ISSN: 2414-1895 DOI: 10.6919/ICJE.202006_6(6).0021

Factor Analysis based Similarity Dimension Reduction of Dynamic Time Warping

Pengcheng Zhao[1, a], Shibao Sun[1, b] and Fengkui Xu[1, c]

[1]School of Information Engineering, Henan University of Science and Technology, Luoyang 471000, China.

[a]zpchaust@163.com, [b]sunshibao@haust.edu.cn, [c]xfkhaust@163.com

Abstract

In a multivariate time series, key features are selected in high dimensional data and used as input information. This is critical to the accuracy of the time series model prediction. Dynamic time warping is often used to measure the temporal feature similarity. After feature selection, a model is built to improve the performance of the time series data model. However, this method does not suppress noise and outliers in the time series data. The method is less adaptable to temporal offsets and the computational complexity is too high. In view of this, this paper proposes a dynamic time warping (DTW) similarity dimension reduction method based on factor analysis (FA). First, multi-time data preprocessing is performed. Then, Augmented Dickey-Fuller test (ADF) is performed. This paper derives the symmetry of LB_Keogh. It proposes a dynamic time warping lower bound function LB_H with symmetry, and uses it to construct a distance matrix as a factor analysis covariance matrix. Finally, the output common factor is used to construct the Autoregressive Integrated Moving Average (ARIMA) model. The effectiveness and feasibility of the method are verified by actual data.

Keywords

Dynamic time warping, Factor analysis, Feature similarity, Itakura, Multiple time series.

1. Introduction

The time series is a set of random values arranged in chronological order, which is a quantitative representation of the short-term and long-term changes in one or more statistical indicators. Currently, there are time series applications in industrial production and economic data[1]. With the diversification of application requirements and data, the scale and dimension of timing are gradually increasing [2]. It is easy to generate a large amount of noise and irrelevant feature variables in the data, which easily cover the important features and increase the difficulty of constructing the prediction model. In order to improve the multi-time data mining ability, the data dimension reduction method is usually used to reduce the complexity of the original data, and then the prediction can be modeled in a low-dimensional space.

Therefore, in the multivariate time series, the observed similarity between features is mined from the data. Further, a correlation feature having a large contribution rate to the prediction model is calculated as model input information. Such an approach is currently a mainstream solution [3]. At present, in most multi-time sequence application fields, the dynamic time warping (DTW) method is mainly used to calculate the similarity of data features [4]. The advantages of the method are: ①It can solves the defect of "one-to-one" matching of Euclidean distance, and can match the "one-to-many" time point; ② It can measure the similarity between features in non-equal length time series data[5]. However, the DTW algorithm also has shortcomings in the face of high-dimensional time

ISSN: 2414-1895 DOI: 10.6919/ICJE.202006_6(6).0021

series data: ①The time calculation complexity is high, O(n*m) time complexity is required for n*m time series data; ②The pairs of high dimensional multivariate time series The noise suppression is poor, and it is easy to produce malformation matching. ③The complex system data do not satisfy the triangle distance inequality [6].

Therefore, the higher computational complexity of DTW often takes the lower bound function of the calculated feature distance metric, and the precise distance measure is performed for the time series data that meet the constraint condition [7]. Distance metric needs to satisfy the following three conditions when calculating the distance between features: ① non-negative, ②symmetry, ③ triangle distance inequality. The commonly used distance lower bound functions is LB_Keogh [8-9]. The best and most commonly used LB_Keogh lower bound function does not satisfy symmetry and triangular inequalities in the strict sense. Scholars have done a lot of research on the problem of reducing the computational complexity of DTW dynamic time warping by multi-sequence data similarity matching. In 2017, Jiang J et al. [10] introduced the morphological factors and proposed a weighted dynamic time warping algorithm based on morphological features (SWDTW) to increase the robustness of the algorithm. Frank Höppner et al. [11] used the integrated preprocessing method to filter and scale the original data, and then measure the similarity to obtain a certain good performance index. BG Costa et al. [12] used KNN-DTW to classify time series data faults, and this method achieved good results. Chen, Yimin et al. [13] used K-Medoids based DTW method to analyze the applicability of urban internal structure. This method provided a basis for planning and evaluation. In 2018, Feng Yubo et al. [14] used piecewise linear regression intercept and distance parameter sets as time series data features to model the dimensionality reduction of the data, and the algorithm efficiency was improved. Zhang Fang et al. [15] used the new method of moving block bootstrap to approximate the statistical significance of the feature of the local similarity score of multivariate time series and reasonably controlled the error rate. Boulnemour et al. [16] proposed the combination of SEA and DTW to apply the QP-DTW method in PTB diagnosis. Zhang R et al. [17] used FA-DTW to normalize features with respect to KNN processing. In 2019, Qiao Meiying et al. [18] combined information entropy with dynamic time warping, minimized the loss function and obtained the global optimal Mahalanobis distance to measure the similarity of multidimensional time series, and improved the model accuracy. Ye Yanqing et al. [19] used the grid representation of time series data to segment and compile it, and then used the dynamic time warping method of the matrix to measure, and the algorithm obtained higher efficiency.

Factor analysis combines data with complex relationships into several minority factors to represent all data information [20]-[21]. Factor analysis shows the correlation between raw time series data and factors. It also clarifies the intrinsic association structure between variables. By combining with factor analysis, it avoids traversing the entire time series data. At the same time, it works well for a time series with a large number of related structures. In summary, the symmetry derivation of LB_Keogh requires the combination of the feature covariance matrix, which is DTW lower bound function in the factor analysis. And then propose a lower bound function LB_H with symmetry. This paper proposes a dynamic time warping similarity dimension reduction method based on factor analysis. The method can reduce the dimensionality of multivariate time series data and construct a covariance matrix between the eigenvalues. Then, the feature similarity is measured by the DTW distance lower bound function LB_H with symmetry, and the common factor is extracted as the input information. This method reduces the computational complexity of the DTW and improves the prediction accuracy of the ARIMA model.

