ISSN 1746-7659, England, UK
Journal of Information and Computing Science
Vol. 15, No. 2, 2020, *pp*.146-153

Research on user behavior recognition based on 2D-CNN

Lei Zhao
School of Mathematics and Statistics, Nanjing University of Information Science & Technology, Nanjing 210044, China
(Received July 10, 2020, accepted August 12, 2020)

Abstract: In this paper, we studied the application of two-dimensional convolutional neural networks to the classification of multivariate time series. Time series sample data is usually a set of measurement values of a single attribute or multiple attributes at continuous time points separated by uniform time intervals. It is a set of structured data, usually non-discrete, time-related between data Features such as sex, feature space, and large dimension. At present, most methods for time series classification problems need to go through an extremely complex data preprocessing process and related feature engineering and do not consider the long pattern information hidden in different time dimensions of time series data, and the different characteristics of multivariate time series data Relevant information in the space dimension between. By converting multivariate time series data into matrix form, this paper proposes an end-to-end deep learning model Pyramid-CNN based on two-dimensional convolutional neural networks, which uses two-dimensional convolution kernels to extract the spatial dimensions of multivariate time series data and the relevant information in the time dimension, and applied it to the user behavior recognition time series data set. The experimental results show that for this data set, compared with the existing methods, the model proposed in this paper has higher performance Accuracy and robustness, with a good classification effect.

Keywords: time series classification, deep learning, convolutional neural network

1 Introduction

Today's society has entered the information age of big data. With the rapid development of information technology, data has exploded in various fields and industries, such as stocks[1-3], currencies, precious metals, futures, and other trading quotations and buyers in the commercial field. And seller information feedback, etc.; in the field of science and technology robot detection records, aerospace information, mechanical control, etc[4][5].; in the field of medical information records[6], medical imaging records, disease monitoring, etc.; in the field of social media network chat content Records, digital images, video, and audio, etc. Every day, billions of searches are performed on the Internet, and hundreds of millions of megabytes of information are transmitted. It can be said that data is everywhere, and mankind has officially entered the era of big data. When faced with a huge amount of information, how to choose the available information for use has become an issue of widespread concern, and data mining research is extremely important.

Time-series mining has always been a hot issue in the field of data mining. In recent years, data mining summit KDD and neural information processing summit NIPS have held special seminars on time series related issues every year. Time series data refers to a kind of structured data formed by recording sample attribute values in chronological order. It often has the characteristics of high dimensionality, a large amount of data, noise, and different sample lengths. Time series analysis is to analyze the data change process and future trends based on historical record values. With the advent of the era of big data and the development of computer hardware technology, a large amount of diverse time series data has been generated in the fields of aviation, finance, medical treatment, and industrial production. For example, in motion recognition, the angle changes of the human skeleton joints are obtained as the characteristics of motion. The model measures the similarity of the change sequence for action recognition; in the financial market, it uses time series clustering to study the financial investment portfolio. Further exploration of the hidden information in the original time series will help discover knowledge, laws, and patterns. The problem of time series classification is one of its important research directions. It has obvious application value and theoretical guiding significance. It has been studied and explored for many years.

Time-series mining has always been a hot issue in the field of data mining. In recent years, data mining has topped the list of issues. Multivariate time series classification has been widely used and more complex than univariate time series data. It has always been a research focus on timing issues. This article mainly studies the problem of predicting user behavior based on the time-series data of the user's mobile phone sensor. This has a strong application background in preventing financial fraud and anti-fraud.

Without loss of generality,a labeled data set with N samples $X = \{(x_i, y_i)\}_{i=1}^{N} \in R^{N \times l}$. among them, (x_i, y_i) represents a sample pair, sample $x_i = \{x_i^1, x_i^2, \ldots, x_i^l\}$has l observations, for multivariate time series data, x_i^j is a $p \times l$ vector, p is the number of sample variables, y_l is its corresponding label, there are usually C possible category values. The time series classification problem is to learn a mathematical model that can predict the corresponding label y for a given new input sample x.

Traditional time series methods are mainly distance-based methods, such as DTW[6], ED[7] (Euler distance) feature-based methods, such as decision trees[8], Bayes[9], etc., and feature-based integrated learning methods, such as RF[10], XGBoost[11], LightGBM[12], etc. . However, traditional methods not only have higher data requirements, such as distance-based methods, the length of the sample must not be too short, and when the amount of data is large, the prediction time is longer and cannot be applied to real-time scenarios. Moreover, the requirements for engineers are also high. For example, for feature-based methods, it is necessary to extract features artificially. This requires engineers to have an understanding of specific business scenarios and the subject knowledge required by the business. This requires the knowledge of engineers. It must have a certain breadth and a certain depth, which requires a higher quality of engineers.

Compared with the current mainstream time series algorithms, deep learning models do not require heavy and complex data preprocessing processes and feature engineering, and the accuracy of classification has reached an advanced level. The time series algorithms based on deep learning have attracted the interest of researchers. In this paper, two-dimensional convolutional neural networks are applied to multivariate time series classification problems, and the application of two-dimensional convolution kernels enables the network model to extract local information and cross-combination information between different features. Sequence features have a new perspective, which improves the generalization ability of the model. This provides a new idea for the future application of deep learning technology to solve time series problems, and the development of a new deep learning model framework for time series related problems. It has a promoting effect.

2 Related work

In the research of time series classification, distance-based methods, feature-based methods, model-based methods, integrated learning-based methods, and deep learning-based methods are mainly used.

In the method based on distance measurement, the discriminative distance measurement plays a vital role in the performance of the classification model. First, define the distance function to calculate the similarity between two time-series, and then classify the sequence instances into the corresponding class according to the class of each time series instance and the closest instance in the training data. Commonly used distance measurement methods include Euclidean distance, dynamic time warping distance, longest common subsequence[13], and edit distance[14]. The earliest algorithm based on distance measurement is the method based on Euclidean distance. This method mainly measures the similarity between different time series samples by calculating the Euclidean distance between them, and then uses the KNN method to predict the label. But there is a problem with Euclidean distance. It requires the same length between different samples. In actual scenarios, the lengths between different time series samples are often different. For the case of the unequal length of time series, edit distance and dynamic time warping distance is usually used to solve the problem. There are two forms of edit distance: one is to measure the distance based on the number of conversions used to convert one time series to another; the other is to measure the distance based on the length of the longest common subsequence in two time-series data. Certainly, the edit distance is suitable for dealing with locally disturbing sequences, but not suitable for dealing with samples with severe phase distortion. The nearest neighbor algorithm based on DTW distance has been widely recognized and has been successfully applied to tasks such as time series classification and clustering.

The feature-based classification method usually includes two steps: defining time-series features and then training a classifier based on the defined time-series features for classification. The time series forest [15] algorithm calculates the mean, variance, and slope of the random sequence fragments, takes these statistical features as the characteristics of a fragment and then uses the random forest algorithm to search in a huge