Expert Systems With Applications 182 (2021) 115217

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

A similarity measurement for time series and its application to the stock market

Feng Zhao [a,*], Yating Gao [b], Xinning Li [b], Zhiyong An [a], Shiyu Ge [a], Caiming Zhang [a,c]

[a] School of Computer Science and Technology, Shandong Technology and Business University, 191 Binhai Middle Road, Laishan District, Yantai City, Shandong Province, China
[b] School of Statistics, Shandong Technology and Business University, 191 Binhai Middle Road, Laishan District, Yantai City, Shandong Province, China
[c] School of Software, Shandong University, 27 Shanda Nanlu, Jinan City, Shandong Province, China

ARTICLE INFO

Keywords:
Similarity measurements
Personalization
Multi-perspective
Stock prediction

ABSTRACT

The stock market is a very important financial market, and the prediction of the stock has always been of great interest to many investors. Nowadays, many methods for predicting stocks have been developed and one of the most commonly adopted strategies is to seek similar stocks through historical data to make predictions. The key to this strategy is the construction of a reasonable similarity measurement. In this paper, for accurately describing the similarity between a pair of time series, a novel similarity measurement is proposed, which is named as the dynamic multi-perspective personalized similarity measurement (DMPSM). Specifically, the segmented stock series are weighted according to the principle that the closer to current data, the more weight will be given. Then, Canberra distance is embedded into the dynamic time warping (DTW) to measure the similarity between any pair of time series. By this way, the DMPSM can not only reflect the personalization of stock time series, but also eliminate the impact of singularities and apply to one-to-many matching. To validate the efficiency of DMPSM, experiments utilized 285 stocks from the Shanghai Stock Exchange and the results demonstrated the superiority of the proposed approach over similarity measurements, including Euclidean distance, Canberra distance and DTW.

1. Introduction

Stock is an important way for people to invest and doing financial and asset management. Precise prediction for stocks can not only bring benefits to investors, but also greatly promote the development of the national economy. Stock prediction, with the aim of predicting the future price trend of stocks, plays a key role in seeking maximized profit from the stock investment and has attracted increasing attention from brilliant minds. However, the stock price is affected by many elements, such as interest rates, exchange rates, stock price indexes in other countries, domestic and global economic situations, etc., its prediction is still a challenging task (Zhang et al., 2017).

To date, researchers have developed many methods for predicting stocks accurately from different perspectives (Kuremoto et al., 2014; Preethi & Santhi, 2012; Groth & Muntermann, 2011). Especially, one of the most common adopted strategies is based on historical data, which is inspired by the view that history repeats itself. Its main idea is to make a reference between the historical data and the current data and then get a final prediction from their similarity. The prediction based on historical data consists of two important parts. One is using appropriate similarity measurements to select the most similar series from historical data. The other is choosing proper prediction methods to utilize these similar series for forecasting.

For the aspect of the similarity measurements, there are various existing similarity measurements for time series. The similarity between cryptocurrency feature vectors was computed for predicting the up/down movements of cryptocurrencies (Bai et al., 2019). Although this algorithm has been improved to some degree, it still only adopted Euclidean distance for similarity measurement, and this algorithm was only appropriate for time series at the same time and could not weaken the influence of singularities. Sun et al. (2021) developed a similarity measurement that was based on extreme point bias compensation. However, it had much higher computational complexity and only taken Euclidean distance into consideration. Juszczuk et al. (2020) constructed a similarity measurement based on relative instrument changes to test the predictive capabilities of stock companies, currency pairs, and

* Corresponding author.
E-mail address: zhaofeng1016@126.com (F. Zhao).

https://doi.org/10.1016/j.eswa.2021.115217
Received 29 November 2020; Received in revised form 24 March 2021; Accepted 13 May 2021
Available online 18 May 2021

stock indexes. One of the major drawbacks of this approach is that this method is applied to three different data groups, but it only showed advantages over the DTW measure in the case of currency pairs. For applying to measure time series at different times, a Mahalanobis distance-based Dynamic Time Warping (MDDTW) measure for classification was proposed. But it was not always calculated successfully due to the instability of the covariance matrix when computing Mahalanobis distance, which means that the MDDTW had high standards with data (Mei et al., 2016). Tsinaslanidis (2018) constructed an algorithmic approach by using mainly the DTW algorithm and two of its modifications: subsequence DTW and derivative DTW. Although it was able to describe the similarity between the sequences at different times, it could not demonstrate the importance of time. Overall, stock prices exhibit dynamic, non-linear, non-parametric and chaotic properties in nature (Han et al., 2020). However most of the existing methods only measure the stock series from a single visual angle, but cannot take the multiple characteristics of the stock series into account at the same time. Therefore, how to construct a reasonable similarity measurement to suitably reflect the characteristics of stock time series, is still an issue for the historical-data-based precision stocks prediction.

To address the issue, we construct a novel similarity measurement, which is called the dynamic multi-perspective personalized similarity measurement (DMPSM), for accurately describing the similarity between a pair of segmented time series which refer to the sequences of closing stock price in this paper. Specifically, after segmenting the whole stock sequences into small pieces of time series, we firstly weigh the segmented series according to the principle of that the closer the factor to the current time, the greater the weight. Secondly, we embedded Canberra distance into the DTW to measure the similarity between any pair of time series. In other words, we adopt Canberra distance to construct the DTW matrix. Finally, we will select the most similar time series for prediction.

The DMPSM has three major characteristics. (1) Personalization. The DMPSM can reflect the personalization of stock time series, that is, the influence of time series on the current data decreases with the advance of time. Therefore, we weigh the time series where the section closer to the current time has a greater weight while the section farther away has a smaller weight (Li et al., 2009). (2) Elimination of singularity influence. Canberra distance is not sensitive to singularities which refer to the points in a sequence that are too large or too small. Therefore, by employing Canberra distance as the main measurement, the DMPSM can eliminate the impact of singularity. (3) Application to the situation of time shifts and warpings. DTW can minimize the distance between two time series by constructing an optimal warping path (Han et al., 2020). Therefore, by adopting DTW as the measuring framework, the DMPSM possesses the dynamic property that can flexibly measure the similarity between any pair of time series and thus can cope with time shifts and warpings in stock time series. For verifying the efficiency of DMPSM, two experiments are designed by using 285 stocks from the Shanghai Stock Exchange. The experimental results show that the performance of DMPSM is obviously better than others, such as Euclidean distance, Canberra distance and DTW, meanwhile, it is more suitable for flat data.

For the aspect of prediction methods, a lot of prediction methods based on similar measurement have been developed in recent years. Fenghua et al. (2014) proposed an SSA-SVM combination prediction, which had better predictive effect in stock prediction and provided investors with a certain value for stock forecasting. Khoojine and Han (2020) developed a Stock Price Network Autoregressive Model for stock prediction which had high accuracy and performance. Zhang and Lou (2020) applied neural network and back-propagation (BP) algorithm onto the classification and prediction of stock price patterns and had better prediction effect.

In fact, even if using the same forecasting method, different forecasting schemes will have different forecasting results. To obtain a better prediction performance, we provide four alternative prediction schemes. Specifically, Scheme 1 is only utilizing the sample series to predict while the similar series are not used; Scheme 2 is selecting the most similar series and averaging their last value to represent the prediction value; Scheme 3 is choosing the most similar series to perdict; Scheme 4 is picking out the most similar series and averaging theses series to obtain a new time series for prediction.