2. Related Works

2.1 Multivariate Time Series

Dynamic time warping (DTW) is the best similarity in time series in the field of data science [22]. For a given multiple time series:

$$D_1 = \{a_1, a_2, \cdots, a_n\} \tag{1}$$

ISSN: 2414-1895 DOI: 10.6919/ICJE.202006_6(6).0021

$$D_2 = \{b_1, b_2, \cdots, b_m\} \tag{2}$$

According to the determined m × n time series matrix, the Euclidean distance is shown in formula (3).

$$D(n,m) = \sqrt{(a_1 - b_1)^2 + \cdots + (a_n - b_m)^2} \tag{3}$$

As shown in Fig. 1(a), the Euclidean distance can only measure the similarity between equal-length time series data. It does not reflect the similarity between the two sequences. Essentially it refers to the inter-data synchronization correlation [23]. The DTW showed in Fig. 1(b) can measure the similarity of time-series data features such as non-equal length and distortion.

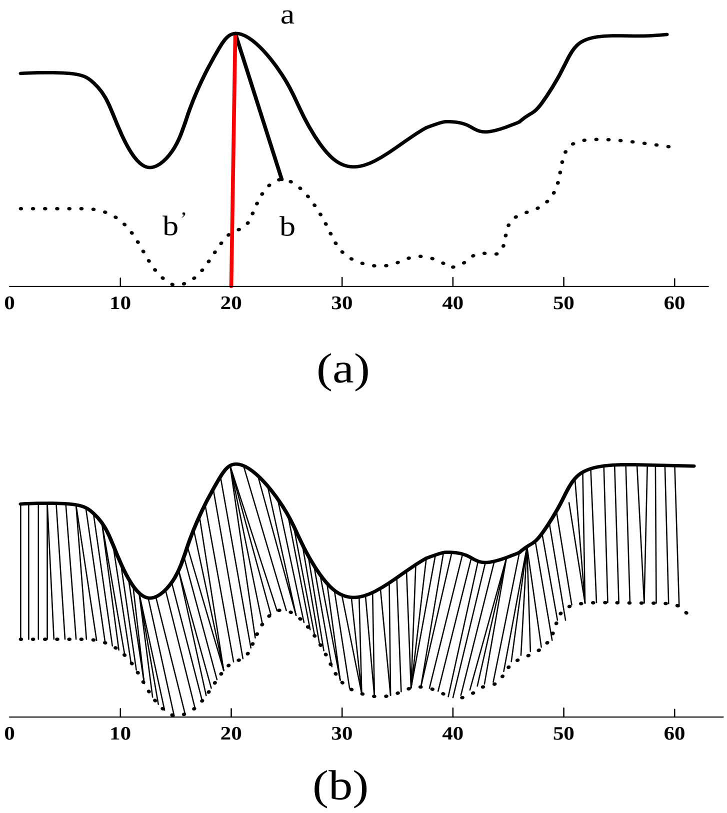

Figure 1. Comparison of Euclidean Distance and DTW.

For the multivariate time series data $D_{n\times m}$.The shortest bending path $P_{\min}$ is found by the dynamic time warping. The distance $D_{n\times m} = d(a_i, b_j) = \sqrt{(a_i - b_j)^2}$, $(1 \le i \le n, 1 \le j \le m)$, $P_{\min}$ is expressed as:

$$P_{min} = (p_1, p_2, \cdots, p_k)(max\{n,m\} \le k \le n + m - 1) \tag{4}$$

Where $p_k = (i,j)$ in the formula (4), the cumulative distance $\varphi(n,m)$ is the smallest, and the effective path P satisfies three requirements in the DTW:

① Monotonic, for $p_k = (i,j)$ and $p_{k+1} = (i',j')$, there is always $j' \ge j$;

② Boundary, $p_1 = (1,1)$, $p_k = (i,j)$;

③ Continuity, $i' \le i + 1$, $j' \le j + 1$ can be obtained from $p_k = (i,j)$, $p_{k+1} = (i',j')$.

Finally, DTW is expresse as:

$$DTW(D_1, D_2) = \min(\sum_{k=1}^{k} p_k) = \varphi(n,m) \tag{5}$$

The dynamic programming method is used to solve the formula (5), and the time series cost matrix φ is constructed. The cumulative distance in the matrix can be expressed by the following formula:

ISSN: 2414-1895 DOI: 10.6919/ICJE.202006_6(6).0021

$$\varphi(i,j) = D(i,j) + \min\begin{cases}\varphi(i-1,j-1);\\ \varphi(i-1,j);\\ \varphi(i-1,j-1).\end{cases} \tag{6}$$

Equation (6) represents the cumulative distance value of the current variable of the time series data group. That is, the current distance plus the minimum value of the accumulated distance values of the adjacent three variables. Where $\varphi(0,0) = 0$, $\varphi(i,0) = \varphi(0,j) = \infty$. The best choice path is the path where the cumulative distance reaches the minimum when $p_k = (i,j) = (1,1)$.

In multivariate time series data, the distance metric between feature variables is an important component. However, the computational complexity of DTW is O(n*m). In practical applications, the DTW distance lower bound function needs to be calculated first. The distance is measured by the qualified time series data, and then the calculation process of the algorithm is reduced by this method [24].

As a measure of distance similarity of MTS, LB_Keogh calculates a time series that exceeds the boundary of another time series. It uses the boundary approximation as the lower bound of the distance of the DTW [25].$[d_i^{min}, d_i^{max}]$ is the maximum and minimum of the MTS boundary. Then, LB_Keogh is expressed as:

$$LB\_Keogh(D,D') = \sum_{i=1}^{n}\sqrt{\begin{cases}(d_i - d_i^{\min})^2, d_i > d_i^{\min}\\ 0, others\\ (d_i - d_i^{\max})^2, d_i < d_i^{\max}\end{cases}} \tag{7}$$

However, the lower bound function LB_Keogh has asymmetry. It limits the scope of research on time series prediction and classification problems in applications. The function does not satisfy the distance triangle inequality. The inability to quickly retrieve data is the shortcoming of the current algorithm to be solved [26].

