



Normalization and Bias in Time Series Data

Aishwarya Asesh^(✉)

Adobe, Mountain View, USA
a.asesh@gmail.com

Abstract. Data normalization is an important preprocessing step in data mining and Machine Learning (ML) technique. Finding an acceptable approach to deal with time series normalization, on the other hand, is not an easy process. This is because most standard normalizing approaches rely on assumptions that aren't true for the vast majority of time series. The first is that all time series are stationary, which means that their statistical characteristics, such as mean and standard deviation, do not vary over time. The time series volatility is assumed to be uniform in the second assumption. These concerns are not addressed by any of the approaches currently accessible in the literature. This research provides theoretical and experimental evidence, that normalizing time series data, can prove to be of utmost value by trimming non necessary data points and achieving minimum information loss, by using the concept of Minimal Time Series Representation (MTSR).

Keywords: Time series · Normalization · Classification · Time series length · Minimal time series representation

1 Introduction

The University of California, Riverside (UCR) Time Series Classification Archive [1], has grown into a valuable resource for the time series data mining community, with over a thousand articles citing at least one data set from the repository. While the classification accuracy demonstrated by predictive models on UCR data is undeniable, it is critical to look into the impact of data normalization approaches on classification accuracy. Because data normalization procedures are known to have a substantial influence on prediction accuracy for many classifiers, a knowledge of the impact of UCR's approaches is required to validate the accuracy of classification models.

Due to the bias incorporated into time series classification approaches, created and evaluated on a single benchmark dataset, as discussed by Keogh and Kasetty [3], there is a clear need for broader testing on real-world data. However, it is important to validate if it's not just the data that causes methods to become over-trained, but also the normalization that goes into producing such datasets.

With the aid of raw unprocessed and non-normalized UCR data provided by Geoff Web, Anthony Bagnall and Eamonn Keogh, this research study focuses

on normalization techniques and understanding the influence of normalization approaches on classifier (and regressor) accuracy.

2 Normalization and Time Series Data

Normalization methods are recognized to have a significant impact on classification accuracy in multivariate data sets. When data from two distributions with vastly different means and variances exist, normalization becomes critical in guaranteeing that each variable does not bias prediction. This may be less essential in univariate datasets [4, 5]. Despite this, the multidimensional issue space becomes simpler to train in a variety of predictive models, including neural networks and support vector classifiers, and a number of mathematical functions rely on normalized data. The choice of activation functions in neural networks is greatly influenced by this fact, with sigmoid activations becoming essentially useless until input is in the 0–1 range. If the hyperplanes used in class separation can be fitted most precisely, support vector machines will require a standardised problem space. While this is a more complicated topic in and of itself, this aforementioned research concentrates on time series length standardisation.

While each data set in the UCR Time Series Data Archive has a drastically varied duration, it is crucial to identify the influence this has on categorization. The varying rates at which events occur throughout a number of occurrences is one reason why Dynamic Time Wrapping (DTW) distance measurements are so useful. When an event occurs in a specific length of time, it is possible that the same event will occur in a greater time frame in another occurrence, euclidean distance measurements will not match in these cases [6, 7]. Although altering the time series length will not address this problem, it is critical in evaluating how much information is necessary in a specific time series to get best outcomes. The remainder of the raw time series is not necessary, if using a very short time series length is able too achieve high classification accuracy. This concept is comparable to early detection, which is a distinct field of time series classification. One can regulate which portion of the information and how much of it classifiers may utilize for prediction by altering the length of time series, both where they begin and where they stop [8, 9]. While having more data provides for higher prediction accuracy, it also introduces noise, and the larger the data, the longer it takes to classify it. The information gain/loss, as well as the slow-down/speed-up associated with it, may be understood through changes in time series length.

3 Experiments

The data from the UCR Time Series Classification Archive is not only standardized using z-score normalization, but it is also divided into training and testing subsets. The raw data required to create the UCR datasets includes CricketX, CricketY, CricketZ, GesturesX, GesturesY, GesturesZ, and Wafer. These datasets will be referred to as Cricket, Gestures, and Wafer. Two techniques are used to assess the accuracy of categorization on this data.

The first involves creating a distribution of classification accuracy for a particular data set using random train/test splits of the same size. The second technique is used to see whether there are any discrepancies between the raw data and the data in the UCR repository. To evaluate if there are any major discrepancies between the two datasets, each time series in the raw data is matched with its closest matching time series in the UCR repository data.