In brief, we mainly make the work of two following aspects in this paper. On the one hand, we propose a new similarity measurement that overcomes some difficulties for time series. It not only can reflect the time characteristics, but also can eliminate the negative impact of singularities and cope with time shifts and warpings. Meanwhile, considering that different predictions will get different results, we also provide some alternative prediction schemes.

The rest of the paper is organized as follows. Section 2 briefly introduces the related similarity measurements and prediction methods. In Section 3, the DMPSM method will be constructed to measure the similarity between time series and the prediction schemes will be described in detail. The experimental settings and results analysis are reported in Section 4. Finally, we conclude this paper and discuss some possible future directions in Section 5.

2. Related work

The prediction based on historical data can be mainly divided into two parts. The first part is utilizing reasonable measurements to find the similar series from historical data (see in Fig. 1(a)). Then, the second part is adopting appropriate prediction methods to obtain the prediction value by using the similar series (see in Fig. 1(b)). Fig. 1 shows the flowchart of these two parts, which will be briefly introduced in the following. Specifically, three similarity measurements, Euclidean distance, Canberra distance and DTW, are introduced in Section 2.1. Then, two common prediction methods are introduced in Section 2.2, containing autoregressive model and neural network.

2.1. Similarity measurements

Before continuing, a note on mathematical notations is given as follows. We use upper case letter (e.g., X, Y, N) to denote time series, lower case letters with the subscript (e.g., p_k, ω_i) to denote an element from a matrix or a series, upper case bold letters (e.g., $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{C}$) to denote matrixes, the letter d with the superscript to denote distances (e.g.,d^E denotes Euclidean distance).

(1) Euclidean distance

Euclidean distance (Anton, 1993) refers to the true distance between two points in m-dimensional space, or the natural length of the vector, i.e., the distance from the point to the origin. And its formula between $X = (x_1, x_2, \cdots, x_n)$ and $Y = (y_1, y_2, \cdots, y_m)$ can be described as follows:

$$d^E(X, Y) = \sqrt{\frac{\sum_{i=1}^{n} (x_i - y_i)^2}{n}}, n = m \tag{1}$$

Euclidean distance has been widely applied to time series since the advantages of fast computing speed and low complexity. Qiang and Vasileios (2007) utilized Euclidean distance to improve a dimensionality reduction technique for time series analysis which significantly improved the efficiency and accuracy of similarity searches. Chen et al. (2015) used Euclidean distance to construct an algorithm for time series similarity matching and this algorithm had a significant superiority in efficiency and accuracy.

However, the biggest drawback of Euclidean distance is that it very sensitive to singularities, making a negative impact on similarity measuring of time series. For better comprehension and comparison, an example of Euclidean distance is shown in Fig. 2(a), where Series A, B and C represent different stock time series. In Fig. 2(a), Series A and B are flat, but B has a singularity, and C has an obvious fluctuation, but $d^E(A, B) = d^E(A, C) = 0.85$, indicating that Euclidean distance is

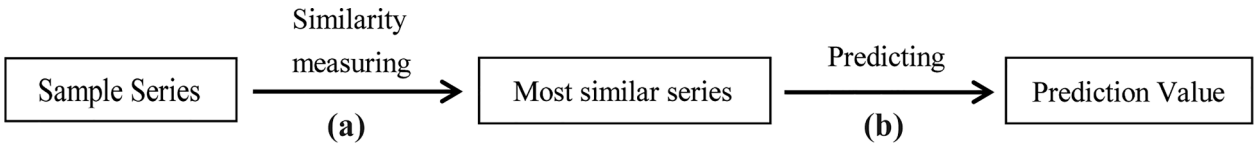

Fig. 1. The flowchart of prediction based on similarity measurement.

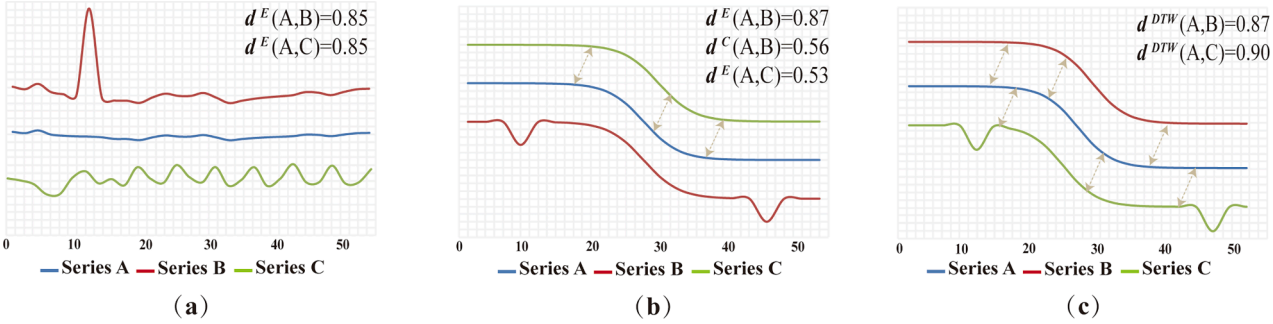

Fig. 2. The related examples of the three measurements.

sensitive to singularities.

(2) Canberra distance

In order to overcome the shortcoming of Euclidean distance, many methods have been proposed. And one of the most effective approaches is Canberra distance (Bijnen, 1973), whose formula is as follows.

$$d^C(X,Y)=\frac{1}{n}\sum_{i=1}^{n}\frac{|x_i-y_i|}{|x_i|+|y_i|}, n=m \tag{2}$$

From Eq. (2), we can see that Canberra distance is a dimensionless quantity and is insensitive to singularities. By taking this advantage of Canberra distance, Faisal et al. (2020) utilized Canberra distance into inter-centroid K-means and made a comparative analysis of Euclidean Distance and Manhattan Distance, find that the Canberra method is superior to Euclidean and Manhattan on their dataset.

However, the biggest disadvantage of Canberra distance is that it cannot cope with time shifts and warpings. For example, in Fig. 2(b), compared with Series A, B has two singularities, and C has time shifts and warpings. By calculating, $d^E(A,B)=0.87$, $d^C(A,B)=0.56$ and $d^C(A,C)=0.53$, which means that Canberra distance can effectively eliminate the impact of singularities, but the measuring effect will become very poor once time shifts and warpings occurs.

(3) DTW

For solving the problem of time shifts and warpings, many approaches have been developed in recent years, and the most common one is DTW. DTW (Sakoe and Chiba, 1978; Sharabiani et al., 2018) is can measure the similarity of time series with different lengths, which minimizes the distance between two segmented series by constructing an optimal warping path. There are two steps of DTW. The first step is computing the distance matrix $\boldsymbol{D}(\boldsymbol{p_k})$ $(1\le k\le K)$.

$$\boldsymbol{D}(\boldsymbol{p_k})=\begin{bmatrix} d^E(x_1,y_1) & \cdots & d^E(x_1,y_m) \\ \vdots & \ddots & \vdots \\ d^E(x_n,y_1) & \cdots & d^E(x_n,y_m) \end{bmatrix}_{n\times m} \tag{3}$$

where $d^E\left(x_i,y_j\right)=\sqrt{\left(x_i-y_j\right)^2}\, i=1,2,\cdots,n; j=1,2,\cdots,m$.A warping path $P=(p_1,p_2,\cdots,p_K)$ means the mapping information derived from the two time series. $K(max(n,m)\le K\le n+m-1)$is an integer representing length of the path, and p_k denotes the matching relation between two points x_i and y_j, i.e., $d(p_k)=d^E\left(x_i,y_j\right)$. The second step is to find the shortest path in P, whose formula is as follows.

$$d^{DTW}=min\left\{\sum_{k=1}^{K}d(p_k)\right\} \tag{4}$$

Different from traditional similarity measurements, DTW can match data by a "one-to-many" mechanism to cope with time shifts and warpings, which makes it been widely applied in many fields. For instance, Wiemann and Amman (2020) chose a non-parametric DTW technique in an application to examine the temporal alignment and similarity across economic time series.