3. Method

3.1 Factor Analysis

Factor analysis (FA) is one of the most widely used algorithms for dimension reduction in complex systems [27]. The factor analysis first assumes that there are n samples with M evaluation indicators, and the data set is represented by matrix X as $D = (D_1, D_2, \cdots D_m)^T$. Let the obtained covariance matrix $cov(Z) = 1$. The obtained factor model is shown in equation (8).

$$\begin{cases}D_1 = a_{11}Z_1 + \cdots + a_{1m}Z_m + \theta_1\\ D_2 = a_{21}Z_1 + \cdots + a_{2m}Z_m + \theta_2\\ \cdots\cdots\\ D_m = a_{i1}Z_{1j} + \cdots + a_{ij}Z_{ij} + \theta_m\end{cases} \tag{8}$$

Where θ_m is a special factor, a_{ij} is expressed as a factor load, and the essence is the correlation coefficient between the common factor and the variable. Finally, the cumulative contribution rate $(\mu \geq 85\%)$ is calculated. The found common factor $Z = (Z_1, Z_2, \cdots Z_j,)^T$ is used as input information. The information examines the actual meaning of the common factors and builds the model.

3.2 Design and Analysis of Algorithms

Aiming at the problem of dimensionality reduction by similarity in multivariate time series data, this paper improves the DTW lower bound function LB_Keogh based on factor analysis dimension reduction and completes the distance metric. The algorithm in this paper uses the output factor to build a model to reduce the computational time complexity. The algorithm framework is shown in Figure 3.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202006_6(6).0021

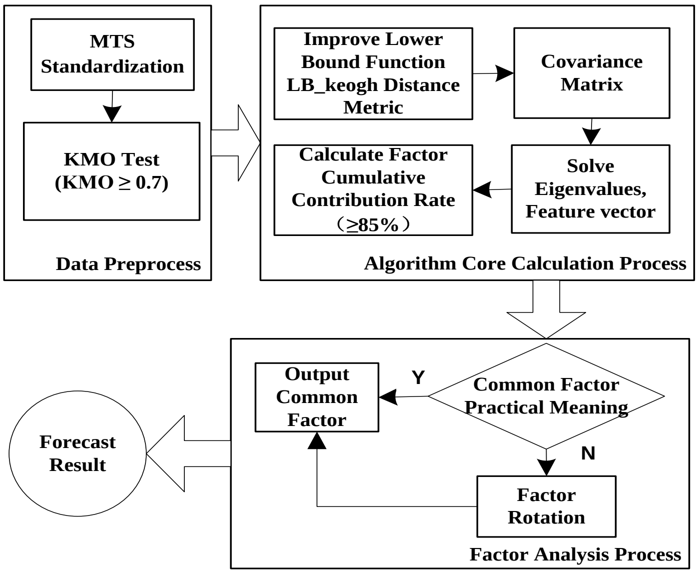

Figure 2. FA-DTW algorithm framework.

The main steps of the mathematical derivation of the dynamic time-bending similarity dimension reduction method based on factor analysis are as follows:

(1) This step preprocesses the MTS data. In view of this, the MTS is normalized before the multi-sequence similarity measure is performed. This article uses the "max-min" processing method [28]. Assuming that the multivariate timing is $D_{n\times m}$, the entire time series data maximum value d_{max} and minimum value d_{min} are calculated. Then normalize it:

$$\bar{d}_{ij} = \frac{d_{ij} - \min\{d_{ij}\}}{\max\{d_{ij}\} - \min\{d_{ij}\}}, (1 \le i \le n, 1 \le j \le m) \tag{9}$$

(2) This step checks if it meets the criteria for factor analysis. Firstly, the sample size should not be too small; then each feature should have a certain correlation, through KMO test (KMO ≥ 0.7); finally, the obtained common factor has practical significance.

(3) This step solves the multivariate time series data components. According to the multivariate time series data $D_{n\times m}$, it is assumed that n is a time series unit (year, month, day, hour, minute, etc.), m is expressed as the number of items, and d_{ij} is referred to as the year is i (1≤i≤n) in the j feature d_j represents the matrix formed, ie:

$$D_{nm} = \begin{pmatrix} d_{11} & d_{12} & \cdots & d_{1m} \\ d_{21} & d_{22} & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \cdots & d_{nm} \end{pmatrix} \tag{10}$$

(4) The covariance matrix is solved by the DTW distance lower bound function of the combination factor analysis. Distance is often used to describe the degree of correlation between two matrices. This correlation is constructed using the DTW distance lower bound function LB_Keogh. The distance correlation formula is calculated as following:

$$\left(Z_i, Z_j\right) = \begin{cases} 1 - d(Z_i, Z_j), i \ne j \\ 1, i = j \end{cases} (1 \le i \le n, 1 \le j \le m) \tag{11}$$

For $d(Z_i, Z_j)$ in equation (9) is the distance between the features of the covariance matrix Z_k. The inter-feature distance is expressed as a value of correlation.

When deriving the symmetry lower bound function, we introduce a curved path to constrain it and limit it to the vicinity of the diagonal of the matrix, and set the constant s [29]. As shown in Figure 3, the Itakura quadrilateral is shown, and s is a characteristic variable with respect to i.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202006_6(6).0021

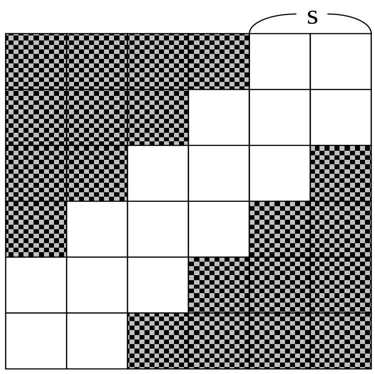

Figure 3. Itakura schematic.

