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

4.1 Normalization Techniques

4.2 Effect of Normalization on Time Series Classification Accuracy

4.3 Shapelets

Shapelets are in ever increasing use in the field of time series classification. While NN classification models are still renowned for their high accuracy de-

spite their simplistic nature, shapelets offer not only competitive classification but provide human-readable results [1, 2, 3, 4]. Shapelets aim to determine a key subsequence within a class of classifiable time series' which is representative of that class [3]. They have been shown to be robust to noise due to the shapelet defining a common subsequence in a given class. Additionally being robust to time series length given they represent a common subsequence, length normalization may be omitted in favour of more information rich data.

The use of shapelets in classificaton models involves the creation of decision-trees using the most likely fitting shapelet as a feature [1,2,3]. As a result, in n-class classification, the resulting number of extracted features will be n. These shapelets are produced by maximising occurrences of a subsequence (shapelet) in a given class and minimising occurrences of the same subsequence in other classes. However there still exists a level of information loss due to this, there exist a number of cases where multiple shapelets may define a class to a greater extent than a single subsequence [6,7].

4.4 Distance Measure

Euclidean distance is one of the most common distance measures for both NN classification as well as shapelet based classification methods [1,2,3,4]. Dynamic Time Warping (DTW) has in recent years been shown in a number of cases to improve classification accuracy in NN models [4,5]. The use of shapelets in time series classification has utilised Euclidean distance measures for a large area of research [1,2,3] however DTW as a measure of shapelet subsequence distance has been shown to outperform 1-NN-DTW and existing shapelet models in a number of real world and UCR datasets [4].

In research that uses euclidean distance measures, shapelets are still competitive with regard to classification accuracy [1,2,3,4,6]. Techniques have been developed that are orders of magnitude faster than traditional NN (using numerous distance measures) and exhaustive shapelet search (ES) [1,7]. Despite DTW being a more accurate distance measure in NN classification, techniques involving top k shapelets [6,7] result in lower error associated with euclidean distance measures.

5 Method

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/*