Although DTW solved the problem of point-to-point matching of time series, it could not reduce the negative influence of singularities since it still adopts Euclidean distance to construct the distance matrix. For instance, in Fig. 2(c), Series B and C have the time shifts and warpings compared with A, and C has two singularities at the same time. The computing results are $d^{DTW}(A,B)=0.61$ and $d^{DTW}(A,C)=0.90$, since DTW can measure the similarity of time series at different time but cannot eliminate the impact of singularities.

To summarize, the above three measurements have their respective advantages and disadvantages. For clearly presenting, Table 1 listed these strengths and weaknesses. Here, it is worth pointing out that all the above methods cannot reflect the characteristics of time, that is, the influence of time series on the current data decreases with the advance of time. Therefore, a novel similarity measurement is proposed in this paper and the details are shown in Section 3.

2.2. Prediction methods

Nowadays, more and more forecasting methods have been developed and applied to the stock market (Fenghua et al., 2014; Khoojine and Han, 2020; Zhang and Lou, 2020). In this paper, for verifying the performance of our similarity measurement, we take the autoregressive

Table 1
Three characteristics for similarity measurement.

	Eliminating the impact of singularity	Coping with time shifts and warpings	Reflecting time characteristics
Euclidean distance	weak	weak	weak
Canberra distance	strong	weak	weak
DTW	weak	strong	weak

(AR) model and neural network as the prediction framework since they are two of the most common and basic methods.

The autoregressive (AR) model (Chatfield, 1975) is a statistical way of dealing with time series, which is widely used in the prediction of economics, information science and so on. If p denotes the order of the model, the calculation of AR(p) is as follows.

$$x_t = \varphi_1 x_{t-1} + \varphi_2 x_{t-2} + \cdots + \varphi_p x_{t-p} + \varepsilon_t \tag{5}$$

where $\varphi_1, \varphi_2, \cdots, \varphi_p$ denotes the model parameter, x_t is the dependent variable, i.e., it can be expressed by the previous period values, $\{\varepsilon_t\}$ is a white noise sequence, whose mathematical expectation is zero and variance is σ^2. It also represents random factors that cannot be modeled.

Neural network (Müller and Reinhardt, 1990) is a dynamic system with a directed graph topology, which carries out information processing by making state response to continuous or intermittent input. BP neural network, as one of the most common neural networks, shown in Fig. 3, is a multi-layer feed-forward network trained by error inverse propagation algorithm and it is composed of input layer, hidden layer and output layer.

3. Method

In this section, we detailly introduce the proposed similarity measurement and some prediction schemes. Specifically, Section 3.1 gives a detailed description of our measurement firstly. Subsequently, Section 3.2 shows the prediction schemes based on AR models and neural network.

3.1. Dynamic multi-perspective similarity measurement

To address the limitations of the above approaches, we propose a novel similarity measurement, which is called the dynamic multi-perspective personalized similarity measurement (DMPSM), and the specific steps are as follows.

Firstly, we assign weights to the segmented series for reflecting the time characteristic after segmenting the whole stock sequences into small pieces of time series. The principle of weight is that the closer the factor to the current time, the greater the weight. In other words, for the time series $X = (x_1, x_2, \cdots, x_n)$, each of the factors is given a weighting on a scale of ω_1 to ω_N, where $\omega_1 < \omega_2 < \cdots < \omega_n$and $\omega_1 + \omega_2 + \cdots + \omega_n = 1$. That is,

$$X' = (\omega_1 x_1, \omega_2 x_2, \cdots, \omega_n x_n) = (x'_1, x'_2, \cdots, x'_n) \tag{6}$$

For example, suppose the length of sample stock series $X = (x_1, x_2, \cdots, x_n)$ is n, and the weight number is k, for the weight series,

$$W = \left(\frac{1}{1 + \sum_{i=2}^{n}(i+k)}, \frac{2+k}{1 + \sum_{i=2}^{n}(i+k)}, \cdots, \frac{n+k}{1 + \sum_{i=2}^{n}(i+k)} \right) \tag{7}$$

From Eq. (7), it is obvious that the greater the value of k, the greater the weight of the latter part, which can reflect the time characteristics.

Secondly, Canberra distance is embedded into the DTW not only for eliminating the impact of the singularities, but also for coping with time shifts and warpings. Specifically, we utilize Canberra distance to construct the DTW matrix between the weighted series $X' = (x'_1, x'_2, \cdots, x'_n)$and $Y' = (y'_1, y'_2, \cdots, y'_m)$, that is,

$$D'(p_k) = \begin{bmatrix} d^C(x'_1, y'_1) & \cdots & d^C(x'_1, y'_m) \\ \vdots & \ddots & \vdots \\ d^C(x'_n, y'_1) & \cdots & d^C(x'_n, y'_m) \end{bmatrix}_{n \times m} \tag{8}$$

where $d^C\left(x'_i, y'_j\right) = \frac{|x'_i - y'_j|}{|x'_i| + |y'_j|} (i = 1, 2, \cdots, n; j = 1, 2, \cdots, m)$.

Then, the calculation formula of DMPSM can be described as

$$d^{\mathrm{DMPSM}} = min\left\{ \sum_{k=1}^{k} \boldsymbol{d}'(\boldsymbol{p_k}) \right\} \tag{9}$$

where $d'(p_k) = d^C\left(x'_i, y'_j\right)$ (here, the meaning of the symbols in the equation is same with above).

The DMPSM has three major advantages. For better explanation, some examples are given in Fig. 4, where Series A and Series B are two different time series. (1) In Fig. 4(a), it is obvious that Series B has a singularity, but $d^{DMPSM}(A, B) = 0.35$, indicating that DMPSM can eliminate the impact of singularities. (2) In Fig. 4(b), Series B has time shifts and warpings with A, however, the calculating result of DMPSM shows that it can cope with time shifts and warpings. (3) In Fig. 4(c), the latter part of the two series is more similar than that of the first half part, and by computing, the result of DMPSM is the minimum, which means that DMPSM can reflect the personalized characteristics of time series. In addition, since the DMPSM is flexible, there is no limit of range for our measurement, and different adjustments can be made according to different situations.