Assuming that the lengths of the multivariate timings D and D' are v and v'. Both of these are non-equal length multi-time sequences. The global constraint value is set to s. Then, if there are the following conditions for multivariate timing:

$$d_i^{max} \in (d_{i-s}, d_{i+s}); d_i^{min} \in (d_{i-s}, d_{i+s}); \quad d_i^{'max} \in (d_{i-s}, d_{i+s}); d_i^{'min} \in (d_{i-s}, d_{i+s}). \tag{12}$$

Then define the LB_H lower bound function as：

$$LB\_H(D,D') = \sum_{i=1}^{n} \sqrt{\begin{cases} \left(d_i^{max} - d_i^{'min}\right)^2, d_i^{max} < d_i^{'min} \\ 0, others \\ \left(d_i^{'max} - d_i^{min}\right)^2, d_i^{'max} > d_i^{min} \end{cases}} \tag{13}$$

$$LB\_H(D',D) = \sum_{i=1}^{n} \sqrt{\begin{cases} \left(d_i^{'max} - d_i^{min}\right)^2, d_i^{'max} < d_i^{min} \\ 0, others \\ \left(d_i^{max} - d_i^{'min}\right)^2, d_i^{max} > d_i^{'min} \end{cases}} \tag{14}$$

In summary, it can be known from the square of the parentheses that $\forall i(1 \le i \le v \text{ or } 1 \le i \le v')$, $LB\_H(D, D') = LB\_H(D', D)$, This lower bound function has symmetry. Since the DTW distance metric derives LB_Keogh, the symmetrical LB_H is located at the lower bound of LB_Keogh. In view of this, it is easy to find the following restrictions：

$$0 \le LB\_H \le LB\_Keogh \le DTW \tag{15}$$

Then you can get：

$$LB\_H(D,D') = \sum_{i=1}^{n} \sqrt{\begin{cases} \left(d_i^{'min} - d_i^{max}\right)^2, d_i^{'min} > d_i^{max} \\ 0, others \\ \left(d_i^{min} - d_i^{'max}\right)^2, d_i^{min} > d_i^{'max} \end{cases}} \tag{16}$$

(5) This step is solves the eigenvalues and eigenvectors of the correlation matrix. It can be seen that $|Z - \lambda E| = 0$, E is a unit matrix, and the eigenvalue of the matrix Z is solved.

$$Z_{ij} = \frac{1}{n}\sum_{k=1}^{n} d_{ki} d_{kj} (i,j = 1, \cdots, m) \tag{17}$$

Unit features vector a_{ij} corresponding to the feature root Z_{ij} of the matrix Z is solved. a_{ij} is the factor load. It expresses as the degree of correlation between common factors and characteristic variables.

(6) This step is solves calculation and derive the cumulative contribution rate of the factor [30]. It can assume that the cumulative contribution rate is μ%. (In statistics, it is generally considered that the cumulative contribution rate is 85%).

ISSN: 2414-1895 DOI: 10.6919/ICJE.202006_6(6).0021

$$\frac{1}{m}\sum_{i=1}^{n-1}Z_{ij} < \mu \le \frac{1}{m}\sum_{i=1}^{n}Z_{ij} \quad (18)$$

(7) If the common factor has practical significance, a time series model is established. It is judged whether the comprehensive finding of the common factor has practical significance. The time-series data common factor is output, and a time-series model is established for prediction.

Through the factor analysis in the time series, the original data is dimension-reduced. In the process of constructing the distance matrix by solving the similarity of variables, the modified DTW distance lower bound function LB_H is used to measure the symmetry. It can describe the algorithm using code in Table1.

Table 1. Dynamic Time Warping Similarity Reduction Method Based on Factor Analysis

Input: time series training set $D = (d_{nm})^T$; Constraint value is s; $\text{diff} = \begin{cases} d_i'^{min} - d_i^{max} \\ 0 \\ d_i^{min} - d_i'^{max} \end{cases}$ Output: $LB\_H(D, D')$ and ARIMA
1: input s and def D 2: for D in "max-min" 3: endfor 4: if $D_{KMO} \ge 0.7$ then 5: cov(Z)=1, $Z = [Z_1, Z_2, \cdots, Z_m]$ 6: endif 7: $DTW_{LB}(D, D')$ 8: for each $D_{ij} > 0$ then 9: $LB\_H(D, D')=\sum_{i=1}^{n}\sqrt{diff_i^2}$; $LB\_H(D', D)=\sum_{i=1}^{n}\sqrt{diff_i'^2}$ 10: def $LB_H(D, D') = LB\_H(D', D)$ 11: $DTW_{LB_H}(D, D') \to [Z_{ij}]$ 12: if $\frac{1}{m}\sum_{i=1}^{n-1}Z_{ij} < \mu \le \frac{1}{m}\sum_{i=1}^{n}Z_{ij}$ and $\mu \ge 85\%$ 13: output $[Z_{ij}]$ 14: output ARIMA 15: endif 16: end

4. Experiments And Results

This chapter will use the actual data set to verify the feasibility and effectiveness of proposed method. The city's 18-year air quality data simulation will be used. The method is compared with the DTW distance measurement method combining information entropy in the literature [18] and the matrix-based DTW regularization method in the literature [19]. The quantitative and qualitative analysis results will be finally indicated. Experimental results is determines by quantitative and qualitative analysis.

