ponta and S. Cincotti, "Traders' networks of interactions and structural properties of financial markets: An agent-based approach," *Complexity*, vol. 2018, pp. 1–9, Jan. 2018.

[28] H. Xiao, X.-F. Feng, and Y.-F. Hu, "A new segmented time warping distance for data mining in time series database," in *Proc. Int. Conf. Mach. Learn. Cybern.*, Aug. 2004, pp. 1277–1281.

[29] G. Dorgo and J. Abonyi, "Learning and predicting operation strategies by sequence mining and deep learning," *Comput. Chem. Eng.*, vol. 128, pp. 174–187, Sep. 2019.

[30] C. Faloutsos, M. Ranganathan, and Y. Manolopoulos, "Fast subsequence matching in time-series databases," *ACM SIGMOD Rec.*, vol. 23, no. 2, pp. 419–429, Jun. 1994.

[31] L. Fang, H. Zhao, P. Wang, M. Yu, J. Yan, W. Cheng, and P. Chen, "Feature selection method based on mutual information and class separability for dimension reduction in multidimensional time series for clinical data," *Biomed. Signal Process. Control*, vol. 21, pp. 82–89, Aug. 2015.

[32] X. Zhong and D. Enke, "Forecasting daily stock market return using dimensionality reduction," *Expert Syst. Appl.*, vol. 67, pp. 126–139, Jan. 2017.

[33] C. Yan, J. Fang, L. Wu, and S. Ma, "An approach of time series piecewise linear representation based on local maximum minimum and extremum," *J. Inf. Comput. Sci.*, vol. 10, no. 9, pp. 2747–2756, Jun. 2013.

[34] J. Yin, Y.-W. Si, and Z. Gong, "Financial time series segmentation based on turning points," in *Proc. Int. Conf. Syst. Sci. Eng.*, Jun. 2011, pp. 394–399.

[35] Y.-W. Si and J. Yin, "OBST-based segmentation approach to financial time series," *Eng. Appl. Artif. Intell.*, vol. 26, no. 10, pp. 2581–2596, Nov. 2013.

[36] G. Luo, K. Yi, S.-W. Cheng, Z. Li, W. Fan, C. He, and Y. Mu, "Piecewise linear approximation of streaming time series data with max-error guarantees," in *Proc. IEEE 31st Int. Conf. Data Eng.*, Apr. 2015, pp. 173–184.

[37] Y. Lin and Z. Wang, "Time series piecewise linear representation method based on first-order filtering," *Comput. Eng.*, vol. 42, no. 9, pp. 151–157, Sep. 2016.

[38] E. Keogh, "A Fast and robust method for pattern matching in time series databases," *Proc. WUSS*, vol. 97, no. 1, p. 99, 1997.

[39] J. Yi, Y. Xue, L. Chen, A. Compare, and X. Mao, "Facial expression recognition considering individual differences in facial structure and texture," *IET Comput. Vis.*, vol. 8, no. 5, pp. 429–440, Oct. 2014.

[40] J. Yi, A. Chen, Z. Cai, Y. Sima, M. Zhou, and X. Wu, "Facial expression recognition of intercepted video sequences based on feature point movement trend and feature block texture variation," *Appl. Soft Comput.*, vol. 82, Sep. 2019, Art. no. 105540.

[41] B. Wang, W. Chen, and Y. Zhu, "An empirical study of the relationship between the stock discussion board's posting numbers and stock trading volume," in *Proc. 10th Int. Conf. Service Syst. Service Manage.*, Jul. 2013, pp. 884–888.

[42] L. Li, X. Su, Y. Zhang, Y. Lin, and Z. Li, "Trend modeling for traffic time series analysis: An integrated study," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 6, pp. 3430–3439, Dec. 2015.

[43] S. Pravilovic, M. Bilancia, A. Appice, and D. Malerba, "Using multiple time series analysis for geosensor data forecasting," *Inf. Sci.*, vol. 380, pp. 31–52, Feb. 2017.

[44] B. Zheng, S. W. Myint, P. S. Thenkabail, and R. M. Aggarwal, "A support vector machine to identify irrigated crop types using time-series landsat NDVI data," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 34, pp. 103–112, Feb. 2015.

