

A Survey on Pre-processing Methods for Time Series Data

Liyu Zhu*

School of Computing, Guangdong University of Science & Technology, Dongguan, China

Abstract

This paper reviews the latest development of time series data analysis methods. Firstly, the characteristics of high-dimensional, high-redundancy, autocorrelation and concealment of time series data are introduced. Then, according to the existing problems of time series data, the corresponding pre-processing methods such as data dimensionality reduction, data augmentation and data denoising are given respectively. After that, combined with the instantiation of time series data in different fields, the application of time series data in anomaly detection, prediction, diagnosis and other practical tasks is discussed. Finally, the future research trends are summarized and prospected.

Keywords

Time series data; Dimensionality reduction; Data augmentation; Denoising.

1. Introduction

With the rapid development of Internet of Things technology and application, there are a lot of time series data in the real world, such as traffic data [1], medical data [2], meteorological data [3], and earthquake data. Time series is a set of numerical sequences in which the observed variables of the system are arranged in time order, which can describe the dynamic characteristics of complex systems [4]. Time series data have temporal dependence in the time dimension, through which the evolution law of complex systems can be revealed.

Scholars try to discover the dynamic evolution law of complex systems from various angles of time series data and apply the evolution law to practical tasks such as data anomaly detection and prediction. Pan et al. [5] put forward the research on time series data, pointing out that statistical research mainly focuses on the representation, measurement, and prediction of accurate or discrete data; Machine learning research involves pattern discovery and prediction of discrete sequences; The research of database involves the storage and indexing of multidimensional data and the similarity query of large data sets. Therefore, analyzing time series data is an important research topic concerned by academic circles.

In the era of big data, time series data sources are extensive, and the accumulation time is extended, which leads to data redundancy, data missing, and noise interference in the dataset, which seriously affects the quality of time series data. Therefore, it is necessary to improve the data quality to a certain extent through effective means to prepare for the subsequent mining, analysis and modeling of time series data.

This paper first introduces the characteristics of time series data, then gives the corresponding pre-processing methods for the problems existing in time series data, and finally introduces the application of time series data in different fields.

2. Characteristics of Time Teries Data

In the era of big data, time series data mainly has the characteristics of massiveness, high-dimensional, high-redundancy, autocorrelation and concealment.

2.1. Massiveness

With the rapid development of Internet of Things technology and sensors, various types of sensors collect and transmit massive amounts of data in the fields of engineering quality control, monitoring and security. Time series data come from a wide range of sources and accumulate for a long time, which makes the whole world tend to be digitized.

2.2. High-dimensional and High-redundancy

Most of the data detected in the real world comes from complex systems, and the data presents high-dimensional and high-redundancy. High-dimensional refers to the collection of data that contains multiple different attributes and different positions in a complex system [6]. This makes it difficult for many classical data mining algorithms to function and increases the difficulty of research on mining algorithms. How to reduce redundant features and extract main features is the primary goal of time series data analysis.

2.3. Autocorrelation and Concealment

In the time dimension, the same attribute has a high degree of autocorrelation, which is also the most common data characteristic of time series data. There are often some hidden features in multivariate time series data, which are not obvious, so they are easy to be ignored in data analysis. However, if the hidden features in the time series data are abused or ignored, the subsequent model establishment will be affected and the model performance will be negatively affected. Hidden features are still an indispensable part of the representation of things, so it is a great challenge to discover hidden features and their evolution in massive time series data.

3. Pre-processing Method

At present, the main problems of time series data include data redundancy, data missing and noise interference, which seriously affect the data quality of time series. Data preprocessing improves data quality to a certain extent by practical means and prepares for the subsequent analysis, mining, and modeling of time series data. In view of the above three main problems, we summarize three pre-processing methods, which are data dimensionality reduction, data augmentation, and data denoising.