4.1 Model Evaluation Method

The comparison of the experimental results will use the root mean square error (RMSE), average absolute error (MAE) algorithm running time (s) and prediction accuracy of the machine learning model as the evaluation index. The root means square error is defined as following:

ISSN: 2414-1895 DOI: 10.6919/ICJE.202006_6(6).0021

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}\left(X_{obs,i} - X_{model,i}\right)^2}{n}} \tag{19}$$

RMSE refers to the square of the deviation between the predicted value and the true value. The smaller the value, the higher the accuracy of the model. Similarly, the MAE reacts to the true case of the error between the predicted values. The smaller the value, the better the model. Where X_{model} is expressed as a predicted value and X_{obs} is represented as a true value. The mathematical formula can be expressed as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|\left(X_{model,i} - X_{obs,i}\right)\right| \tag{20}$$

4.2 Algorithm Performance Comparisons

This set of experiments is designed to verify the effectiveness and superiority of the proposed method. It is compared with the current representative algorithm LB_Keogh based on DTW distance metric, and combined with information entropy in [18] and grid line distance metric in literature [19].

The experimental data set of this group selects the air quality data of a city. It contains hourly and daily air quality data recorded for 2001-2018. In this data source, the timing length of each type of air quality particle type is the same, so that 80% is used as training set data (Query time series set) and 20% is used as detection set data (Candidate time series set). In the training process, the ten-fold intersection method is used for verification.

First, the experimental data are standardized and reprocessed. The multivariate time series model requires data sequence integrity and KMO test. The original data satisfies the normal distribution. The test results are shown in Table 2.

Table 2. KMO and Bartlett's test results.

Kaiser-Meyer-Olkin Measure of Sampling Adequacy		0.871
Bartlett`s Test of Sphericity	sig	0.001
	P-Value	0.001

In view of the above test results, the annual cycle is defined as a time variable. The obtained data sequence is subjected to distance metric measurement using dynamic time warping under factor analysis, and the common factor of the cumulative contribution rate $\mu \geq 85\%$ in the obtained factor is selected as input, and the time series stationary test is performed, because the time series data are not stable. After the differential conversion, the four sets of sequence P-value rejection null hypotheses can be obtained. The numerical results are shown in Table 3. The ARIMA model is then constructed by the common factor of the output to predict the air quality.

Table 3. Sequence stationary test results.

$P < \alpha(0.05)$ Reject the null hypothesis	
Type	P-value
LB_Keogh	0.017
Information entropy	0.034
Grid representation	0.038
LB_H	0.021

The results of the ADF smoothing test are shown in the above table. Since the experiment selects the air quality time series data for each time from 01 to 18, the maximum lags=30 is assumed. In order to verify and compare the performance of the algorithm, the experimental group uses lags=30 to conduct experiments when constructing ARIMA models with different similarity methods. The

ISSN: 2414-1895 DOI: 10.6919/ICJE.202006_6(6).0021

ARIMA model is constructed by LB_Keogh and the methods of [18] and [19] and the qualitative experimental results of this method are shown in Figure 4-7.

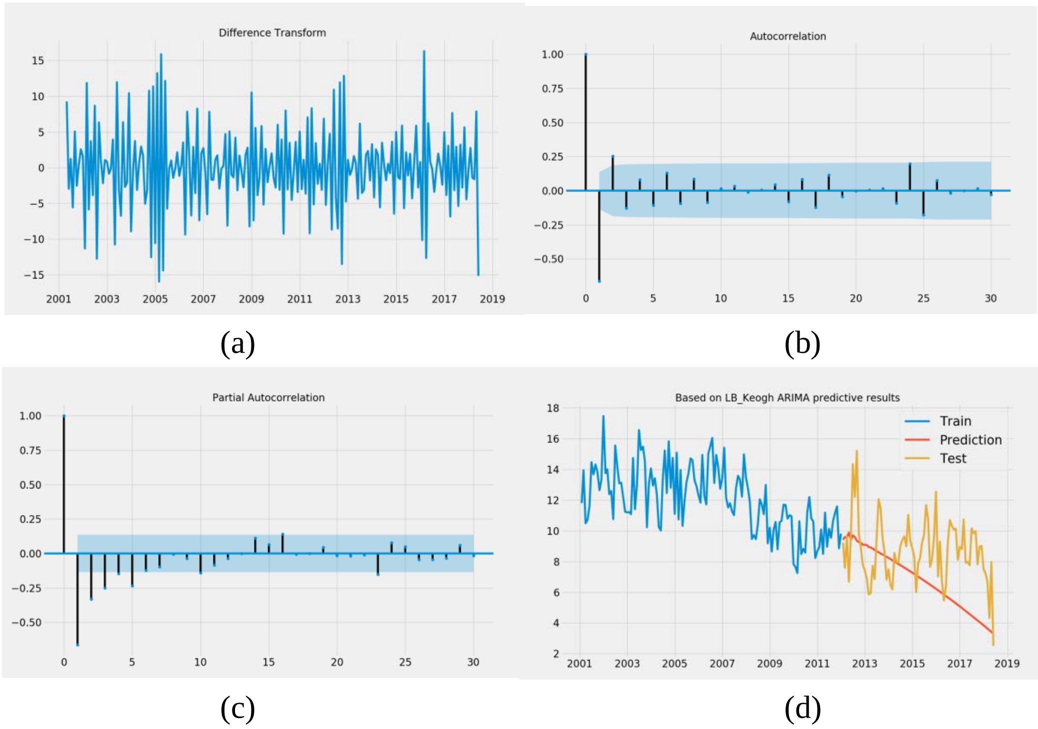

Figure 4. ARIMA model with LB_Keogh. (a) Stationarity test. (b) ACF of time series. (c) PACF of time series. (d) Predictive model results.

