Advances in Computer, Signals and Systems (2023)
Clausius Scientific Press, Canada

DOI: 10.23977/acss.2023.070816
ISSN 2371-8838 Vol. 7 Num. 8

Simulation of Multidimensional Time Series Data Analysis Model Based on Deep Learning

Lixia Liu

College of Information Engineering, Engineering University of PAP, Xi'an, Shaanxi, 710086, China

Keywords: Deep Learning; Time Series; ResNet

Abstract: By analyzing time series, we can realize functions such as prediction and detection to save manpower and material resources. However, time series data are usually accompanied by noise and data loss, which greatly restricts our use and analysis of time series data. In this paper, the current situation of time series classification research is comprehensively analyzed, and a multi-dimensional time series data analysis model based on deep learning is proposed. The feature extraction part of the model consists of a hollow convolution space pyramid structure and two residual blocks, and the residual blocks follow the structure of ResNet classification model. The pyramid structure of empty convolution space can be used as a basic module structure and a part of other types of neural network structures to obtain rich feature information, or it can be simply stacked many times and used as an independent network structure. Experimental results show that the proposed model has similar and good classification performance. Compared with other algorithms, the end-to-end deep learning algorithm designed in this paper has greatly improved the accuracy and solved the problem of the accuracy of multi-dimensional time series classification.

1. Introduction

Time series is a series of observed data values collected in time sequence. According to the number of observed variables, it can be divided into multivariate (single) time series, and according to the number of observed time steps, it can be divided into equal-length (variable-length) time series. In recent years, with the rapid development of sensor equipment and the continuous improvement of data storage capacity, a large number of high-dimensional time series data have been generated and saved in many fields such as biology, medicine, finance and public transportation[1]. The analysis and mining of time series data can be divided into two types: prediction and classification. For example, the classification of ECG series data will help doctors diagnose patients' heart diseases, while for continuous traffic volume and numerical weather data, regression prediction will be more conducive to the travel planning of urban travelers and the management decision of urban managers. Therefore, the problem of time series classification and prediction has penetrated into all aspects of production and life, and it has high research value in academic and industrial scenarios.

In the past few decades, the technology related to machine learning has made great progress, which has aroused the interest of researchers in different fields, and its influence has been involved in many industrial fields, such as autonomous driving, medical care, finance, manufacturing, energy and many other fields[2-3]. Usually, these processes are carried out automatically through the algorithms themselves, so these algorithms can be regarded as automatic machines and belong to the category of artificial intelligence. At present, people usually prefer to use deep learning algorithms to solve more complex problems that traditional machine learning is difficult to deal with.

Time series data usually have the complex characteristics of high dimension, mass, real-time, high noise and multiple outliers[4-6]. This makes it difficult for us to directly obtain valuable potential information in the series, and it is even more difficult to dig out the accurate commonness and difference characteristics between time series. In this paper, the current situation of time series classification research is comprehensively analyzed, and a multi-dimensional time series data analysis model based on deep learning is proposed.

2. Research method

2.1. Model structure

The multi-dimensional time series in this paper has the characteristics of non-linearity and non-stationarity, that is, we can't guarantee that it is produced by the linear superposition of several attributes or the values of previous moments, nor can we guarantee that its statistical characteristics will not change with time. Because of the different physical meanings in reality, the numerical range and changing trend of data in each dimension are also different. There may also be correlation between the data of each dimension[7]. By analyzing time series, we can realize functions such as prediction and detection to save manpower and material resources. However, time series data are usually accompanied by noise and data loss, which greatly restricts our use and analysis of time series data.

In traditional time series prediction and repair methods, whether based on autoregressive models and their derivative models, or models based on changes in data rate, they cannot extract the relationships between different dimensions of data at the same time, and deep learning methods have more complex models, Structures such as CNN (Convolutional Neural Network) and RNN (Recurrent Neural Network) are also more conducive to extracting relationships between data of different dimensions. With the development of hardware such as GPU, deep learning is applied more and more in all aspects, and new network structures are constantly proposed. Deep learning has also become an important means to solve time series problems.

With the important breakthrough and wide application of deep learning in many fields, the end-to-end deep learning model is gradually concerned and studied by the time series community [8-9]. Based on the end-to-end time series classification model ResNet with good performance, this paper makes an improvement study. Figure 1 is the improved ResNet classification model with a pyramid structure of empty convolution space.

Hole convolution enlarges the receptive field without reducing the image resolution [10]. The feature extraction part of the model consists of a hollow convolution space pyramid structure and two residual blocks, and the residual blocks follow the structure of ResNet classification model. The pyramid structure of empty convolution space can be used as a basic module structure and a part of other types of neural network structures to obtain rich feature information, or it can be simply stacked many times and used as an independent network structure.