Finally, the whole flowchart of the proposed method framework is displayed in Fig. 5, comprising the following steps: (a) weighting the segmented series after segmenting the whole stock sequences into small pieces of time series, that is, the closer the factor to the current time, the greater the weight. (b) the similarities are measured by DMPSM between

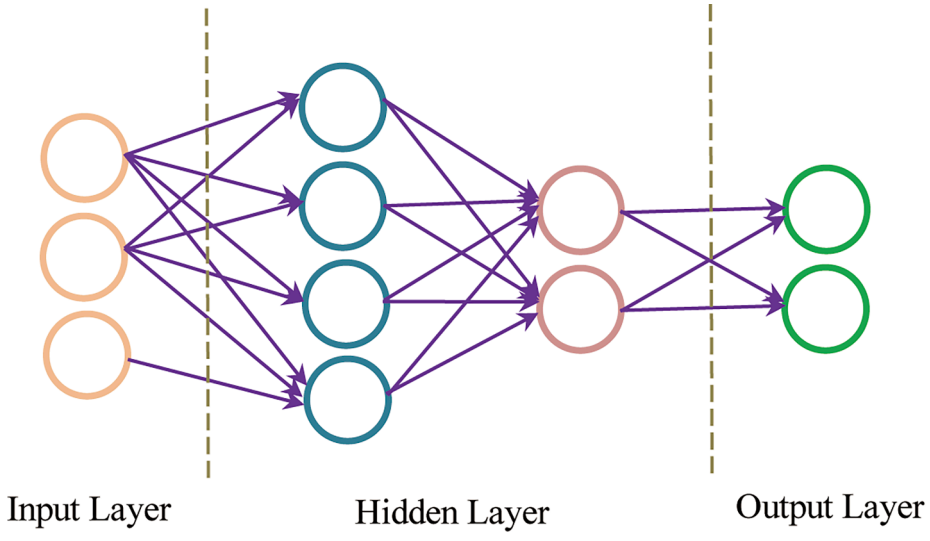

Fig. 3. The infrastructure of the BP neural network.

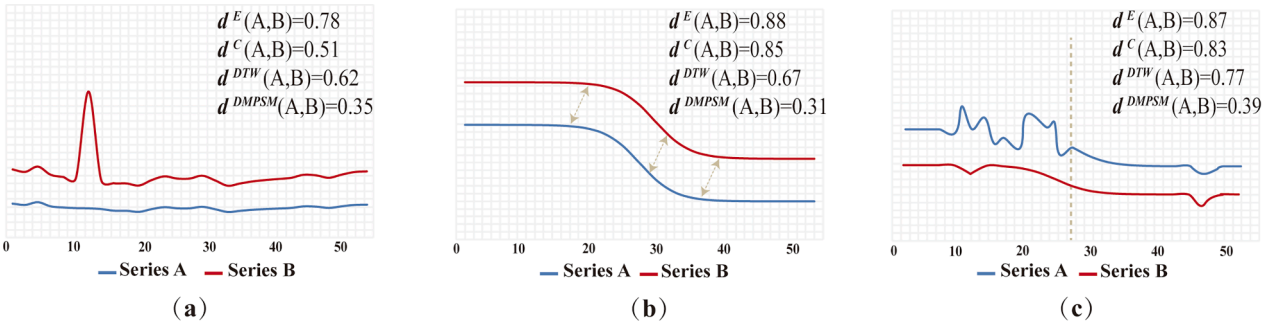

Fig. 4. The three advantages of DMPSM.

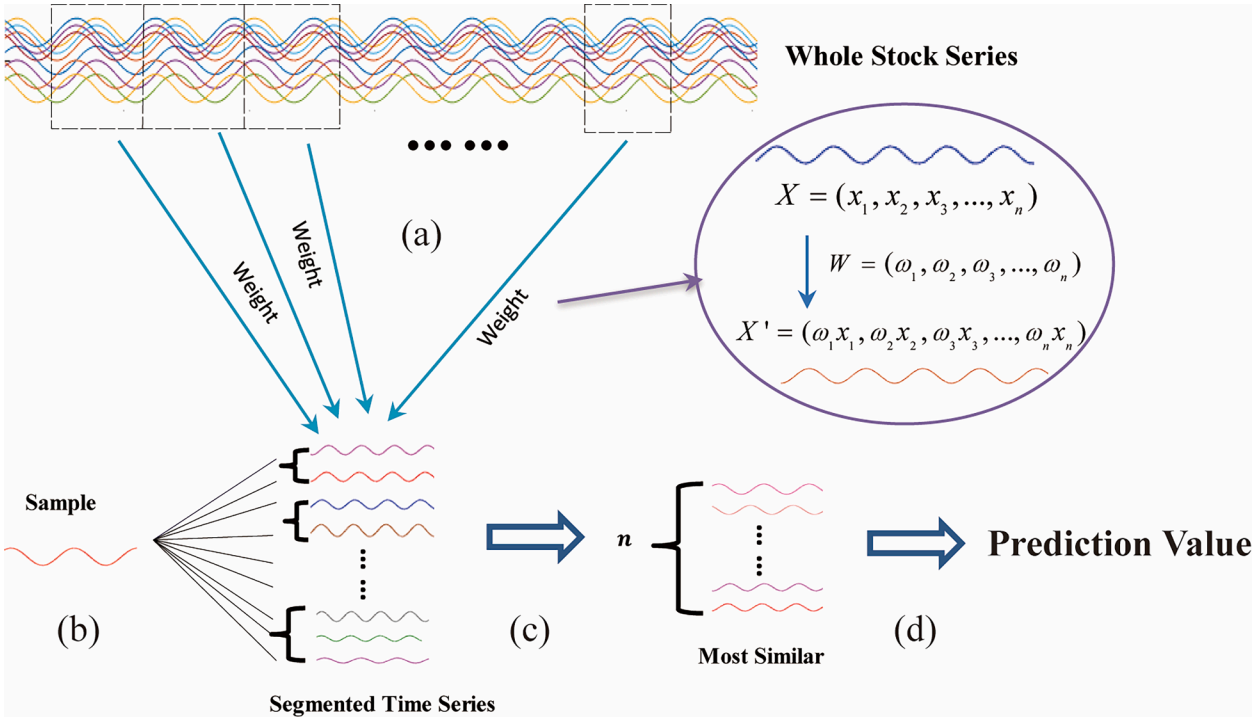

Fig. 5. Overview of the proposed method framework. Each color of line represents a whole stock series. $W = (\omega_1, \omega_2, \omega_3, \cdots, \omega_n)$ represents the weight series, $X = (x_1, x_2, x_3, \cdots\cdots, x_n)$ represents a segmented time series, and $X'(\omega_1 x_1, \omega_2 x_2, \omega_3 x_3, \cdots, \omega_n x_n)$represents the weighted time series.

the sample series and segmented time series. (c) the top n most similar time series with the sample series are picked out. (d) these selected n time series are used for prediction.

3.2. The prediction schemes

Based on the above prediction models, we adopt four kinds of prediction schemes for better prediction performance, which is shown in Fig. 6. Let $X_{sample} = (x_1, x_2, x_3 \cdots\cdots x_m)$ denote the sample series, $X_i = (x_{i1}, x_{i2}, x_{i3} \cdots\cdots x_{im})$ $(1 \le i \le n)$ denote the top n time series which are most similar with the sample time series, $\overline{X} = \left(\overline{x}_1, \overline{x}_2, \overline{x}_3, \cdots\cdots, \overline{x}_{m-1}, \overline{x}_m\right)$ denote the average series of $X_1, X_2, \cdots, X_n$, $x_{i(m+1)}$ denote the real value of $X_i$on the $(m+1)$-th day, $\widehat{x}_{m+1}$, $\widehat{x}_{i(m+1)}$ and $\widehat{\overline{x}}_{m+1}$ denote the prediction value of X_{sample}, X_i and $\overline{X}$ respectively. The four kinds of schemes are as follows.