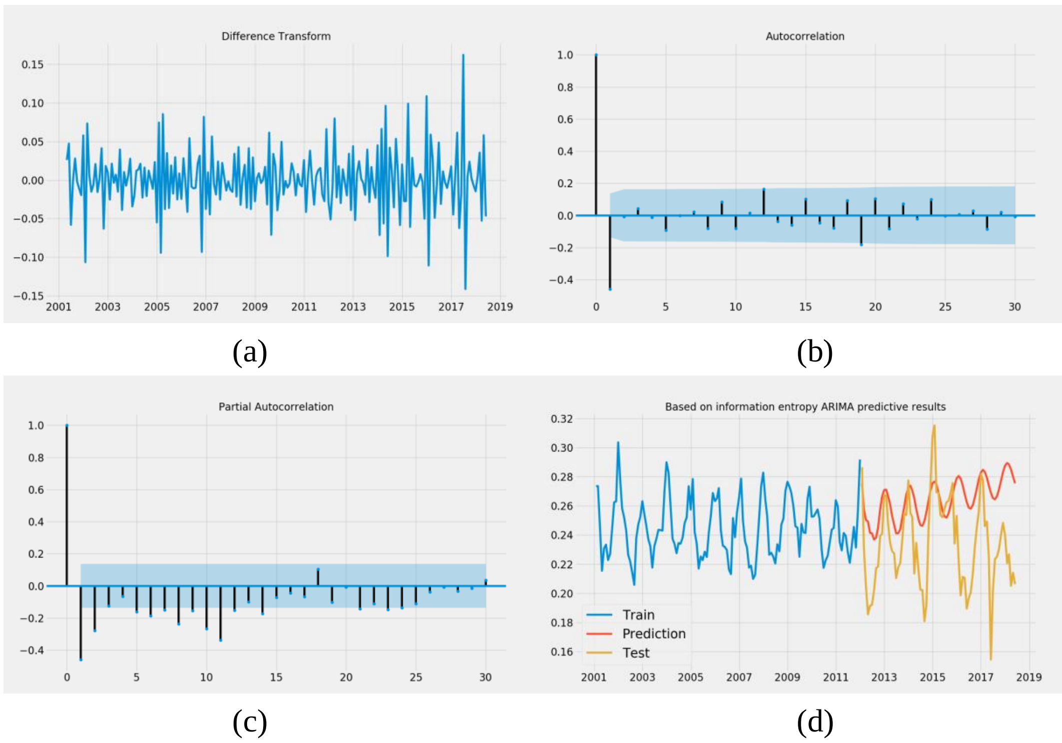

Figure 5. ARIMA model with information entrop. (a) Stationarity test. (b) ACF of time series. (c) PACF of time series. (d) Predictive model results.

ISSN: 2414-1895
DOI: 10.6919/ICJE.202006_6(6).0021

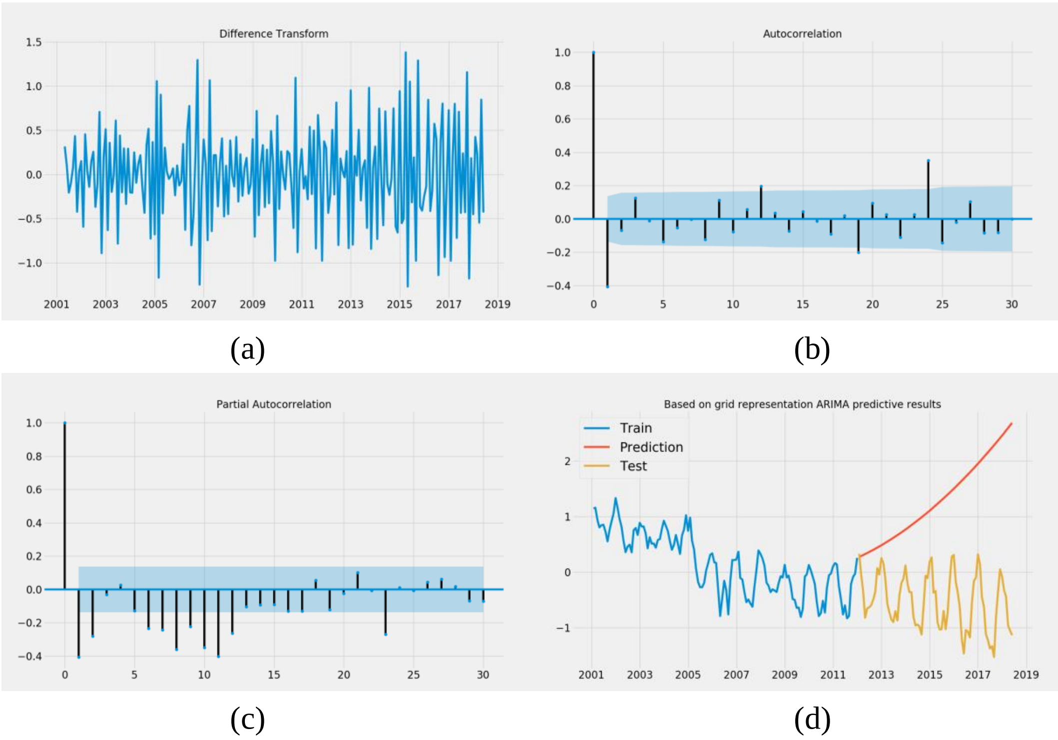

Figure 6. ARIMA model with grid representation. (a) Stationarity test. (b) ACF of time series. (c) PACF of time series. (d) Predictive model results.