3.1 Normalization

Scalar Normalization. The data is normalized using the z-score method. The two most prevalent scale normalizing approaches, z-score normalization and min-max normalization, are used. When the data corresponds to a normal distribution, Z-score normalization is the most frequent and most representational of the original raw data. The process of Z-score normalization entails turning each data point into a positive or negative number that represents how many standard deviations the data point is from the mean.

Min-max normalization includes removing the minimum and dividing by the difference between the time series datapoint minimum and maximum values to transform data into values between 0 and 1. These normalization approaches are straightforward and widely used to eliminate bias in variables with larger values when compared to data with lower values.

Time Series Length Normalization. The data is organized so that each time series in a dataset is the same length. There is no need for stringent normalization techniques with regard to time series length, as in datasets like Cricket, the lengths of time series have a relatively little deviation from the mean. Despite this, when compared to the run durations of the closest neighbor technique, time series length minimization gives a significant computational speedup. The impact of both increasing and lowering the length of time series is an important aspect. Both approaches entail shrinking/expanding a n -length time series to a m -length time series.

For a time series T of length n , the i -th data point in T is represented as T_i . T is converted to a time series of length m , which is denoted as S . The new time series S where each data point in S is as follows:

$$S_i = T_j \quad j = \lfloor n \times \frac{i}{m} \rfloor$$

Information loss is observed when m is less than n , the impact of which will be detailed in the coming sections.

3.2 Classification Techniques

Using a variety of approaches, the UCR Time Series Classification Archive defines the lowest classification error achievable.

- 1-Nearest Neighbor classifier - 1-NN Euclidean - This is the error produced by utilizing one nearest neighbor algorithm with a euclidean distance metric.

- 2 1-NN DTW Best Warping Window - In the NN-DTW classification technique, the warping window is a hyper-parameter that has been computed for each data set, coupled with the error attained with this ideal window.
- 3 1-NN DTW DTW DTW DTW DTW DTW No Warping Window - When the NN-DTW classifier has no warping window, this error occurs.

The correctness of normalized data can be determined by comparing these three approaches. Because of their strength in time series classification [2], the universality and simplicity of the algorithms, these three approaches are chosen. While there are alternative approaches that may yield more accurate classifications, using NN classifiers allows for a better understanding of the homogeneity and relatedness of the normalized data within a particular class.

4 Results and Discussion

While the datasets are extremely diverse in origin, there are obvious parallels in terms of classification accuracy for both Euclidean and DTW distance metrics (Fig. 1).

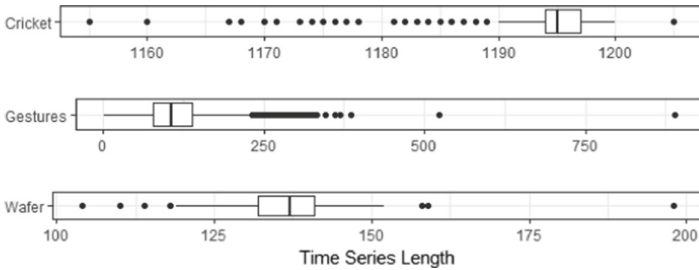


Fig. 1. The distribution of time series length for each of the three datasets.

As the time series length grows, the classification accuracy approaches a maximum, as seen in Fig. 2 and Fig. 3. However, the maximum classification accuracy is obtained at a relatively short time series length and remains almost constant as the duration grows.

The concept of a Minimal Time Series Representation (MTSR), which minimizes data information loss (along with maximizing classification accuracy) is established. It can be observed that there is a minimum information loss in the instance of *Cricket* at a time series length close to the beginning values of the time series, 1200. While this reduces information loss, one can obtain the same classification accuracy with a time series length of 40 and a slightly greater classification accuracy distribution with a time series length of 90. The primary advantage of an MTSR is that it allows for quicker classification owing to the lower computing burden of conducting Euclidean distance measurements on a smaller time series.

