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

Konrad Cybulski

Research Proposal
FIT2082 Research Project

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 Effect of Normalization on Time Series Classification Accuracy

Methods of normalization are known to greatly affect the accuracy of classification in multivariate data sets. When there exists data from two distributions which have greatly different means and variances, normalization becomes an important factor in ensuring each variable does not skew prediction. In univariate datasets this may be less important. However despite this, in a number of predictive models, including neural networks and support vector classifiers, the multidimensional problem space becomes not only easier to train for, but additionally a number of mathematical functions depend on normalized data. In neural networks, the chosen activation functions depend heavily on this fact, with sigmoid activations becoming almost meaningless unless input is within the 0-1 range. Support vector machines require a standardised problem space if the hyperplanes used in the separation of classes can be fitted most accurately. While this itself is a more complex problem, we will focus on time series length standardisation, which is a more pressing and time series related problem.

The UCR Time Series Data Archive's [1] cleaned data has a fixed time series length in each data set, while each data set has a vastly different length, it is important to recognize the effect this has on classification. One reason DTW distance measures are so effective is due to the different rates at which events occur across a number of occurrences. When an event may be registered in given amount of time, the same event might occur in a larger time frame in another occurrence, which using euclidean distance measurements would not match. This event cannot be solved by changing the time series length, however time series length is integral in determining how much information is required in a given time series to have optimal results. If a very small time series length is required to achieve high classification accuracy, the remainder of the raw time series is not required. This idea is similar to the notion of early detection, which is a separate area of time series classification. By changing time series length, both where they begin, and when they end, we control what section of the information and how much of it we allow classifiers to use for prediction.

While more information allows for greater predictive accuracy, it may also introduce noise to the data, and additionally, the larger the data, the longer classification takes. Changes in time series length allows us to understand the information gain/loss, and the slow-down/speed-up associated with it.

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 [1] 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 representative of the original raw data when it conforms to a normal distribution. Z-score normalization involves converting each data point into a positive or negative value representing how many standard deviations away from the mean the data point is. *Min-max* normalization involves converting data into values between 0 and 1 by subtracting the minimum and dividing by the difference between the minimum and maximum values of the time series. These normalization techniques are simple and are very commonly used in order to remove the bias involved in variable with large values compared to other smaller valued data.

5.2 Classification Techniques

The UCR Time Series Classification Archive [1] specifies the lowest classification error possible using a number of techniques.
1-NN Euclidean - This is the error using a one nearest neighbour algorithm with a euclidean distance measure.
1-NN DTW Best Warping Window - The warping window is a hyper-parameter in the NN-DTW classification method which has been determined for each data set along with the error achieved with this optimal window.
1-NN DTW No Warping Window - This is the error when no warping window is used in the NN-DTW classifier.
With solely these three techniques we will determine the accuracy of normalized data. We choose to use these three techniques alone due to their power in time series classification [2] and the generality and simplicity of the algorithms.

While other methods exist and may provide more accurate classifications, the use of NN classifiers will allow a better understanding of the homogeneity and relatedness of the normalized data within a given class.

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	-	-			●	●						
Testing scalar normalization	-	-					●	●				
Testing time series length	-	-							●	●	●	
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/