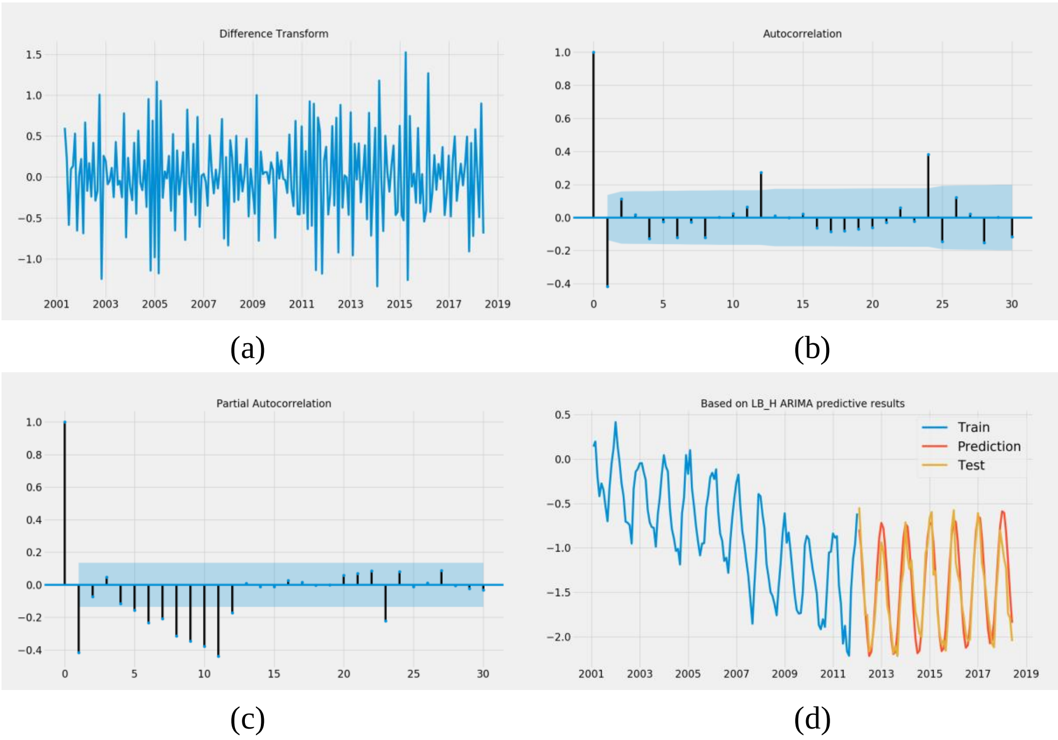

Figure 7. ARIMA model with LB_H. (a) Stationarity test. (b) ACF of time series. (c) PACF of time series. (d) Predictive model results.

As be shown in Figure 4-b above, it can be seen that the data ADF stationarity test is obtained from the correlation coefficient with a large correlation in the order of 1-5. The maximum order of the PACF partial correlation coefficient is 5. In summary, Similarly, the maximum order of the model constructed by the information entropy of Figure 5-b/c is 2.Figure 6-b/c can get the maximum order of grid lines is 2. In this paper, the maximum order of the model construction model smoothing transformation is 1.

The model is constructed by the smoothed processed time series data. The qualitative results of the training set and the test set are shown in Figure 4-7-(d). It can be seen that the proposed algorithm perform optimally. The results of quantitative analysis under different methods can also be clearly obtained from Table 4. The method obtained in this paper has achieved good results. Under the constraint of similarity measure symmetry, it is obvious that the computational complexity of LB_Keogh is about twice that of the proposed method, which proves the correctness of mathematical theory.

Table 4. Comparison of experimental under different methods.

Type	RMSE	MAE	ACC	Times
LB_Keogh	0.416	0.337	0.891	56.67
Information entropy	2.058	1.821	0.852	45.24
Grid representation	3.078	2.548	0.730	44.19
LB_H	0.219	0.173	0.956	31.32

5. Conclusion

In this paper, the dynamic time bending method distance metric is used to calculate the defect with high time complexity. And time series model predictions are inaccurate and improved. Extensive experimentation, this paper proposes a dynamic time warping similarity dimension reduction method based on factor analysis. In the multi-dimensional time series data metrics, the dimension reduction is based on factor analysis. Then, it is derives the DTW lower bound function distance metric method with symmetry LB_H. While solving the problem of distance measurement complexity, the LB_Keogh is improved to not satisfy the symmetry defect. The DTW lower bound function with symmetry is derived for the multivariate time series data distance metric and the common factor is generated, and then the model is constructed. The RMSE=0.219 and MAE=0.173 are obtained, and the accuracy and algorithm computation time complexity are significantly improved. The in-depth study of the next step is based on the multivariate time series similarity dimension reduction method with rough sets.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China (NSFC) under grant no. 51474095, in part by Henan Provincial Key Science and Technology Research Projects under grant no.152102210277, and CERNET Innovation Project under grant no.NGII20180313/NGII20160517,and in part by the Program for Science & Technology Innovative Research Team in the University of Henan Province under grant no.17IRTSTHN010, and in part by the Progarm for Science & Technology Innovative Research Team In University of Henan Province under grant no.2015XTD011, and the Henan University of Science and Technology Major Research, Education and Production Cooperation Cultivation Fund under grant no.2015ZDCXY03.

References

[1] K. Yang, M. Pan, Y. Luo, K. Chen, Y. Zhao, and X. Zhou, “A time-series analysis of urbanization-induced impervious surface area extent in the dianchi lake watershed from 1988–2017,” International journal of remote sensing, vol. 40(2019), p. 573–592.

[2] H. Feng, H. Wang, C. Yin, and X. Lu, “The set sfmbdr of fuzzy time series forecasting models,” in 2017 2nd International Conference on Modelling, Simulation and Applied Mathematics (MSAM2017). Atlantis Press, 2017.

[3] X. Guan, C. Huang, G. Liu, X. Meng, and Q. Liu, “Mapping rice cropping systems in vietnam using an ndvi-based time-series similarity measurement based on dtw distance,” Remote Sensing, vol. 8, no. 1, p. 19, 2016.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202006_6(6).0021

[4] J. Galbally and D. Galbally, “A pattern recognition approach based on dtw for automatic transient identification in nuclear power plants,” Annals of Nuclear Energy, vol. 81, p. 287–300, 2015.

[5] S. Oehmcke, O. Zielinski, and O. Kramer, “knn ensembles with penalized dtw for multivariate time series imputation,” in 2016 International Joint Conference on Neural Networks (IJCNN). IEEE, 2016, p. 2774–2781.

[6] H. Chen and X. Gao, “Time series similarity measure and cluster analysis based on wave characteristics,” Statistics and Decision, vol. 35, no. 11, p. 17–22, 2019.