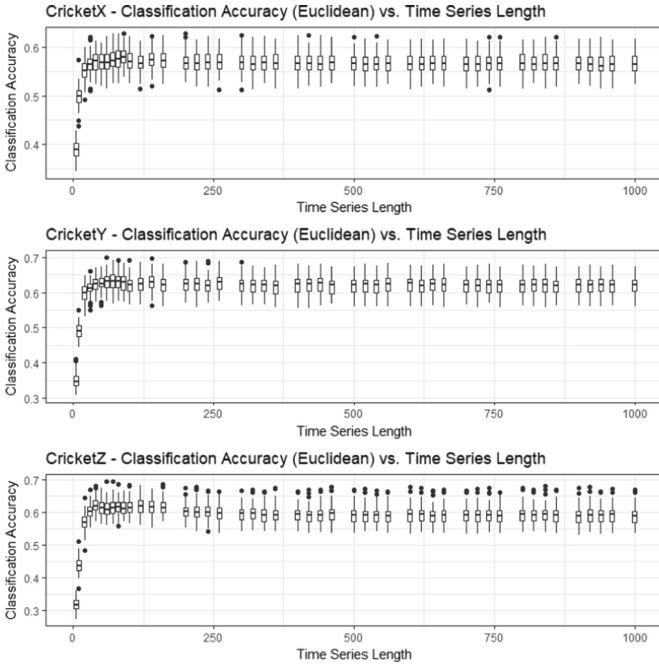


Fig. 2. The distribution of classification accuracy for the *Cricket* dataset as a function of time series length using the Euclidean distance measure in NN prediction.

Similar results are observed for *Wafer* dataset, as illustrated in Fig. 3, where the classification accuracy stays steady as the time series length increases. One can detect a clear maximum classification accuracy for a length of 30 in this dataset. With a time series length of 10–20, a similar accuracy can be achieved as with a length of 300 (Fig. 4).

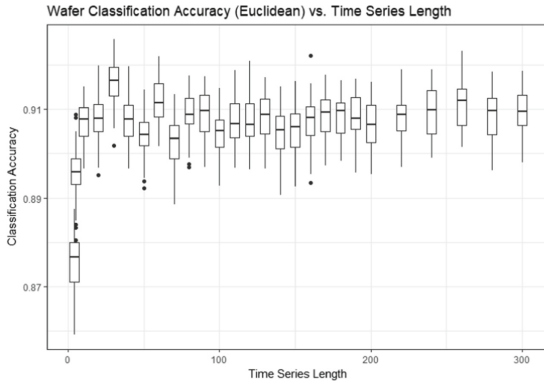


Fig. 3. The distribution of classification accuracy for the *Wafer* dataset as a function of time series length using the Euclidean distance measure in NN prediction.

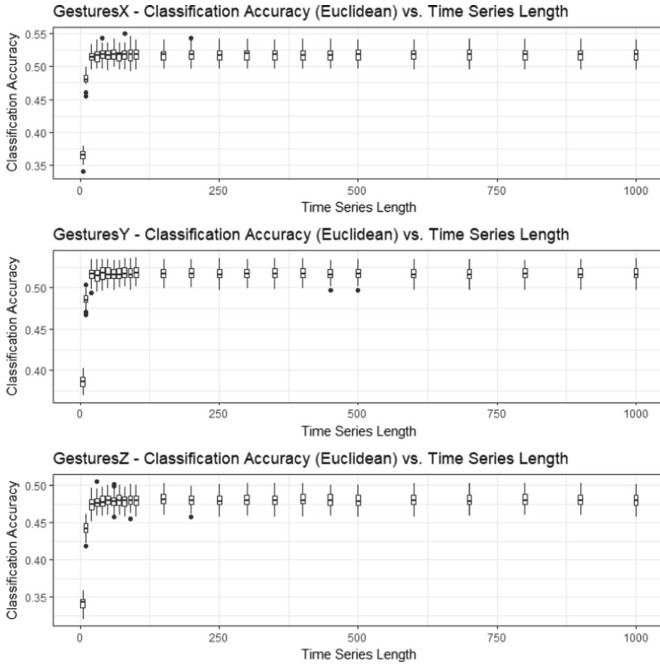


Fig. 4. The distribution of classification accuracy for the *Gestures* dataset as a function of time series length using the Euclidean distance measure in NN prediction.

