Application of Multivariable Time Series Classification and Clustering Algorithm

Yuqi Wang

Hangzhou Normal University, School of Mathematics Hangzhou, Zhejiang 311121, China.

Abstract: With the advent of the era of big data and the development of artificial intelligence technology, more and more fields need to use time series analysis. In order to solve the clustering problem of multivariate time series, a deep learning-based algorithm is adopted, which is optimized by deep neural network. The algorithm uses different types of networks to complete the work of feature extraction, feature selection and classifier design in the process of data classification, so as to realize the effective clustering analysis of multivariate time series. At present, multi-source data clustering class analysis based on deep learning has become a research hotspot, but there is no mature and stable theoretical framework and perfect and practical technical scheme. This paper proposes an innovative algorithm, a multi-variable time series clustering algorithm based on deep learning, to solve two major problems faced by existing models and algorithms.

Keywords: Multivariate Time Series Classification; Clustering Algorithm; Applied Research

Introduction

Clustering is a very important technology in data mining. By defining the distance measurement method, we can obtain the similarity between unclassified label samples, so as to obtain the clustering information of data. Cluster analysis has become the most active branch in data analysis and mining because it can effectively find the potential law in data distribution. The processing of multivariate time series data is a widely used task in many fields such as finance, medicine and audio or video recognition. The traditional density-based clustering algorithm has the disadvantages of heavy computation and easy to fall into local extreme points. With the vigorous development of information technology, the amount of multivariate time series data generated every day is huge, and the cost of manually labeling these data is too high, so the research of multivariate time series clustering algorithm has extremely important theoretical significance and practical value.

1. Overview of clustering algorithms for multivariate time series

At present, the existing time series clustering algorithms are mainly suitable for single variable time series, while the clustering algorithms for multivariate time series are relatively lacking. At present, there is not a unified and clear standard on the clustering of multivariate time series, and there is also a lack of effective classification of different types of data. Since the complexity of multivariate time series samples is higher than that of univariate time series, both dimensions of time step and variable need to be considered at the same time, which greatly increases the difficulty of algorithm design, so the research on this part of algorithms is relatively weak [1]. So far, these relatively few multivariable time series clustering efforts have been summarized into two broad classes of methods, each with unique characteristics and advantages:

1) Traditional clustering method: it is based on the similarity measurement of the original data, and clustering is carried out according to the measurement results;

2) Use deep learning technology: Use deep neural network to extract features from the original data, then measure the similarity of the extracted features, and finally perform cluster analysis according to the measurement results.

2. Disadvantages of deep learning multivariate time series clustering algorithm

2.1 ignored feature extraction between variables

When extracting the features of MTS, Encoder only considers the information extraction on the time step dimension, but ignores the feature extraction between variables. In addition, the dimensions of different MTS datasets vary greatly, so it is difficult to design a fixed network model to meet the requirements.

2.2 The downstream clustering task is not considered

In the process of learning representations, the downstream clustering task is not considered, and only simple dimensionality reduction is carried out, which fails to effectively constrain the compactness within the learned representations and the separation between classes. Therefore, clustering directly on such representations will inevitably have a negative impact on the performance of the algorithm.

2.3 Limitations of the algorithm

Although K-means algorithm has a wide range of applications in clustering, its direct use on the learned representation will still cause some limitations on the clustering performance.

3. Multivariate time series clustering based on deep learning

3.1 Algorithm Framework

The new Encoder algorithm can not only extract features from the two dimensions of time step and variable, but also flexibly adjust the network structure according to the dimension of the data set, so as to achieve more efficient feature extraction. In addition, the algorithm is applied to the classification method based on deep learning and compared with the existing similar work. Secondly, by jointly optimizing the reconstruction loss and triplet loss of Encoder, and improving the existing problems in triplet loss, the model can achieve interclass compactness and inter-class separation in low-dimensional space while containing as much original data information as possible, so as to better support the development of downstream clustering tasks [2]. The DCMTS network model is shown in Figure 1.

