Multivariate time series prediction of high dimensional data based on deep reinforcement learning

Xin Ji[1], Haifeng Zhang[1], Jianfang Li[1], Xiaolong Zhao[1], Shouchao Li[2] and Rundong Chen[2*]

[1] Big Data Center of State Grid Corporation of China, Beijing 100052, China
[2]Beijing Sgitg Accenture Information Technology Center Co., Ltd, Beijing, 100052 China

Abstract. In order to improve the prediction accuracy of high-dimensional data time series, a high-dimensional data multivariate time series prediction method based on deep reinforcement learning is proposed. The deep reinforcement learning method is used to solve the time delay of each variable and mine the data characteristics. According to the principle of maximum conditional entropy, the embedding dimension of the phase space is expanded, and a multivariate time series model of high-dimensional data is constructed. Thus, the conversion of reconstructed coordinates from low-dimensional to high-dimensional can be kept relatively stable. The strong independence and low redundancy of the final reconstructed phase space construct an effective model input vector for multivariate time series forecasting. Numerical experiments of classical multivariable chaotic time series show that the method proposed in this paper has better forecasting effect, which shows the forecasting effectiveness of this method.

1Introduction

The evolution of any component of complex high-dimensional data multivariate time series can be determined by other components determined by high-dimensional data multivariate time series. Therefore, the development process of any component of high-dimensional data multivariate time series contains the information influence of other components[1-2]. Data reconstruction is carried out through the component development characteristics of high-dimensional data multivariate time series, so as to fully reflect the original characteristics. Because of the complexity of structure and dynamic system, it is difficult to reconstruct the phase space from one-dimensional time series by delay embedding method. Therefore, it can not guarantee that any given single variable time series in the actual system is enough to reconstruct the original system Unification[2]. The multivariable time series contains more abundant and complete system information, so it can reconstruct more accurate data vector space. Therefore, the research of multivariable chaotic time series has received more and more attention. Time series can explore the law of its development and change to predict some phenomena[3]. With the deepening of time series analysis, a multivariable time series prediction method based on deep reinforcement learning for high-dimensional data is proposed, and has achieved good application results, but most of the methods are proposed for single variable time series prediction. However, the time series data collected in real life are often not determined by a single factor, and need to consider a variety of factors[4]. Therefore, the research on multivariate time series prediction has more practical significance. Because multivariate time series data have different characteristics from univariate time series: multi noise, multi-scale, variable correlation and so on, the existing univariate time series prediction methods can not directly predict multivariate time series, so the research on multivariate time series prediction has important theoretical value.

2Multivariate time series prediction of high dimensional data

2.1multivariate time series association algorithm for high dimensional data

High dimensional data is a feedforward learning data processing model, and its structure is very similar to radial basis function. Compared with the RBF data processing model, the generalized high-dimensional data has more advantages in approximation ability and convergence speed[5-6]. The multivariable time series pre prediction method uses the correlation between multiple time series to improve the overall prediction accuracy. The key of forecasting multivariable time series is how to accurately capture the complex time series pattern and the dependence between variables.

*E-mail: chenrundong@sgitg.sgcc.com.cn

$$E(Y|\ X)=\frac{\int_{-\infty}^{+\infty} Yf(Y,X)dX}{\int_{-\infty}^{+\infty} f(Y,X)dY} \tag{1}$$

Where X is an n-dimensional input variable, $X=(X_1, X_2, \ldots, X_n)$; Y is a k-dimensional output variable, $Y=(Y_1, Y_2, \ldots, Y_n)$,F (x, y) is the joint probability density function of X and Y, and $E(Y|\ X)$ is the expected value of the output variable Y given the input variable X. For univariate time series $x(n)(n=1, 2, \ldots, N)$After determining the embedding dimension m and time delay τ, the phase space is reconstructed

$$\mathbf{X}(n)=[x(n),\cdots,x(n-(m-1)\tau)]^{\mathrm{T}}\in\mathbf{R}^m\ \ (n=N,\cdots,(m-1)\tau+1) \tag{2}$$

Let X_1, X_2, …, X_K, where k is the number of observed variables, $X_i=[x_i(1), x_i(2), \ldots, x_i]$ denotes the time series of the ith variable. If each variable chooses the appropriate time delay τ and embedding dimension m_i (i = 1, 2…n). Among them, τ i and mi are the key to the success or failure of phase space reconstruction[7-8]. Through the above formula, the regression estimation is obtained, and the predicted time series are obtained. According to the above system functional requirements, the C / S three-tier architecture is adopted.