Scheme 1: Only utilizing the sample series to predict while the similar series are not used. Specifically, the segmented time series are predicted by constructing AR models according to Eq. (5) or a multilayer BP neural network to build the function relation. This scheme is denoted as AR or BP and shown in Fig. 6(a).

Scheme 2: The top n similar series $X_1, X_2, \cdots, X_n$ are picked out and their mean on the $(m+1)$-th day represents the predicted value of the sample time series on the $(m+1)$-th day. The specific calculation is as follows. This scheme is denoted as AVG and shown in Fig. 6(b).

$$\widehat{x}_{m+1} = \frac{\sum_{i=1}^{n} x_{i(m+1)}}{n} \tag{10}$$

Scheme 3: We select the top n similar series $X_1, X_2, \cdots, X_n$ and construct autoregressive models respectively. Then, the mean of their autoregressive value on the $(m+1)$-th day represents the predicted value. The calculation formula is as follows. This scheme is denoted as AR-AVG or BP-AVG and shown in Fig. 6(c).

$$\widehat{x}_{m+1} = \frac{\sum_{i=1}^{n} \widehat{x}_{i(m+1)}}{n} \tag{11}$$

Scheme 4: The top n similar series $X_1, X_2, \cdots, X_n$are selected and be averaged to obtain a new time series $\overline{X} = \left(\overline{x}_1, \overline{x}_2, \overline{x}_3, \cdots\cdots, \overline{x}_{m-1}, \overline{x}_m\right)$. Its calculation formula is as follows.

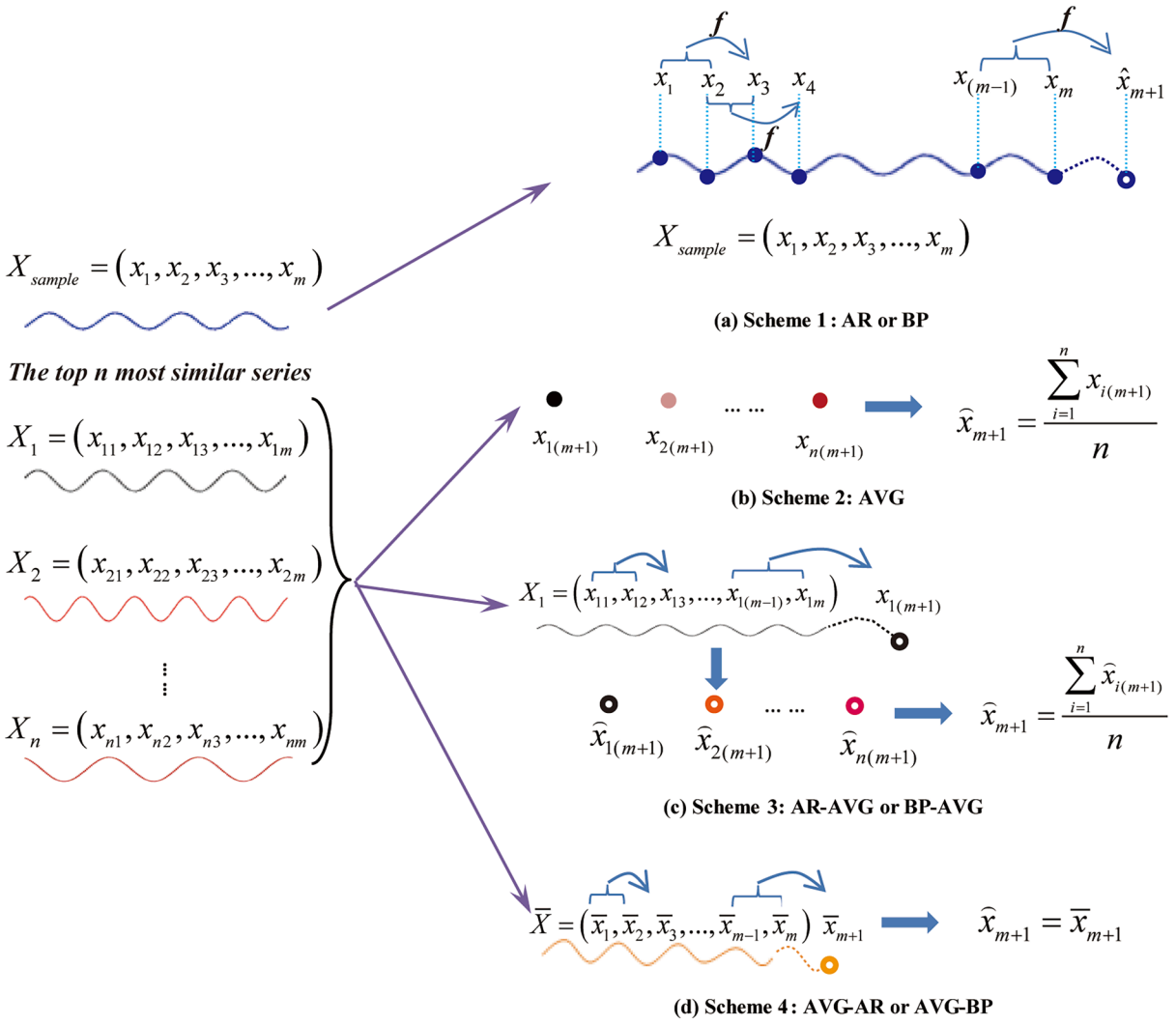

Fig. 6. Four kinds of prediction schemes. Each color represents a segmented stock series. The hollow points represent the prediction values and the solid points represent the real values.

$$\overline{x}_i = \frac{\sum_{i=1}^{n} x_{it}}{n}, t = 1, 2, \cdots, m \tag{12}$$

Afterward, the new time series is predicted by constructing autoregressive models and the predicted value of X_{sample} is represented by the autoregressive or BP value of $\overline{X}$, that is, $\hat{x}_{m+1} = \overline{x}_{(m+1)}$. This scheme is denoted as AVG-AR or AVG-BP and shown in Fig. 6(d).

4. Experiments and analysis

This section compares the performance of DMPSM and the performance based on different similarity measurements. Firstly, we introduce the experiment data in Section 3.1. Then, two experiments are designed and the experiment results and analysis are shown in Section 3.2 and Section 3.3. Finally, Section 3.4 discusses the strengths and weaknesses of DMPSM. The evaluation demonstrates that the fusion of Canberra distance and DTW can increase the accuracy of similarity measurement and further improve the prediction performance. At the same time, the similarity measurement of DMPSM with the prediction of the average of the similar time series after autoregression can achieve the best result.

4.1. Data acquisition and preprocessing

The data for our experiments is 285 stocks from the Shanghai Stock Exchange, and the total number of each stock is 1210 collected from trading days ranged from January 1st, 2016 to November 3rd, 2017.

At the same time, considering that each stock has a different price and some even vary widely from stock to stock in price, the data preprocessing is necessary in our experiment: (1) Data normalization. We first have carried on the normalized data processing, that is, using the day's stock price minus the stock price of the previous day, and then divided the former day's stock price, so that the growth rate of each stock every day will be computed, and the specific calculation formula is as follows.

$$x'_i = \frac{x_{i+1} - x_i}{x_i}, i = 1, 2, \cdots \tag{13}$$

where x'_i denotes the growth rate of stocks and x_i denotes the close price of the stocks on $i-th$ day. (2) Data screening. The Shanghai stock exchange opened only on weekdays, thus the weekend will be displayed for Friday's closing price, which leads to the growth rate become zero in the weekend. So that, we take out the values of zero for countering this effect.