The results and findings utilizing a DTW distance measure differ considerably from those obtained using Euclidean distance measurements. There is a definite maximum classification accuracy in each of the three datasets. With the following optimum time series lengths and classification accuracies for each dataset:

Dataset	Avg. accuracy	Optimal length	Optimal accuracy
CricketX	0.6299	100	0.7756
CricketY	0.7063	100	0.8353
CricketZ	0.6888	100	0.8061
GesturesX	0.5768	60	0.6459
GesturesY	0.5708	40	0.6498
GesturesZ	0.5293	50	0.5975
Wafer	0.9142	325	0.9214

There is a considerable absolute gain in classification accuracy in both the *Cricket* and *Gesture* datasets. While in case of *Wafer*, it appears to be merely a minor gain in classification accuracy, it is 0.77 standard deviations above the mean accuracy across all duration. This rise in accuracy over the mean is true for both *Cricket* and *Gesture*, with optimal time series length accuracy in *Cricket*

X, Y, and Z being 1.66, 1.36, and 1.24 standard deviations above the mean, respectively. *Gesture* X, Y, and Z have ideal time series length accuracy of 1.12, 1.23, and 1.15 standard deviations above the mean, respectively. These are all considerable gains over average accuracy for various time series lengths, but the most interesting finding is that there is an ideal time series length for DTW accuracy. DTW has an unambiguous maximum with regard to time series length, unlike Euclidean measures, which have a minimal representation of the time series that nevertheless maximizes classification accuracy.

Because of the single maximum, there is considerable symmetry, with lengths on each side of the optimal length resulting in the same accuracy. In *Cricket*, for example, time series lengths of 30 and 550 produce extremely equal classification accuracy, but picking the length of 30 would be optimum for speed of calculation.

There is no simple answer to determining a generic minimal time series length, as there is with practically all sample size questions. It is determined by the number of model parameters to be evaluated as well as the quantity of data randomness. With the number of parameters to be estimated and the level of noise in the data, the sample size required grows (Figs. 5 and 6).

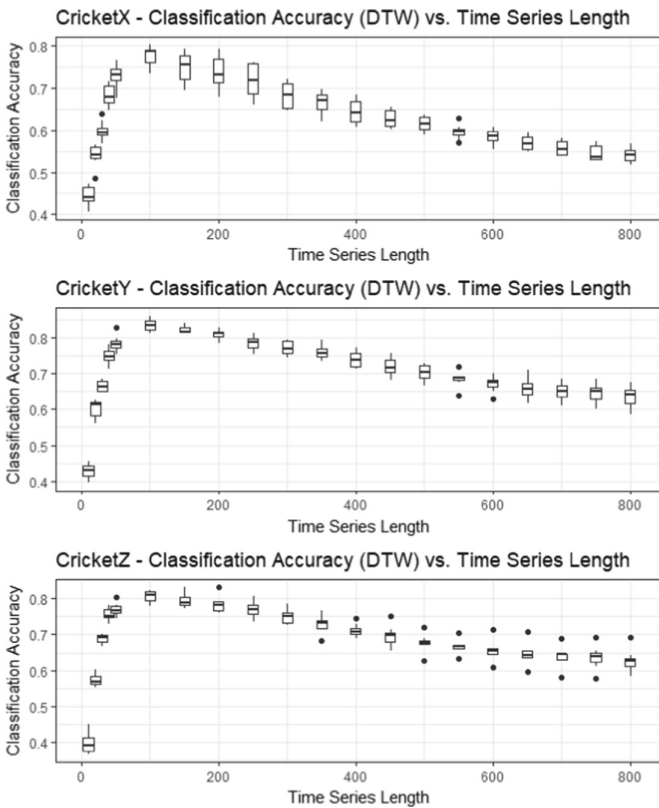


Fig. 5. The distribution of classification accuracy for the *Cricket* dataset as a function of time series length using the DTW distance measure in NN prediction.