3.1. Data Dimensionality Reduction

The traditional method of time series modeling focuses on the future trend of a variable in the system. When modeling multivariate time series, scholars pay more attention to the evolution law of the whole complex system. Due to the high-dimensional and high-redundancy of the data, the dimension of time series data is reduced, and essential features are extracted mainly utilizing feature selection and feature extraction.

feature selection [7] refers to eliminating irrelevant and redundant variables in the original high-dimensional input dataset, retaining related variables, and screening out an optimal subset containing relevant information among variables to replace the original high-dimensional input feature set. Guyon I et al. [8] divided variable selection into four steps: subset generation, subset evaluation, termination condition judgment, and subset verification. At present, it mainly focuses on the research of subset evaluation, which can be divided into filtering, encapsulation, and embedded methods according to the criteria of subset evaluation.

Feature extraction refers to mapping or transformation to transform the original high-dimensional data into the low-dimensional feature space to achieve the purpose of dimension reduction. In the data of heterogeneous sensors, many scholars apply principal component analysis [9], convolutional neural network [10], residual shrinkage network [11], and other methods to high-dimensional time series.

3.2. Data Augmentation

Many time series tasks do not have enough tag data in a complex system. Data augmentation is considered a method of constructing more samples based on a small number of samples to expand the amount of data [12]. Data augmentation is an effective way to improve the scale and quality of training data. The basic idea of data augmentation is to generate a synthetic dataset with new input space while maintaining the correct label.

There is temporal dependency between the series data samples, so we should pay special attention to the temporal dependency between the newly generated samples when enhancing the series data. There are many types of data, such as time series data with continuous values, text data with discrete values, audio data, and so on. Although the data types are different, the applied data augmentation methods have common design ideas. Traditional methods commonly include geometric transformation, window clipping or window slicing, adding noise, transformation in the frequency domain, and anomaly label extension [13]. Data augmentation methods based on deep learning include generative adversarial networks (GAN), GAN based on convolutional neural networks, and GAN based on recurrent neural network [14].

3.3. Data Denoising

In a complex system, due to the influence of observation errors, sensor failures, data storage and transmission failures, and extreme events of the system, it is inevitable that some noise will be involved in the collected time series data. Therefore, different methods can be used to deal with noise according to the type of noise, which can help the subsequent system modeling. When noisy time series data is input into the model, it may reduce the accuracy of system modeling and even lead to model failure. Therefore, we should choose different methods to deal with noise according to the type of noise and provide help for the subsequent system modeling. Usually, the filtering method is used to denoise time series data, divided into time-domain filtering methods and frequency-domain filtering methods [15]. The time-domain filtering method usually uses the sliding window as the filter and replaces it with the statistical value in the sliding window to separate the signal and noise data. Time-domain filtering methods commonly include mean filtering, median filtering, and adaptive filtering. The frequency-domain filtering method transforms the signal data from the time domain to the frequency domain and distinguishes between high and low-frequency signals by setting appropriate thresholds. Low-frequency signals belong to valid data, and high-frequency signals belong to noise data. Traditional frequency domain filtering methods mainly include Kalman, Wiener, and low-pass filtering [16].

4. Application

Time series data analysis is widely used in intelligent transportation, medical treatment, weather detection, and other fields. We should fully consider the characteristics of different time series data and make the corresponding pre-processing to provide higher quality data for practical tasks in different fields.

Intelligent transportation system is thriving. Traffic flow is one of the foremost parameters to reflect the road running state. Traffic flow data is highly nonlinear and complex time series data. If the characteristics of traffic flow data can be fully considered and preprocessed, more accurate input data can be provided for traffic flow prediction [17], traffic event prediction [18], and traffic congestion prediction. Grooming vehicles in advance can improve the operational capacity and efficiency of the road network and reduce traffic accidents, thus reducing people's travel time and cost and environmental pollution.

In the medical field, patients' physical condition can be continuously detected by various technologies such as functional magnetic resonance imaging, electroencephalogram, and

functional near infrared spectroscopy. For such complex time series data, if we can fully consider the characteristics of medical data and construct data models of different disease types, we can better help solve many problems in the medical field. For example, classification based on brain activation [19], classification based on brain function network and other diseases classification and diagnosis tasks [20].

5. Conclusion

This paper introduces the characteristics of time series data and its common problems. Aiming at the problems existing in time series data, corresponding preprocessing methods such as data dimensionality reduction, data augmentation and data denoising are given. We also discussed the application of time series data in different fields. With the deepening of the research on complex systems, we think that incremental and high-order mining can be deeply studied to discover the hidden rules in time series data and their dynamic evolution rules. In addition, aiming at multivariate time series data, we can deeply explore the correlation between multivariate and the future evolution trend of the whole system to prepare for the decision-making and control of complex systems.