4.2. The comparison of different similarity measurements

In order to prove the performance of DMPSM, we conduct an experiment, which compares the predictive capabilities of DMPSM,

Euclidean distance, Canberra distance and DTW. Firstly, we cut the 285 stocks into segmented series and choose the number 10 as the length of segmented stock series, that is, $n = 10$. At the same time, we take the last segment series of each stock as the test series, and predict the last stock price of the 285 stocks respectively. Secondly, DMPSM, Euclidean distance, Canberra distance and DTW are used to measuring the similarity between test series and other segmented series. In particular, for the weight of DMPSM, we select three weight numbers, i.e.,$k = 1,2,3$, and according to the weight principle in Eq. (7), we can get three weight series:-
$W_1 = (1,2,\cdots,10)/55; W_2 = (1,3,\cdots,19)/100; W_3 = (1,11,\cdots,91)/460$. Thirdly, the four prediction schemes in Section 3.2 are utilized to forecast and obtain the final predicted value. Fourthly, for comparing the results more directly, the mean absolute error (MAE) and root mean square error (RMSE) are selected as evaluation metrics to judge the prediction performance. Their definitions are given in Table 2, where $X = (x_1, x_2, \cdots, x_n)$ denotes the real time series, $\widehat{X} = \left(\widehat{x}_1, \widehat{x}_2, \cdots, \widehat{x}_n\right)$ denotes the predicted time series. The smaller values of these indexes, the more accurate of prediction.

Table 3 demonstrates the experiment results, where AR and BP represent Scheme 1 that only utilizing the sample series to predict, AVG represents Scheme 2 which refers to the average of the similar time series, AR-AVG and BP-AVG represent Scheme 3 that is the average of the similar time series after AR or BP processing, AVG-AR and AVG-BP represent Scheme 4 which refers to the AR or BP processing after averaging the similar time series. In the Table 3, the columns represent the similarity measurements, rows represent the prediction methods, and the bold digits in each row represent the smallest value of the evaluation metrics, which refers to the best measurement under this prediction method.

Based on Table 3, we can make the following observations. (1) Different evaluation metrics shows different results. However, the DMPSM generally works better than other measurements. (2) Especially, for this experiment, the DMPSM with W_3 is the best similarity measurement among the four kinds of methods. (3) At the same time, even if utilizing the same prediction methods, different prediction schemes also have different results. The AR-AVG has the best prediction performance among all the schemes.

To further clarify the reason that why the DMPSM can get better performance, we randomly select a sample series and extract the top five similar series calculated by using DMPSM. The curves of these series are shown in Fig. 7, where the red line denotes the sample series, while others denote the similar series and Table 4 shows the measurement results between different parts of one series by using the DMPSM, where the first half refers to the similarity between the first five points, the second half refers to the similarity between the latter points and the bold digits represent the better performance of the two parts.

From the figure and the table, we can conclude that: (1) The DMPSM can reflect that different time has different impact on current data. From Table 4, it is obvious that most similarity values of the second half are smaller than the first half, which indicates that the latter part of series are more similar than the former part. It can be explained by that different weights are assigned to the series before calculating the similarity and the weighted principle is that the closer the data to the current data, the greater the weight. (2) The DMPSM can effectively eliminate the impact of singularities. Fig. 7(b) shows the trend of the sample series and Series 2. Compared the two series, we can find that there is a singularity (the seventh point) on Series 2, which will increase the direct distance by adopting Euclidean distance. That is to say, the reason why Series 2 is more similar with the sample is that DMPSM has effectively eliminated the influence of the singularity due to the adaptation of Canberra distance. (3) The DMPSM can cope with time shifts and warpings. Fig. 7(c) shows the warping path of the sample series and Series 2, where the horizontal axis represents the points in sample series, the vertical axis represents the points in Series 2, the red areas represent the DMPSM path and the yellow areas represent the point-to-point path. As shown in Fig. 7(c), it is obvious that since employing DTW as the framework, the DMPSM can automatically find the optimism warping path so that it can measure the divergence between two time series with different phases and lengths. Based on above analysis, we can see that the reason why the DMPSM has the best performance may be that the DMPSM fully considers the characteristics of stock series, that is, it can eliminate the impact of singularities, reflect time characteristics and cope with time shifts and warpings.

Table 2
The definitions of the two evaluation metrics.

Metrics	Expression
MAE	$\frac{1}{n}\sum_{i=1}^{n}\left\|\widehat{x}_i - x_i\right\|$
RMSE	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\widehat{x}_i - x_i)^2}$

4.3. The analysis of DMPSM performance

In order to find out what kind of sequence type the DMPSM is most applicable to, we design another experiment. Firstly, a section from each of the 285 stocks are cut as test data, and then the DMPSM is used to pick out 5 series with the lowest error and 5 series with highest error. The curves and boxplots are shown in Fig. 8. In Fig. 8, (a) and (b) represent the curves and boxplots of the 5 series with the lowest error, while (c) and (d) represent the 5 series with the highest error. In the boxplots, we can see the upper margin, upper quartile, median, lower quartile and lower margin clearly from top to bottom. Meanwhile, the outliers are also shown in boxplots.

For each curve and boxplot visible in Fig. 8, we can find that the series in Fig. 8(a) are flatter than that in Fig. 8(c), which is also demonstrated in Fig. 8(b) and Fig. 8(d). Compared Fig. 8(b) with Fig. 8(d), it is obvious that there is only one outlier in Fig. 8(b), while there are 4 outliers in Fig. 8(d). Besides, the boxes in Fig. 8(b) are more compact than that in Fig. 8(d), which indicates that the data of the series with lowest error have little fluctuation while others are more volatile. Therefore, it can be concluded that the DMPSM has a good prediction effect for stationary series, but the effect will be significantly reduced for excessively volatile data.

4.4. The strengths and weaknesses of DMPSM

From the above experiments and results, compared with Euclidean distance, Canberra distance and DTW, it is obvious that the DMPSM has a good prediction ability for stock time series. Firstly, it can reflect the personalization of time by giving different weights to different time, thus solving the problem of time sensitivity in stock series. Then, it can eliminate the impact of singularities by adopting Canberra distance, so that it can reduce the negative effect of singularities on stock prediction. Moreover, it can cope with time shifts and warpings by taking DTW as the framework of measurement, which solves the matching problem of stock sequences with different lengths and time periods.

Although the DMPSM overcomes these three problems in stock prediction, it still has high computing complexity and for the length of segmented series in this paper, just value it according to experience, without doing relevant test.