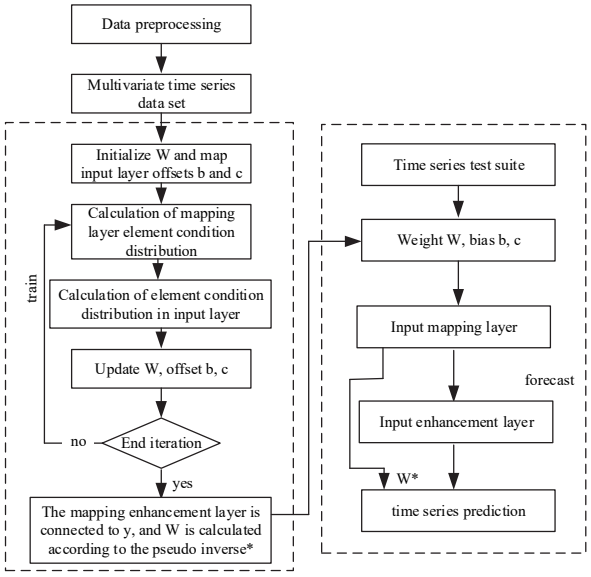

Fig. 1 Processing flow of multivariate time series of high dimensional data

The significance test of the model mainly tests the validity of the model to see whether the model has fully and effectively extracted all the information, that is, to ensure that the residual sequence is the significance test of white noise model parameters, is to test whether each parameter in the model is significantly different from zero, and its purpose is to make the model more concise and accurate.

2.2Realization of multi pass time series prediction based on high dimensional data

From the perspective of mathematical statistics, time series can be divided into long-term trends, cyclical changes and irregular items. Among them, the long-term trend is a relatively stable part of the data for time changes. It is often a stable rise or fall, and it is a predictable part. The dominant cyclical fluctuation of the data depends on the position of the data in the cycle Historical data location predicts the periodicity of a future location. Cyclical change is a kind of invisible periodic performance, most of which are hidden in the trend part, so it is usually not predicted separately. Irregular term usually refers to the randomness of data, which is an unpredictable part. Therefore, when predicting time series, time series is usually divided into three parts: trend (T), data (S) and remainder (R). The common decomposition modes are addition mode and multiplication mode.

$$X_t=T_t+S_t+R_t \tag{3}$$

Among them, the classical time series decomposition method is widely used, but there are some problems[9-10]. The specific implementation process is as follows:

Step 1: establish data volume prediction model. Through the known data time series, the data volume at a certain time point in the future can be predicted, and the formula is as follows:

hypothesis $Y=\{y_1,y_2,...,y_{t-1},y_t,y_{t+1},...\}$ is a time series in a fixed time period, the prediction model is

$$\hat{y}_{t+1}=\sum k(1-k)Y+(1-k)s \tag{4}$$

In the formula, k is the smoothing parameter and s is the initial value of dynamic smoothing;

Step 2: calculate the deviation between the predicted data and the actual data. The formula is as follows:

$$w_t=|Y_t-U_t| \tag{5}$$

In the formula, W is the deviation at time t, Y is the predicted data at time t, and U is the actual data at time t.

Step 3: judge whether the deviation exceeds the set threshold range. If it exceeds the threshold, it is considered that there is an abnormal phenomenon in the data volume. Otherwise, it is considered normal and no abnormal condition occurs. The mathematical formula is described as follows:

$$\begin{cases} w_t\in[n,m], & \text{abnormal} \\ w_t\notin[n,m], & \text{normal} \end{cases} \tag{6}$$

In the formula, $[n,m]$ according to the Takens embedding theory, if the embedding dimension and delay time are selected reasonably, the phase space can be reconstructed and the prediction effect is ideal. The method is used to forecast the short-term load based on the data series of short-term load influencing factors and load time series collected by an actual power grid.