[7] T. Janyalikit, P. Sathianwiriyakhun, H. Sivaraks, and C. A. Ratanamahatana, “An enhanced support vector machine for faster time series classification,” in Asian conference on intelligent information and database systems. Springer, 2016, p. 616–625.

[8] X. He, C. Shao, and Y. Xiong, “Improving time series classification accuracy: combining global and local information in the similarity criterion,” Latin American applied research, vol. 44, no. 3, p. 225–229, 2014.

[9] Y. Li, H. Hu, Y. Wen, and J. Zhang, “Can we speculate running application with server power consumption trace?” IEEE transactions on cybernetics, vol. 48, no. 5, p. 1500–1512, 2017.

[10] J. B. G. Yanqing Ye, Kewei Yang and Y. Dou, “Multivariate time series similarity matching method based on weighted dynamic time warping,” Pattern Recognition And Artificial Intelligence, vol. 30, no. 04, p. 314–327, 2017.

[11] F. Höppner, “Improving time series similarity measures by integrating preprocessing steps,” Data mining and knowledge discovery, vol. 31, no. 3, p. 851–878, 2017.

[12] B. G. Costa, J. C. A. Freire, H. S. Cavalcante, M. Homci, A. R. Castro, R. Viegas, B. S. Meiguins, and J. M. Morais, “Fault classification on transmission lines using knn-dtw,” in International Conference on Computational Science and Its Applications. Springer, 2017, p. 174–187.

[13] Y. Chen, X. Liu, X. Li, X. Liu, Y. Yao, G. Hu, X. Xu, and F. Pei, “Delineating urban functional areas with building-level social media data: A dynamic time warping (dtw) distance based k-medoids method,” Landscape and Urban Planning, vol. 160, p. 48–60, 2017.

[14] X. G. X. Z. Yubo Feng, Chengjun Ding and Q. Liu, “Time series similarity based on moving average and piecewise linear regression,” Computer Science, vol. 45, no. S1, p. 110–113, 2018.

[15] F. Zhang, A. Shan, and Y. Luan, “A novel method to accurately calculate statistical significance of local similarity analysis for high-throughput time series,” Statistical applications in genetics and molecular biology, vol. 17, no. 6, 2018.

[16] I. Boulnemour and B. Boucheham, “Qp-dtw: Upgrading dynamic time warping to handle quasi periodic time series alignment.” Journal of Information Processing Systems, vol. 14, no. 4, p. 851–876, 2018.

[17] R. Zhang and Y. Liu, “Multi feature small sample object recognition method based on dtw algorithm,” in 2018 IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC). IEEE, 2018, p. 2118–2122.

[18] Y. L. Meiying Qiao and H. Tao, “A multidimensional time series similarity measurement algorithm based on information entropy and dtw,” Journal of Sun Yat-sen University (Natural Science Edition), vol. 58, no. 02, p. 1–8, 2019.

[19] Y. Ye, J. Jiang, B. Ge, Y. Dou, and K. Yang, “Similarity measures for time series data classification using grid representation and matrix distance,” Knowledge and Information Systems, vol. 60, no. 2, p. 1105–1134, 2019.

[20] N. E. Watson, S. E. Prebihalo, and R. E. Synovec, “Targeted analyte deconvolution and identification by four-way parallel factor analysis using three-dimensional gas chromatography with mass spectrometry data,” Analytica chimica acta, vol. 983, p. 67–75, 2017.

[21] J. Garcia-Manrique, D. Camas, and A. Gonzalez-Herrera, “Study of the stress intensity factor analysis through thickness: methodological aspects,” Fatigue & Fracture of Engineering Materials & Structures, vol. 40, no. 8, p. 1295–1308, 2017.

[22] D. Pandey and K. K. Singh, “Implementation of dtw algorithm for voice recognition using vhdl,” in 2017 International Conference on Inventive Systems and Control (ICISC). IEEE, 2017,pp. 1–4.

[23] A. R. Anwary, H. Yu, and M. Vassallo, “Gait evaluation using procrustes and euclidean distance matrix analysis,” IEEE journal of biomedical and health informatics, 2018.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202006_6(6).0021

[24] R. Tavenard and L. Amsaleg, "Improving the efficiency of traditional dtw accelerators," Knowledge and Information Systems, vol. 42, no. 1, pp. 215–243, 2015.

[25] J. Wang, C. Liu, L. Ding, H. Luo, and B. Song, "Multi-stage real-time identification for data stream events with drift feature based on dtw," IEEE Access, vol. 7, pp. 89 188–89 204, 2019.

[26] T. Janyalikit, P. Sathianwiriyakhun, H. Sivaraks, and C. A. Ratanamahatana, "An enhanced support vector machine for faster time series classification," in Asian conference on intelligent information and database systems. Springer, 2016, pp. 616–625.

[27] C. S. Fuchs, K. Muro, J. Tomasek, E. Van Cutsem, J. Y. Cho, S.-C. Oh, H. Safran, G. Bodoky, I. Chau, Y. Shimada et al., "Prognostic factor analysis of overall survival in gastric cancer from two phase iii studies of second-line ramucirumab (regard and rainbow) using pooled patient data," Journal of gastric cancer, vol. 17, no. 2, p. 132–144, 2017.

[28] J. Wolfert, C. Verdouw, and A. Beulens, "Future challenges for information integration in multi-dimensional agri-food supply chain networks," in Proceedings of the 4th International Conference on Information and Communication Technologies in bio and Earth Sciences (HAICTA 2008), 18-20 September 2008,Athens, Greece, 2008, p. 196–203.

[29] Y. Nozaki and T. Nakamoto, "Itakura-saito distance based autoencoder for dimensionality reduction of mass spectra," Chemometrics and Intelligent Laboratory Systems, vol. 167, p. 63–68, 2017.

[30] D. Garcia, S. MacDonald, and M. Rapp-Ricciardi, "Factor analysis of the s wedish version of the d ark t riad d irty dozen," PsyCh journal, vol. 6, no. 2, p. 166–167, 2017.