5. Conclusion and future work

In this paper, we propose a new similarity measurement for predict the stock price accurately which is called DMPSM. In order to validate the effectiveness of our method, we firstly segmented and weighted the stock time series. Then, embedding Canberra distance into DTW and finally combining with the autoregression and neural network

Table 3
The results by using different similarity measurements and prediction methods.

		d^E	d^C	d^{DTW}	$d^{DMPSM-W_1}$	$d^{DMPSM-W_2}$	$d^{DMPSM-W_3}$	
MAE	AVG	0.01447	0.01436	0.01526	0.01529	0.01550	**0.00349**
	AR	0.01501	0.01502	0.01502	0.01545	0.01550	**0.00343**
	AR-AVG	0.01417	0.01397	0.01395	0.01453	0.01478	**0.00358**
	AVG-AR	0.01453	0.01422	0.01459	0.01497	0.01534	**0.00345**
	BP	0.11749	0.13221	0.14516	0.10361	0.20303	**0.00431**
	BP-AVG	0.01447	0.01437	0.01526	0.01529	0.01549	**0.00349**
	AVG-BP	0.22559	0.16759	0.18067	0.17172	0.14345	**0.01268**
RMSE	AVG	0.14066	0.14512	0.15087	0.13675	0.13216	**0.13026**
	AR	0.14566	0.14565	0.14566	0.12861	0.12550	**0.12461**
	AR-AVG	0.14285	0.14149	0.14766	0.13398	0.13023	**0.12967**
	AVG-AR	0.14460	0.14262	0.14630	0.12172	0.12047	**0.12025**
	BP	0.56316	1.012812	0.09715	**0.07690**	0.84655	0.23377
	BP-AVG	0.14066	0.14512	0.15087	0.13673	0.13216	**0.13023**
	AVG-BP	2.00531	1.45413	1.50917	**0.33279**	0.189370	0.66783

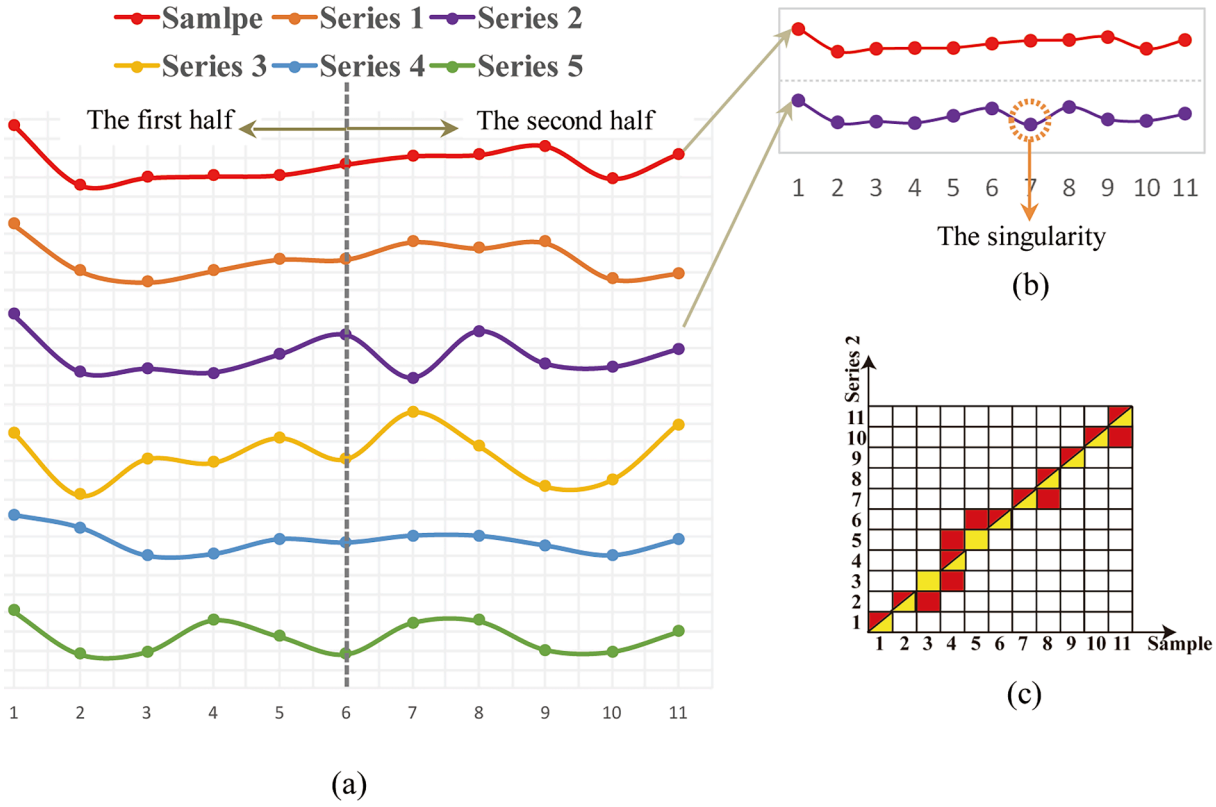

Fig. 7. The most similar series selected by DMPSM.

Table 4
The measurement results between different parts of one series by using the DMPSM.

	Series 1	Series 2	Series 3	Series 4	Series 5
The first half	1.2249	**1.2979**	2.6113	1.5316	1.4458
The second half	**0.6100**	2.1339	**0.8410**	**1.1974**	**1.3439**

prediction methods to forecast the stock price. The experimental results show that: (1) different weights of stock data in different periods can reflect the personalized characteristics of stock time series, and further effectively improve the accuracy of prediction. (2) Integrating Canberra distance into DTW can not only effectively eliminate the negative impact brought by the singularity in the series, but also cope with time shifts and warpings. (3) The similarity measurement of DMPSM and the prediction method of the autoregression after averaging the similar time series have the capability of prediction in short term and the guiding significance for investment in short term. Through the analysis of the test results and the algorithm, it is obvious that our method has a certain application prospect in stock prediction.

However, the similarity measurement and prediction method proposed in this paper may be a preliminary exploration and the future research can focus on the following aspects: (1) For the length of segmented series, it is an important parameter for the DMPSM that can influence the final performance to some extent. Therefore, an in-depth study on the length of segmented series can be conducted in the future work. (2) In this paper, we mainly focus on the improvement of similarity measurement, ignoring the combination of computing complexity. Therefore, we will mainly study on the optimize the algorithm in the next stage.