3Analysis of experimental results

The experiment is implemented in the experimental environment of win7 system and 2G memory with MATLAB 2016. In the experiment, the original data is standardized and the records with missing values are deleted. Then the time series is divided into the first 2 / 3 and the last 1 / 3, and the K-10 sliding window is used to generate the training set and the test set respectively. In order to verify the practical application effect of the model, comparative experiments are carried out. Through OPNET modeler14. 5 network simulation software, establish information transmission system model. Ppleca and HARQ error correction algorithm are used to send information data to D1. The experimental environment settings are shown in table 1.

Tab. 1 Experimental environment setting

parameter	parameter
Host memory	4GB
CPU frequency	3。50GHz
operating system	64 bit
Program running platform	VisualStudio

The network connection bandwidth is 10Mbps and the bandwidth frequency is 4Mbps. In order to improve the simulation speed, C language is used in OPNET Network System. The experimental information is h. 264 standard "akiyo" sequence and QCIF file format. Using 25frame / s, 10s and 25s as the encoding / decoding rate, jm18. 4 as the encoder / decoder of information sequence, the simulation time is set at 1800s. Multivariable chaotic time series are time series with chaotic characteristics generated by chaotic model. Because of the complexity of chaotic system, the long-term prediction effect is not ideal, but its deterministic structure determines the possibility of short-term prediction of the system. Therefore, we predict Rossler, Chen's and Lorentz chaotic time series respectively. The generation model of chaotic time series is as follows: Rossler chaotic time series.

$$\begin{cases} \dfrac{dx}{dt} = -y - z \\ \dfrac{dy}{dt} = x + sy \\ \dfrac{dz}{dt} = r + z(x - b) \end{cases} \tag{7}$$

In order to verify the effectiveness of the algorithm, the latent extreme learning machine proposed in this chapter is compared with the traditional and weighted extreme learning machine. It is the distance between the sample to be tested and the training sample, which gives the training sample the advantages and disadvantages of the algorithm.

$$RMSE_j = \sqrt{\sum_{i=1}^{N} \left(y_{ij} - \hat{y}_{ij} \right)^2 / N} \tag{8}$$

$$MAE_j = \frac{1}{N} \sum_{i=1}^{N} \left| y_{ij} - \hat{y}_{ij} \right| \tag{9}$$

Where N is the root of the j-th variable, which is the actual value of the j-th variable at the i-th time point of the number of test samples, and Y is the predicted value of the j-th variable at the i-th time point. The packet loss rate data in the collection period is used as the initial training sample set. See the table for the specific data.

Tab. 2 Initial training sample set

number	time delay (s)	Simulation speed (s)
1	1.02	0.012
2	0.56	0.062
3	0.33	0.016
4	0.18	0.035
5	0.65	0.042

Based on the information in the above table, the peak signal-to-noise ratio and packet loss rate of the receiver information sequence are calculated by comparing the traditional prediction model with the time series analysis method, and recorded and analyzed. The specific detection results are shown in the following figure:

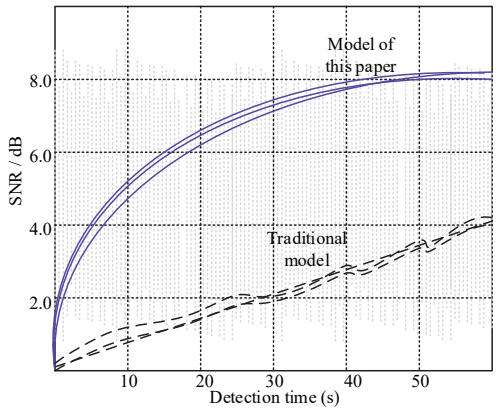

Fig. 2 SNR detection of time series prediction data

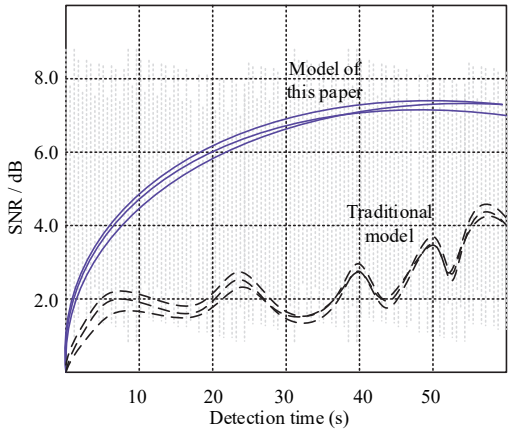

Fig. 3 Packet loss rate detection results of time series prediction data

Based on the analysis of the detection results, compared with the traditional prediction model, the proposed mathematical detection model of network

packet loss rate based on time series analysis method has higher signal-to-noise ratio and lower packet loss rate in the actual application process, which confirms the practical application process of the mathematical detection model of network packet loss rate based on time series prediction method. The results show that the effect is better, which can fully meet the research requirements.

4Conclusions

Time series is a common form of data expression in real life. It has a wide range of applications. Various mathematical models or algorithms are used to mine the internal characteristics of time series, so as to further guide the production and life. In today's era of big data and cloud computing, time series is also showing more and more high-dimensional, long data, high dimensional correlation and information Redundancy and so on. Therefore, it is of great significance to study how to predict multivariable time series through new and efficient analysis and calculation methods. A multivariable time series prediction algorithm based on improved generalized high-dimensional data is proposed. The generalized high-dimensional data algorithm is used to predict multivariable time series, which has less adjusting parameters and saves computing resources. Then, combined with particle swarm optimization algorithm, find the appropriate smoothing factor more accurately. The experimental results show that the prediction accuracy of the algorithm is significantly improved compared with the commonly used mainstream models.

References

1. Btthpa, Cmpca, Cal, et al. (2020) Dynamic time warping-based imputation for univariate time series data - ScienceDirect[J]. Pattern Recognition Letters, 139(10):139-147.
2. Gundersen K, Alendal G, Oleynik A, et al. (2020) Binary Time Series Classification with Bayesian Convolutional Neural Networks When Monitoring for Marine Gas Discharges[J]. Algorithms, 13(6):145.
3. Ambrosino F, L. Thinová, Briestensk M, et al. (2020) Detecting time series anomalies using hybrid methods applied to Radon signals recorded in caves for possible correlation with earthquakes[J]. Acta Geodaetica et Geophysica, 55(3):405-420.
4. Melin P, Daniela Sánchez, Monica J C, et al. (2021) Optimization using the firefly algorithm of ensemble neural networks with type-2 fuzzy integration for COVID-19 time series prediction[J]. Soft Computing, (10):1-38.
5. Churchill R M, Tobias B, Zhu Y. (2020) Deep convolutional neural networks for multi-scale time-series classification and application to tokamak disruption prediction using raw, high temporal resolution diagnostic data[J]. Physics of Plasmas, 27(6):062510.
6. Zhang D, Chen S, Ling L, et al. (2020) Forecasting agricultural commodity prices using model selection framework with time series features and forecast horizons[J]. IEEE Access, PP(99):1-1.
7. Laureniu, Opincariu, Norbert, et al. (2019) Edge computing in space: Field programmable gate array-based solutions for spectral and probabilistic analysis of time series.[J]. The Review of scientific instruments, 90(11):114501-114501.
8. Churchill R M, Tobias B, Zhu Y. (2020) Deep convolutional neural networks for multi-scale time-series classification and application to tokamak disruption prediction using raw, high temporal resolution diagnostic data[J]. Physics of Plasmas, 27(6):062510.
9. Tan W, Xing J, Yang S, et al. (2020) Long Term Aquatic Vegetation Dynamics in Longgan Lake Using Landsat Time Series and Their Responses to Water Level Fluctuation[J]. Water, 12(8):2178.
10. Juan José Vidal Macua, José Manuel Nicolau, Vicente E, et al. (2020) Assessing vegetation recovery in reclaimed opencast mines of the Teruel coalfield (Spain) using Landsat time series and boosted regression trees[J]. Science of The Total Environment, 717(10):137250.