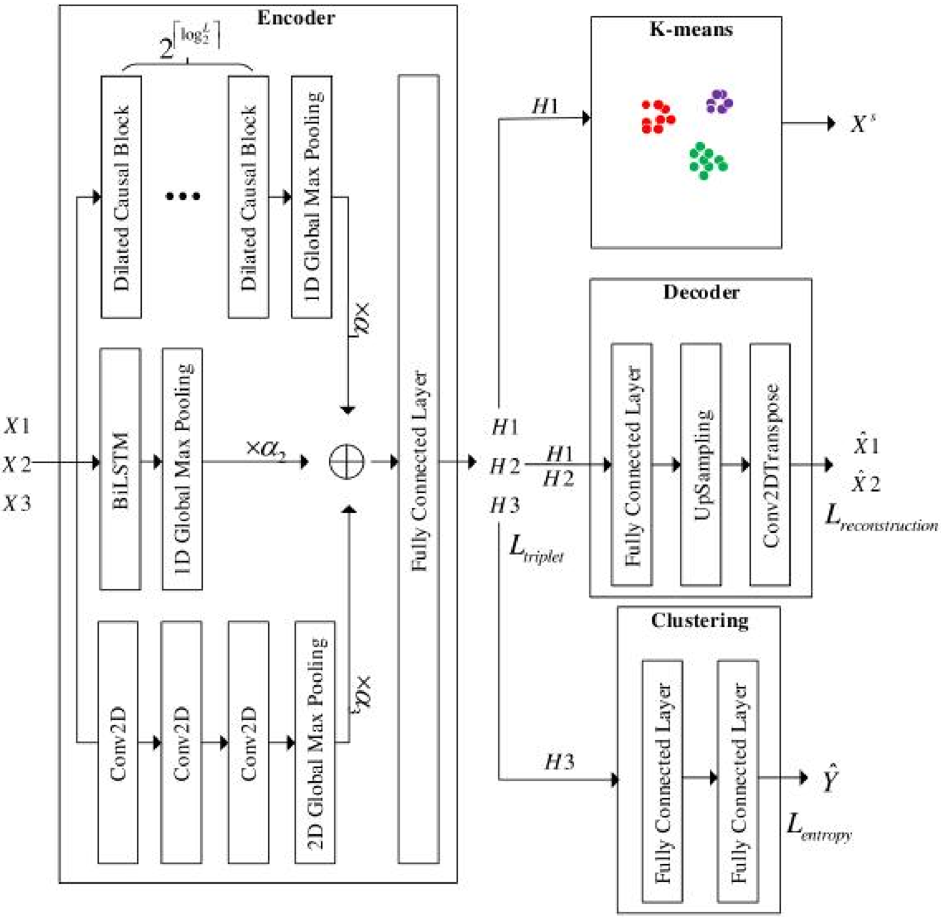

Figure 1. DCMTS algorithm model diagram

3.2 Model loss function

DCMTS algorithm adopts the joint optimization of reconstruction loss and triplet loss of Encoder to obtain a more suitable expression for clustering. This method can reduce the computation amount on the premise of ensuring the clustering effect. In order to obtain better clustering results, the optimization strategy of cross entropy loss is adopted to optimize the performance of the clustering network. The following are the definitions and functions of these three loss functions.

3.2.1 Reconstruction Loss

In order to maximize the information extraction capability of Encoder, an MTS sample is given. Firstly, a nonlinear map with parameter is defined, that is, Encoder network model x. Then, the low-dimensional representation is obtained by Encoder, and the calculation process is shown in formula (1).

$$h_i = f_{enc}\left(x_i, \theta_{enc}\right)$$

Next, define a mapping whose parameters are the Decoder network model. In this model, each node is a function of itself to propagate and receive information. The low dimensional representation obtained by Encoder is then transformed into the high dimensional space for decomposition and synthesis. Reconstructed representations available through Decoder. This function is used to project the original image into the high dimensional space. What is presented is a reconstructed expression whose operation process is shown in formula (2):

$$\hat{x}_i = f_{dec}\left(h_i, \theta_{dec}\right)$$

Finally, the reconstruction loss method Autoencoder was adopted to obtain the optimization. The specific calculation process is shown in formula (3) :

$$L_{reconstruction} = \frac{1}{N}\sum_{i=1}^{N}\left\|x_i - \hat{x}_i\right\|_2$$

In formula (3), the norm of 2 is gradually decreasing, and with the trend of gradually becoming more and more similar to and, more useful information is finally extracted.