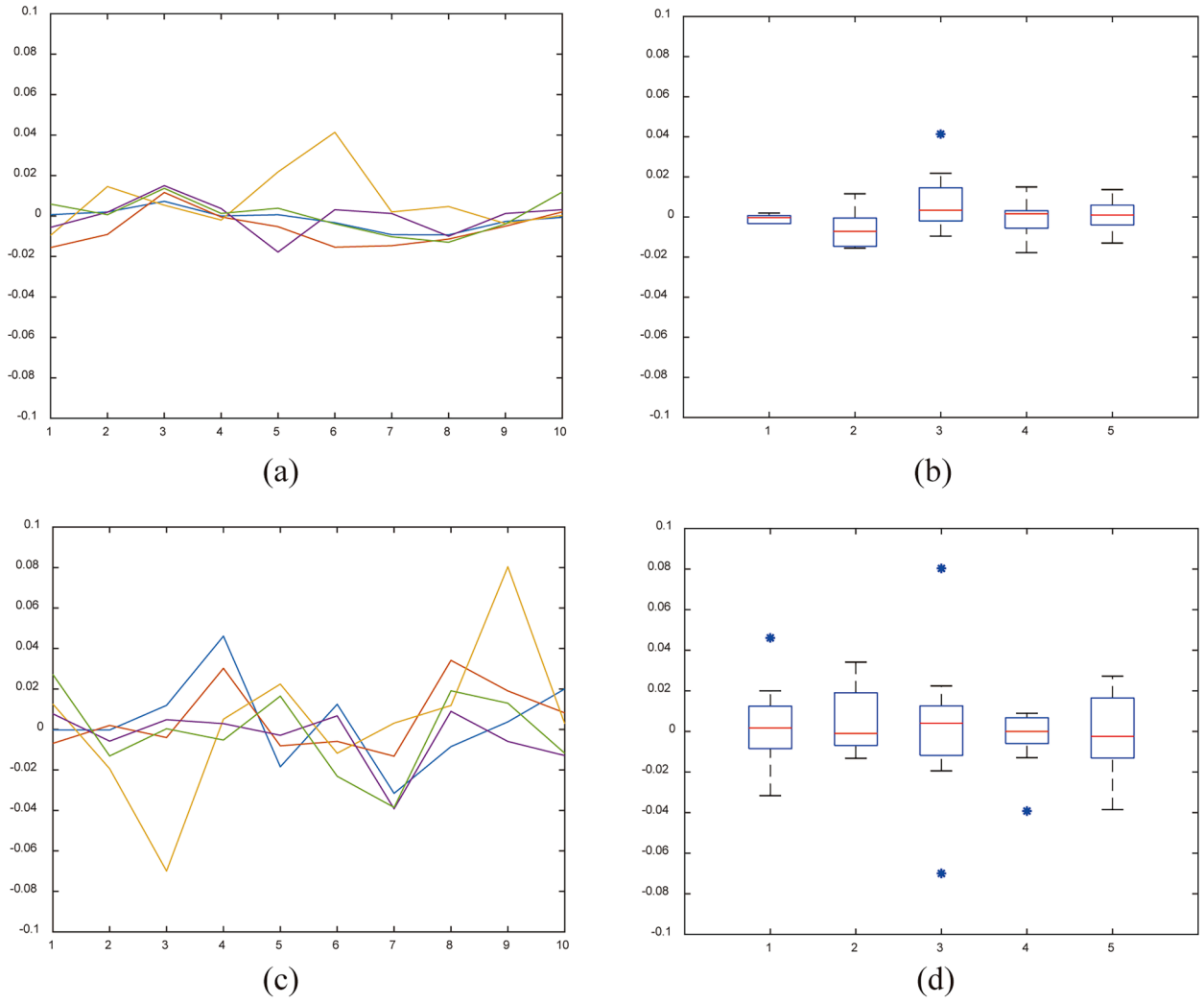

Fig. 8. The curves and boxplots of the 5 series with the lowest errors and 5 series with the highest errors.

CRediT authorship contribution statement

Feng Zhao: Conceptualization, Methodology. **Yating Gao:** Conceptualization, Software, Writing - original draft, Methodology, Formal analysis, Investigation, Validation. **Xinning Li:** Validation. **Zhiyong An:** Writing - review & editing. **Shiyu Ge:** Writing - review & editing. **Caiming Zhang:** Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (61773244, 61976125, 61272319, 61873117, 61972235 and 61976124), Yantai Key Research and Development Program of China (2017ZH065, 2019XDHZ081), Shandong Provincial Key Research and Development Program of China (2019GGX101069), and Wealth management characteristic construction project of Shandong Technology and Business University (2019ZBKY032).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.eswa.2021.115217.

References

Anton, H. (1993). *Elementary linear algebra* (7th ed.). New Jersey: Wiley (Chapter 3).
Bai, C., White, T., Xiao, L., Subrahmanian, V. S., & Zhou, Z. (2019). C2P2: A collective cryptocurrency up/down price prediction engine. 2019 IEEE International Conference on Blockchain (Blockchain), Atlanta, GA, USA.
Bijnen, E. J. (1973). *Cluster analysis* (1st ed.). Dordrecht: Springer (Chapter 2).
Chatfield, C. (1975). *The analysis of time series: Theory and practice* (1st ed.). Boston, MA: Springer (Chapter 4).
Chen, Q., Yan W., Gang, H., & Lei, G. (2015). A Novel Method Based on Data Visual Autoencoding for Time Series Similarity Matching. The 27th Chinese Conference on Control and Decision-Making, Qingdao, China.
Faisal, M., Zamzami, E. M., & Sutarman. (2020). Comparative analysis of inter-centroid K-means performance using Euclidean distance, Canberra distance and Manhattan distance. Journal of Physics: Conference Series, 1566(1):012112.
Fenghua, W., Jihong, X., Zhifang, H., & Xu, G. (2014). Stock price prediction based on SSA and SVM. *Procedia Computer Science, 31*, 625–631.
Groth, S., & Muntermann, J. (2011). An intraday market risk management approach based on textual analysis. *Decision Support Systems, 50*(4), 680–691.
Han, T., Peng, Q., Zhu, Z., Shen, Y., Huang, H., & Abid, N. N. (2020). A pattern representation of stock time series based on DTW. *Physica A*, 124161.
Juszczuk, P., Kozak, J., & Kania, K. (2020). Using similarity measures in prediction of changes in financial market stream data—Experimental approach. *Data & Knowledge Engineering, 125*, 101782. https://doi.org/10.1016/j.datak.2019.101782
Khoojine, A. S., & Han, D. (2020). Stock price network autoregressive model with application to stock market turbulence. *The European Physical Journal B, 93*(7), 1–15.

Kuremoto, T., Kimura, S., Kobayashi, K., & Obayashi, M. (2014). Time series forecasting using a deep belief network with restricted Boltzmann machines. *Neurocomputing, 137*, 47–56.

Li, G., Wang, Y., Zhang, L., & Zhu, X. (2009). Similarity measure for time series based on piecewise linear approximation. International Conference on Wireless Communications & Signal Processing WCSP 2009, Nanjing, China.

Sun, L., Wang, K., Balezentis, T., Streimikiene, D., Zeng, S., & Zhang, C. (2021). Extreme point bias compensation: A similarity method of functional clustering and its application to the stock market. *Expert Systems with Applications, 164*.

Mei, J., Liu, M., Wang, Y.-F., & Gao, H. (2016). Learning a mahalanobis distance-based dynamic time warping measure for multivariate time series classification. *IEEE Transactions on Cybernetics, 46*(6), 1363–1374.

Müller, B., & Reinhardt, J. (1990). *Neural networks* (1st ed.). Berlin, Heidelberg: Springer (Chapter 19).

Preethi, G., & Santhi, B. (2012). Stock market forecasting techniques: A survey. *Journal of Theoretical and Applied Information Technology, 46*(1), 24–30.

Qiang, W., & Vasileios, M. (2007). A dimensionality reduction technique for efficient time series similarity analysis. *Information Systems*.

Sakoe, H., & Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Trans on Acoustics Speech and Signal Processing, 26*(1), 43–49.

Sharabiani, A., Darabi, H., Harford, S., Douzali, E., Karim, F., Johnson, H., & Chen, S. (2018). Asymptotic dynamic time warping calculation with utilizing value repetition. *Knowledge and Information Systems, 57*(2), 359–388.

Tsinaslanidis, P. E. (2018). Subsequence dynamic time warping for charting: Bullish and bearish class predictions for NYSE stocks. *Expert Systems with Applications, 94*(MAR.), 193–204.

Wiemann, T., & Amman, H. (2020). Intertemporal similarity of economic time series: An application of dynamic time warping. *Computational Economics, 56*.

Zhang, D., & Lou, S. (2020). The application research of neural network and BP algorithm in stock price pattern classification and prediction. *Future Generation Computer Systems*.

Zhang, L., Aggarwal, C., & Qi, G. J. (2017). Stock price prediction via discovering multi-frequency trading patterns. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 2141–2149).