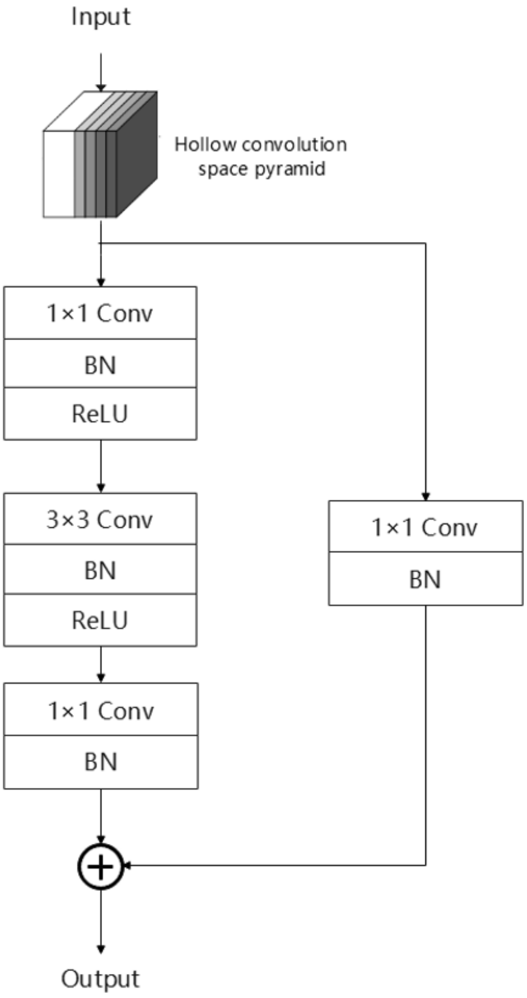

Figure 1: Multi-dimensional time series data analysis model

2.2. Overall description of algorithm

Inter-time series prediction can predict the change of time series in the future by mining the useful information of time series data in the past, and can be applied to retail, cloud computing and energy. Through the investigation of the research situation of time series prediction, we find that although the prediction method based on statistics has a theoretical basis and strong explanatory power, the prediction accuracy is easy to fall into the bottleneck and the performance of the model is poor because of too harsh assumptions on the model. Time series classification refers to inputting time series data as a single sample into the classifier, and the classifier outputs its label category. Traditional time series classification methods are mainly distance-based and feature-based methods, and most of them can't handle irregularly sampled time series data well.

The receptive field is very important for understanding and evaluating the depth of CNN [11-12], and it is often used in the field of computer vision. For time series, receptive field can be regarded as the theoretical value of the maximum visual field of neural network in one-dimensional space. Generally speaking, the larger the receptive field, the better the network will perform in detecting longer sequence patterns.

Suppose that the depth of a network is d, and each layer of it has a one-dimensional convolution kernel with the length $k_i, i \in [l, d]$ and the sliding step l, and its receptive field is defined as:

$$R_i = 1 + \sum_{i=1}^{d} (k_i - 1) \tag{1}$$

It can be seen from the above formula that the receptive field size of the network can be changed by changing the number of layers d of the network and the size k_i of the convolution kernel.

The strategy of choosing the timing of incremental learning can affect the frequency of incremental learning, that is, the update frequency of the model. Reasonable timing can save the calculation cost of model updating and make the model run accurately on new data as long as possible. In this paper, KL divergence is selected as a measure of the difference between the old and new distribution of multi-dimensional time series, so as to select the opportunity to update the model

KL divergence describes the difference between two probability distributions P, Q about the variable x, and the calculation formula is shown in Formula (2):

$$D(p\|q) = \sum_{i=1}^{n} p(x_i) \ln \frac{p(x_i)}{q(x_i)} \tag{2}$$

Where x_i is a different value of x. When x is a continuous variable, the distribution of x is generally obtained by the method of equal interval boxing statistics. Equal interval box counting means that according to the maximum and minimum range Δ of data distribution P, according to a certain interval division number n, Δ is divided into n intervals, that is, n boxes. Then count the number $c_1, \cdots, c_n$ of data in P in each interval of n intervals.

Let the total number of data in P be $|P| = c_1 + \cdots + c_n$, then the continuous distribution P can be regarded as discrete distribution, and the probability distribution of P can be estimated:

$$p(x_i) = \frac{c_i}{|P|} \tag{3}$$

The residual block structure used in this paper consists of three basic convolution structures, BN and ReLU, in which the number of convolution kernels in each convolution operation is equal, and the size of convolution kernels is set to $\{k_1, k_2, k_3\}$ respectively.

It is directly connected to the output of three-layer convolution through the simple addition of shortcut, so as to solve the degradation problem of deep network and improve the training speed of the model. Moreover, when the dimensions of input data and output data are not equal, the convolution kernel with length 1 is used to convert the input into dimensions equal to the output, and then the summation operation is carried out.

For the generation network, we take the mean square error as the loss function, assuming that the prediction result of the generation network for the missing data at time t is $\tilde{x}_t$, then the loss is:

$$l_G(x_t, x_t) = \sum_{d=1}^{D} (1 - m_t^m)(x_t^d - \tilde{x}_t^d)^2 \tag{4}$$

3. Simulation analysis

In order to verify the performance of the model, the experimental data in this paper are taken from the SIS real-time/historical database of a power plant. There are five categories of samples, representing different states of equipment, and the number of samples in each category is relatively balanced. In the experiment, according to the time sequence of sample collection, 800,000 samples were taken as the initial model training set, and the remaining 300,000 samples were taken as the new distribution data. In the experiment, the neural network model is realized by Tensorflow framework, and the model realized under this framework can be accelerated by graphics card GPU.