References

- [1] Zhang Y, Wang S, Chen B, et al. TrafficGAN: network-scale deep traffic prediction with generative adversarial nets[J]. IEEE Transactions on Intelligent Transportation Systems, 2019, 22: 219-230.
- [2] Kawahara J, Brown C J, Miller S P, et al. BrainNetCNN: convolutional neural networks for brain networks; towards predicting neurodevelopment[J]. NeuroImage, 2017, 146: 1038-1049.
- [3] Cheng W, Shen Y, Zhu Y. A neural attention model for urban air quality inference: learning the weights of monitoring stations[C]. Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI, 2018: 1-12.
- [4] Esling P, Agon C. Time-series data mining[J]. ACM Computing Surveys, 2012, 45(1): 12.
- [5] Pan D, Shen JY. Similarity discovery techniques in temporal data mining. Journal of Software, 2007, 18(2): 246-258.
- [6] Min H, Weijie R, et al. Survey of Chaotic Time Series Analysis and Prediction[J]. Information and Control, 2020, 49(1): 24-35.
- [7] Liu H, Yu L. Toward integrating feature selection algorithms for classification and clustering[J]. IEEE Transactions on Knowledge and Data Engineering, 2005, (4): 491-502.
- [8] Guyon I, Elisseeff A. An introduction to variable and feature selection[J]. Journal of Machine Learning Research, 2003, 3: 1157-1182.
- [9] Yi S Y, Lai Z H, He Z Y, et al. Joint sparse principal component analysis[J]. Pattern Recognition, 2017, 61: 524-536.
- [10] Huang Z, Zhu J, Lei J, et al. Tool Wear Monitoring with Vibration Signals Based on Short-Time Fourier Transform and Deep Convolutional Neural Network in Milling[J]. Mathematical Problems in Engineering, 2021(6):1-14.
- [11] Minghang Zhao, Shisheng Zhong, Xuyun Fu, et al. Deep residual shrinkage networks for fault diagnosis[J]. IEEE Transactions on Industrial Informatics, 2020, 16(7): 4681-4690.
- [12] Yizhou Ge, Xiang Xu, et al. Survey on Sequence Data Augmentation[J]. Journal of Frontiers of Computer Science and Technology, 2021, 15(7): 1207-1213.
- [13] Gao J, Song X, Wen Q, et al. RobustTAD: Robust Time Series Anomaly Detection via Decomposition and Convolutional Neural Networks[J]. 2020.
- [14] Esteban C, Hyland S L, Rtsch G. Real-valued (Medical) Time Series Generation with Recurrent Conditional GANs[J]. arXiv, 2017.
- [15] Liu Bo, Wang Mingshuo, et al. Deep Learning for Spatio-Temporal Sequence Forecasting: A Survey[J]. Journal of Beijing University of Technology, 2021, 47(8): 925-941.

- [16] Yasodha, Polisetti, Srinivasulu, et al. Wavelet transform-based methods for removal of ground clutter and denoising the radar wind profiler data[J]. IET signal processing, 2015, 9(5): 440-448.
- [17] Zhang Y, Wang S, Chen B, et al. TrafficGAN: network-scale deep traffic prediction with generative adversarial nets[J]. IEEE Transactions on Intelligent Transportation Systems, 2019, 22: 219-230.
- [18] Ren H, Song Y, Wang J, et al. A deep learning approach to the citywide traffic accident risk prediction[C]. 2018 21st International Conference on Intelligent Transportation Systems (ITSC) . Piscataway: IEEE, 2018: 3346-3351.
- [19] Jang H, Plis S M, Calhoun V D, et al. Task- specific feature extraction and classification of fMRI volumes using a deep neural network initialized with a deep belief network: evaluation using sensorimotor tasks[J]. NeuroImage, 2017, 145: 314-328.
- [20] Kawahar A J, Brown C J, Miller S P, et al. BrainNetCNN: convolutional neural networks for brain networks; towards predicting neurodevelopment[J]. NeuroImage, 2017, 146: 1038-1049.