SaPt-CNN-LSTM-AR-EA: Hybrid Ensemble Learning Framework for Time Series-based Multivariate DNA Sequence Forecasting

Yan Wu[1], Tan Li[2], Meng-shan Li[2], Sheng Sheng[1], Jun Wang[1], and Fu-an Wu[1]

[1]Jiangsu University of Science and Technology
[2]Gannan Normal University

November 2, 2022

Abstract

Biological sequence data mining is a long-term hot spot in bioinformation. Biological sequence can be regarded as a set of characters composed of a number of letters and contain an evolutionary relationship. Time series is a set of numbers arranged according to time and contains the temporal progressive relationship. Time series is similar to biological sequence in terms of both representation and mechanism. Therefore, in the paper, biological sequence is represented with time series to form biological time sequence (BTS). Based on advanced time series methods, hybrid ensemble learning framework (SaPt-CNN-LSTM-AR-EA) for BTS is proposed. Single-sequence and multi-sequence models are constructed based on self-adaption pre-training one-dimensional convolutional recurrent neural network (CNN-LSTM) and autoregressive fractional integrated moving average (ARFIMA) integrated evolutionary algorithm, respectively. In the DNA sequence experiments of six kinds of viruses, SaPt-CNN-LSTM-AR-EA realized the good overall prediction performance, the prediction accuracy and correlation were 1.7073 and 0.9186, respectively. The effectiveness and stability of SaPt-CNN-LSTM-AR-EA were verified through the comparison with other five benchmark models. In addition, compared with other benchmark models, SaPt-CNN-LSTM-AR-EA increased the average accuracy by about 30%. This study opened up a new field of BTS research. The framework proposed in this paper is significant in many disciplines, such as biology, biomedicine, computer science and economics. Especially in sequence splicing, genome, computational biology, bioinformation, theoretical biology, evolutionary biology, signal processing, medicine and health care and other fields, the framework has a wide range of applications.

Hosted file