STAT 6494 Data Science Project Proposal

Shanglin Zhou 2334229

1 Introduction

If any data is collected on base of time and make a series, then that series is called time series. This property makes time series data almost everywhere in our daily life, such as daily temperature we check on our phone everyday, or annual population of the United Stated that publish on the website of Bureau of Statistics every year. We always use time series data to study the past behavior, do forecasting, evaluate the achievement, and also do comparison.

One interesting topic of time series study is the classification of time series data. Example is like when we have a bench of time series datasets collected from the earthquake censor, and we want to identify what kind of features in these dataset indicate the occurrence of earthquake. Then we need to classify which series is recorded just before earthquake and which is recorded during normal time. In this project, I aim to build a different model to push the limit of time series classification limit based on a subset of UCR time series database[1].

2 Data

The datasets are open data from UEA & UCR Time Series Classification Repository [1], it is an ongoing project to develop a comprehensive repository for research into time series classification. I will choose twelve of the datasets which I randomly draw two dataset from each type, but covers all the types of time series data in the repository. Description of each dataset is introduced in the Table 1.

Table 1: Description of Time Series Datasets

Dataset	Train Size	Test Size	Length	No of Classes	Type
Adiac	390	391	176	37	IMAGE
Beef	30	30	470	5	SPECTRO
CinCECGtorso	40	1380	1639	4	ECG
Coffee	28	28	286	2	SPECTRO
CricketX	390	390	300	12	MOTION
Fish	175	175	463	7	IMAGE
ItalyPowerDemand	67	1029	24	2	SENSOR
NonInvasiveFatalECGThorax1	1800	1965	750	42	ECG
SyntheticControl	300	300	60	6	SIMULATED
TwoPatterns	1000	4000	128	4	SIMULATED
UWaveGestureLibraryX	896	3582	315	8	MOTION
Wafer	1000	6164	152	2	SENSOR

3 Methods

Until now, all the time series classification algorithms can be classified into three categories: distance based, feature based, and ensemble based. A milestone for the time series analysis study is the use of neural

[1] `http://timeseriesclassification.com/dataset.php`.

network. FCN (Fully Convolutional Network) [3] uses convolution features for classification and sets a new strong baseline for time series study.

Sometimes it is difficult to capture the specific feature on time domain, but when we convert it to spectrum, those kind of feature may become obvious. Because most of the time series data are non-linear, it is better to convert the time series data to bi-spectrum form[2] and salient the specific feature of different class of time series.

Then, I will try to fit a convolutional neural network (CNN) to the bi-spectrum form of time series data. The CNN is widely known to be good at doing image pattern identification, and hence image classification. Because bi-spectrum form of the time series data is essentially a 2-D image, then the time series classification problem is converted to the image classification problem.

References

[1] A. Bagnall, J. Lines, A. Bostrom, J. Large, and E. Keogh. The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. *Data Mining and Knowledge Discovery*, Online First, 2016.

[2] Jane L Harvill, Nalini Ravishanker, and Bonnie K Ray. Bispectral-based methods for clustering time series. *Computational Statistics & Data Analysis*, 64:113–131, 2013.

[3] Zhiguang Wang, Weizhong Yan, and Tim Oates. Time series classification from scratch with deep neural networks: A strong baseline. In *Neural Networks (IJCNN), 2017 International Joint Conference on*, pages 1578–1585. IEEE, 2017.