3.2.2 Cross entropy loss

In order to optimize the performance of the whole clustering network, DCMTS algorithm adopts cross loss optimization strategy to improve the accuracy and reliability of clustering results. In this method, cross loss refers to using different number of training samples to learn the best weight vector for each class, so as to achieve the purpose of improving the classification accuracy. The specific calculation process is described as follows: After the completion of the first stage of training, the samples with obvious cluster structure and corresponding cluster labels are selected; The representation is processed by the Encoder. And defines a map, which is a model with parameter of, for clustering networks. This method can maintain the relationship information between each data object well and reduce the requirement of the original data set. Through the clustering network, the cluster labels can be obtained. The calculation of this process is shown in equation (4) :

$$\hat{y}_i = f_{cluster}\left(h_i, \theta_{cluster}\right)$$

Then, on the basis of Ys, cross entropy loss is used to optimize the clustering network. The calculation process of the clustering network is shown in equation (5) :

$$L_{entropy} = -\frac{1}{N_S}\sum_{i=1}^{N^s}\sum_{j=1}^{K}1\left\{y_{i,j} = 1\right\}\log\frac{\exp\left(\hat{y}_{i,j}\right)}{\sum_{j=1}^{K}\exp\left(\hat{y}_{i,j}\right)}$$

Formula (5) represents the total number of samples contained in the training set, represents the number of class families contained in the training set, represents the probability that the stored sample belongs to the class, and represents the probability that the predicted sample belongs to the class through the clustering network.

3.3 Actual application of the algorithm

The experimental results show that the algorithm can meet the practical requirements well, with high efficiency and good scalability. In recent years, artificial intelligence technology has achieved great success in solving some intractable problems in the field of life science, especially the rapid development of image recognition technology in recent years, which has shown excellent advantages in the field of medical imaging such as nuclear magnetic resonance, CT and ultrasound. With the continuous maturity of medical image processing and analysis technology, more and more doctors and scholars begin to pay attention to this field. In the field of medicine, the application of artificial intelligence technology has not only brought a significant improvement in the level of medical treatment, but also greatly alleviated the work pressure of medical personnel, however, the time series data involved in this field, such as brain wave data, is far beyond the scope of image data. As one of the important medical data types, electroencephalogram (EEG) contains a lot of information about physiological activities in human body and the characteristics of various biological rhythms. This law can be deeply understood to fill the gaps in relevant knowledge, so as to better benefit the society [3]. With the continuous improvement of science and technology, people are more and more aware that the human brain is also self-regulating, and brainwave signal as a non-contact stimulation mode, can accurately reflect the internal activities of the human body.

Conclusion

In summary, firstly, the existing problems of MTS clustering algorithm are summarized comprehensively, then the proposed DCMTS algorithm is introduced, and the overall framework of the algorithm is introduced in detail. Finally, the design strategies of each part of the algorithm are deeply discussed, and how these strategies can effectively solve the current problems. Finally, the practical application of the algorithm is described. This paper is mainly based on the distribution of large samples and high-dimensional data in real scenarios to build a large-scale classification model that can be applied to different scale and different types of data sets.

References

[1] Ren YY, Wang CJ. Time series classification of remote sensing images with abnormal data [J]. Journal of Computer Applications, 2019,41(3):662-668. (in Chinese)

[2] Wang KY, Du HD, Jia R, Liu H, Liang Y, Wang XY. Short-term interval probability prediction of photovoltaic power based on similar daily clustering and QR-CNN-BiLSTM model [J]. High Voltage Technology, 2019,48(11):4372-4384.

[3] Zhang X, Zhang L, Jin B, Zhang HZ. Research on multivariate time series classification algorithm based on uncertainty [J]. Acta Automatica Sinica, 2023, 49(4): 790-804.