The proposed model has been randomly experimented on data sets for three times, and the average accuracy of each data set is the final result. The overall comparison results between the proposed model and the two algorithms, Gan (Generative Adversarial Networks) and FCN (Fully Convective Networks), are shown in Figure 2.

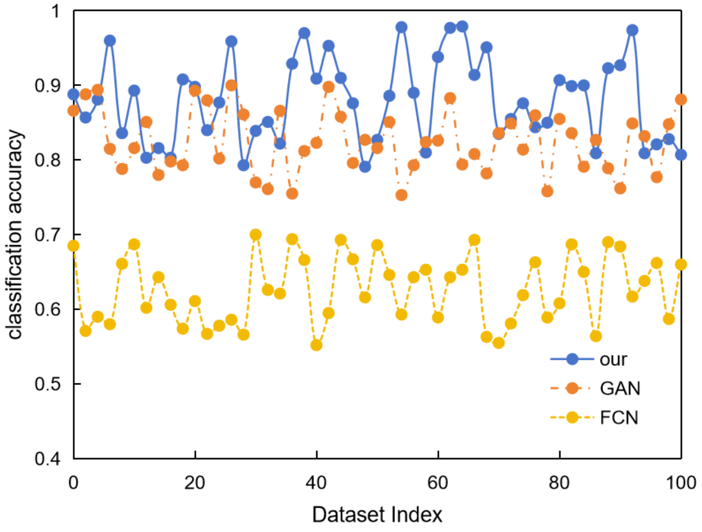

Figure 2: Comparison of classification accuracy of each algorithm

It can be seen that for most data sets, the classification accuracy of the three algorithms is concentrated around 0.8, and only a small number of data sets will have a classification accuracy lower than 0.5, which shows that all five models can achieve good classification results on the whole. It can be concluded that the proposed model has similar and good classification performance with GAN, while the FCN model is not stable but has good classification performance.

4. Conclusions

Time series is a series of observed data values collected in time sequence. According to the number of observed variables, it can be divided into multivariate (single) time series, and according to the number of observed time steps, it can be divided into equal-length (variable-length) time series. In this paper, the current situation of time series classification research is comprehensively analyzed, and a multi-dimensional time series data analysis model based on deep learning is proposed. The experimental results show that the proposed model has similar and good classification performance, and the end-to-end deep learning algorithm designed in this paper has greatly improved the accuracy compared with other algorithms, which solves the problem of the accuracy of multi-dimensional time series classification. However, in terms of efficiency, it can not reach the level required by industry.

References

[1] Aurélie Davranche, Lefebvre, G., & Poulin, B. (2010). Wetland monitoring using classification trees and spot-5 seasonal time series. Remote Sens Enveron, 2010, 114(3), 552-562.

[2] Kim, M. (2013). Semi-supervised learning of hidden conditional random fields for time-series classification. Neurocomputing, 119(7), 339-349.
[3] Zhang, Z., Cheng, J., Li, J., Bian, W., & Tao, D. (2012). Segment-based features for time series classification. Computer Journal, 55(9), 1088-1102.
[4] Yuhei, U. (2017). Time series classification via topological data analysis. Transactions of the Japanese Society for Artificial Intelligence, 32(3), 12.
[5] Bernardi, M. L., Cimitile, M., Martinelli, F., & Mercaldo, F. (2018). Driver and path detection through time-series classification. Journal of Advanced Transportation, 2018(3), 1-20.
[6] Manisha, K., & Arti, K. (2022). Development of adaptive time-weighted dynamic time warping for time series vegetation classification using satellite images in solapur district. The Computer Journal (8), 8.
[7] Huang, W., Yue, B., Chi, Q., & Liang, J. (2019). Integrating data-driven segmentation, local feature extraction and fisher kernel encoding to improve time series classification. Neural processing letters, 49(1), 43-66.
[8] Wang, S., Hua, G., Hao, G., & Xie, C. (2017). A cycle deep belief network model for multivariate time series classification. Mathematical Problems in Engineering, 2017(9), 1-7.
[9] Zhai, Y., & Qu, Z. (2018). Crop classification based on nonlinear dimensionality reduction using time series remote sensing images. Nongye Gongcheng Xuebao/Transactions of the Chinese Society of Agricultural Engineering, 34(19), 177-183.
[10] Fuqun, Z., & Aining, Z. (2016). Optimal subset selection of time-series modis images and sample data transfer with random forests for supervised classification modelling. Sensors, 16(11), 1783.
[11] Lhermitte, S., Verbesselt, J., Verstraeten, W. W., & Coppin, P. (2011). A comparison of time series similarity measures for classification and change detection of ecosystem dynamics. Remote Sensing of Environment, 115(12), 3129-3152.
[12] Deng, H., Runger, G., Tuv, E., & Vladimir, M. (2013). A time series forest for classification and feature extraction. Information Sciences, 239(4), 142-153.