[45] S. M. Idrees, M. A. Alam, and P. Agarwal, "A prediction approach for stock market volatility based on time series data," *IEEE Access*, vol. 7, pp. 17287–17298, 2019.

[46] M. Wen, P. Li, L. Zhang, and Y. Chen, "Stock market trend prediction using high-order information of time series," *IEEE Access*, vol. 7, pp. 28299–28308, 2019.

[47] O. Ait Maatallah, A. Achuthan, K. Janoyan, and P. Marzocca, "Recursive wind speed forecasting based on hammerstein auto-regressive model," *Appl. Energy*, vol. 145, pp. 191–197, May 2015.

[48] M. Lydia, S. Suresh Kumar, A. Immanuel Selvakumar, and G. Edwin Prem Kumar, "Linear and non-linear autoregressive models for short-term wind speed forecasting," *Energy Convers. Manage.*, vol. 112, pp. 115–124, Mar. 2016.

[49] L.-Y. Wei, C.-H. Cheng, and H.-H. Wu, "A hybrid ANFIS based on n-period moving average model to forecast TAIEX stock," *Appl. Soft Comput.*, vol. 19, pp. 86–92, Jun. 2014.

[50] S. Hansun and M. B. Kristanda, "Performance analysis of conventional moving average methods in forex forecasting," in *Proc. Int. Conf. Smart Cities, Autom. Intell. Comput. Syst. (ICON-SONICS)*, Nov. 2017, pp. 11–17.

[51] A. Wiesel, O. Bibi, and A. Globerson, "Time varying autoregressive moving average models for covariance estimation," *IEEE Trans. Signal Process.*, vol. 61, no. 11, pp. 2791–2801, Jun. 2013.

MENGNA ZHOU received the B.S. degree in information management and information system from Hengyang Normal University, Hengyang, China, in 2018. She is currently pursuing the M.S. degree in College of Computer and Information Engineering, Central South University of Forestry and Technology (CSUFT), Changsha, China. Her research interests include deep learning for financial public opinion analysis and financial time series analysis. She is also working on stock analysis system based on financial time series and public opinion analysis.

JIZHENG YI (Member, IEEE) received the B.S. degree in electronic information science and technology from Shandong Agricultural University, Tai'an, China, in 2008, the M.S. degree in pattern recognition and intelligent system from Wuyi University, Jiangmen, China, in 2011, and the Ph.D. degree from the School of Electronic and Information Engineering, Beihang University, Beijing, China, in 2015. He worked and studied at the Medical Image Processing Group (MIPG), School of Medicine, University of Pennsylvania (UPenn), Philadelphia, PA, USA, from 2018 to 2019. He is currently an Associate Professor with the College of Computer and Information Engineering, Central South University of Forestry and Technology (CSUFT), Changsha, China. He is also in charge of the forestry virtual reality technology of the Hunan Forestry Information Research Center and the Intelligent Video Analysis of the Institute for Intelligent 3D Information, CSUFT. He is also working on stock analysis system based on financial time series and public opinion analysis. He is also chairing some research projects, including the National Natural Science Foundation of China, the Hunan Provincial Natural Science Foundation of China, and the Scientific Research Fund of Hunan Provincial Education Department. He is the author of one book and more than 20 articles. His research interests include machine learning for financial public opinion analysis, medical and visual image processing, and affective computing.

JIEQIONG YANG received the B.S. degree in mathematics from Hengyang Normal University, Hengyang, China, in 2008, and the M.S. degree in statistics from Hunan University, Changsha, China, in 2010. She is currently pursuing the Ph.D. degree with the College of Life Science and Technology, Central South University of Forestry and Technology (CSUFT), Changsha. From 2010 to 2019, she worked in a securities company, Changsha, mainly involved in securities investment and analysis. She is also working on stock analysis system based on financial time series and public opinion analysis. Her research interests include financial public opinion analysis and ecological big data analysis.

YI SIMA received the B.S. degree in communication engineering from the Hunan Institute of Science and Technology, Yueyang, China, in 2018. He is currently pursuing the M.S. degree with the Hunan Forestry Information Research Center, College of Computer and Information Engineering, Central South University of Forestry and Technology (CSUFT), Changsha, China. He is also working on face recognition and affective computing. His research interests include human–computer interaction, image processing, and pattern recognition.