Investigating the Effect of Different Data Normalization Techniques on Time Series Classification Accuracy

Konrad Cybulski

Research Proposal
FIT2082 Research Project

Faculty of Information Technology
Monash University
Australia
15/08/2017

Contents

1 Introduction

The UCR Time Series Classification Archive [1] is a data repository with over 1200 downloads and hundreds of references. While the classification accuracy shown by predictive models on the UCR data is irrefutable, we aim to investigate the effect on classification accuracy the methods used in data normalization have had. With data normalization techniques known to have significant impacts on prediction accuracy on many classifiers, in order to verify the accuracy of classification models an understanding of the impact of techniques used by UCR is necessary.

The UCR Time Series Classification Archive is used as a benchmark dataset for hundreds of time series classification publications. As explored by Keogh & Kasetty [3], there exists a real need for larger testing on real world data due to the bias introduced into time series classification techniques developed and tested on a single benchmark dataset. We aim to determine however whether it is not only the data on which techniques may be over-trained, but the normalization involved in creating such datasets.

The raw unprocessed and non-normalized UCR data has been collected from Anthony Bagnall and Eamonn Keogh with the help of Geoff Web.

2 Aims

This research aims to investigate the effects of methods of data normalization in the context of time series classification. We aim to investigate this particularly on the UCR Time Series Data Archive's [1] raw data. The aspects of data normalization investigated will include scalar data transformations (such as Z-score normalization and *Min-Max* scaling) and the effects of time series length standardization. With regard to time series length standardization, current respective preprocessed UCR time series lengths will be used as a reference and a benchmark for classification accuracy.

3 Research Question

What effect do times series length and methods of data normalization have on classification accuracy?

4 Background

This background focuses on the techniques used in normalization, and literature in the area of the effect of normalization techniques on classifier (and regressor) accuracy.

4.1 Normalization Techniques

4.2 Effect of Normalization on Time Series Classification Accuracy

5 Research Methodology

This method briefly states the platforms which will be used to investigate the proposed research as introduced above.

5.1 Normalization Techniques

The original UCR Time Series Classification Archive data is normalized using z-score normalization. We aim to investigate the two most common scale normalization techniques: z-score normalization and *min-max* normalization. Z-score normalization is most common and is most robust and similar to the original raw data when it conforms to a normal distribution.

6 Timeline

Week:	1	2	3	4	5	6	7	8	9	10	11	12
Initial meeting	-	-	•									
Review of literature	-	-	•	•								
Writing proposal	-	-	•	•								
Replication of current research	-	-		•	•	•						
RSAX Development & Testing	-	-				•	•	•				
Shapelet Scoring & Pruning	-	-							•	•	•	
Project presentation	-	-										•
Final report	-	-	•	•	•	•	•	•	•	•	•	•

References

[1] Yanping Chen, Eamonn Keogh, Bing Hu, Nurjahan Begum, Anthony Bagnall, Abdullah Mueen and Gustavo Batista (2015). The UCR Time Series Classification Archive. URL *www.cs.ucr.edu/ eamonn/time*$_s$*eries*$_d$*ata/*

[2] Petitjean, F., Forestier, G., Webb, G. I., Nicholson, A. E., Chen, Y., & Keogh, E. (2016). Faster and more accurate classification of time series by exploiting a novel dynamic time warping averaging algorithm. *Knowledge and Information Systems*, 47(1), 1-26.

[3] Keogh, E., & Kasetty, S. (2003). On the need for time series data mining benchmarks: a survey and empirical demonstration. *Data Mining and knowledge discovery*, 7(4), 349-371.

www.cs.ucr.edu/ eamonn/time$_s$*eries*$_d$*ata/*