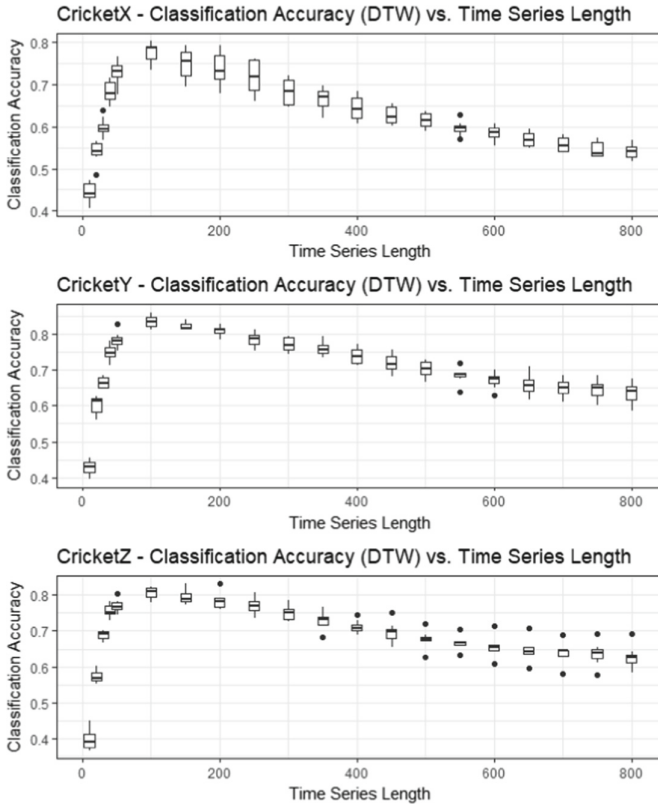


Fig. 6. The distribution of classification accuracy for the *Cricket* dataset as a function of time series length using the DTW distance measure in NN prediction.

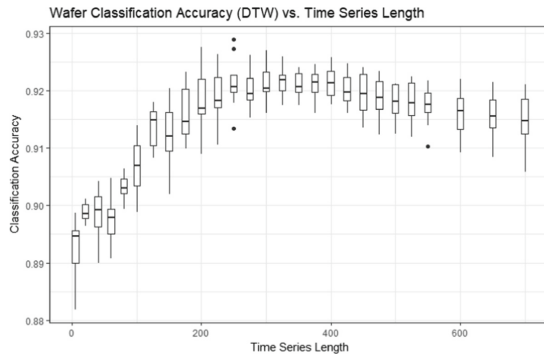


Fig. 7. The distribution of classification accuracy for the *Wafer* dataset as a function of time series length using the DTW distance measure in NN prediction.

5 Conclusions and Future Work

While acknowledged to be less accurate than DTW in time series classification, Euclidean distance measurements have shown to be more stable than DTW. In comparison, DTW has a time series length that is optimal for highest classification accuracy.

The findings are limited to the datasets mentioned. While results about time series length and normalization approaches apply to various datasets, the nature of the data matters when it comes to classification accuracy. As a result, additional research is needed to determine the impact of time series length normalization and scalar normalizing on data from the UCR Time Series Classification Archive and other sources.

In addition, more research into multivariate time series is necessary. The information loss associated with the reduction in time series length has an impact on classification accuracy, as addressed in this paper (both positively and negatively). More study is needed to develop more complex models for determining the smallest time series representations with the least amount of information loss (Fig. 7).

Acknowledgement. All that I am, or ever hope to be, I owe to my angel mother.

References

1. Chen, Y., et al.: The UCR Time Series Classification Archive (2015). www.cs.ucr.edu/~eamonn/time_series_data/
2. Petitjean, F., Forestier, G., Webb, G.I., Nicholson, A.E., Chen, Y., Keogh, E.: Faster and more accurate classification of time series by exploiting a novel dynamic time warping averaging algorithm. *Knowl. Inf. Syst.* **47**(1), 1–26 (2015). <https://doi.org/10.1007/s10115-015-0878-8>
3. Keogh, E., Kasetty, S.: On the need for time series data mining benchmarks: a survey and empirical demonstration. *Data Min. Knowl. Discov.* **7**(4), 349–371 (2003)
4. Deng, J., Chen, X., Jiang, R., Song, X., Tsang, I.W.: ST-Norm: spatial and temporal normalization for multi-variate time series forecasting. In: *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 269–278 (2021)
5. Dette, H., Kokot, K., Volgushev, S.: Testing relevant hypotheses in functional time series via self-normalization. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **82**(3), 629–660 (2020)
6. Boubrahimi, S.F., Hamdi, S.M., Ma, R., Angryk, R.: On the mining of the minimal set of time series data Shapelets. In: *2020 IEEE International Conference on Big Data (Big Data)*, pp. 493–502. IEEE (2020)
7. Denison, R.N., Carrasco, M., Heeger, D.J.: A dynamic normalization model of temporal attention. *Nat. Hum. Behav.* 1–12 (2021)
8. Passalis, N., Tefas, A., Kannianen, J., Gabbouj, M., Iosifidis, A.: Deep adaptive input normalization for time series forecasting. *IEEE Trans. Neural Netw. Learn. Syst.* **31**(9), 3760–3765 (2019)

9. Chonbodeechalermroong, A., Ratanamahatana, C.A.: Robust scale-invariant normalization and similarity measurement for time series data. In: Sieminski, A., Kozierekiewicz, A., Nunez, M., Ha, Q.T. (eds.) *Modern Approaches for Intelligent Information and Database Systems. SCI*, vol. 769, pp. 149–160. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-76081-0_13

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

