Using Reduced SAX Representations for Pruning Randomly Generated Shapelets

Konrad Cybulski

Research Proposal
FIT2082 Research Project

Faculty of Information Technology
Monash University
Australia
10/08/2017

Contents

1 Introduction

In the field of time series classification, Nearest Neighbour (NN) classification models provide highly accurate methods for binary and multiclass classification. However due to the large time complexity associated with NN models, an increasingly common technique involves Nearest Centroid approximations of NN. A recent promising concept in this field are time series shapelets which have shown to be orders of magnitude faster than NN classifiers with comparable accuracy.

As noted by Rakthanmanon & Keogh (2013), in comparison with NN, a "*lazy* classifier", shapelets lead to *eager* classifiers. Current shapelet discovery methods (and models utilising shapelets for classification) aim to determine class representative shapelets. Resultantly there is some amount of information loss when a single shapelet is used as a class identifier [6,7].

This research aims to utilise sequential aggregate approximation (SAX) for shapelet distance approximation to aid randomly generated shapelet pruning. This research is based on the principle proposed by Wisstuba et al. (2015) that a class representative shapelet will occur often in a dataset. Under this assumption the random generation of shapelets will produce class representative shapelets for a given dataset. The algorithm designed by Wisstuba et al. (2015) converts minimum shapelet distances from time series data into a transformed feature set. This transformation involves calculating minimum subsequence distance over numerous shapelets which becomes infeasible for large datasets or large numbers of shapelets (which invariably produce more accurate results). Using approximate distance measures to determine class representative shapelets will thus reduce the necessary computation to attain accurate prediction from the *ultra-fast shapelet* algorithm.

We additionally propose an approximate time series representation based on SAX to aid in fast shapelet scoring and pruning. In preliminary experiments this reduced SAX form has shown comparable accuracy to conventional SAX representation and unsimplified time series representations in shapelet based classification. These results, coupled with faster shapelet discovery and shapelet scoring demonstrate the need for further investigation.

2 Background

2.1 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 despite 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].

2.2 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.

3 Method

Below we outline the methods by which both the shapelet discovery algorithm and reduced SAX representations will be investigated and tested both experimentally and formally.

3.1 Reduced Sequential Aggregate Approximation (RSAX)

3.2 Shapelet Discovery

3.2.1 Shapelet Scoring and Pruning

4 Preliminary Experimental Results

In order to provide motivation for further investigation of the ideas mentioned in the introduction to this research, preliminary experiments have been conducted with the following results.

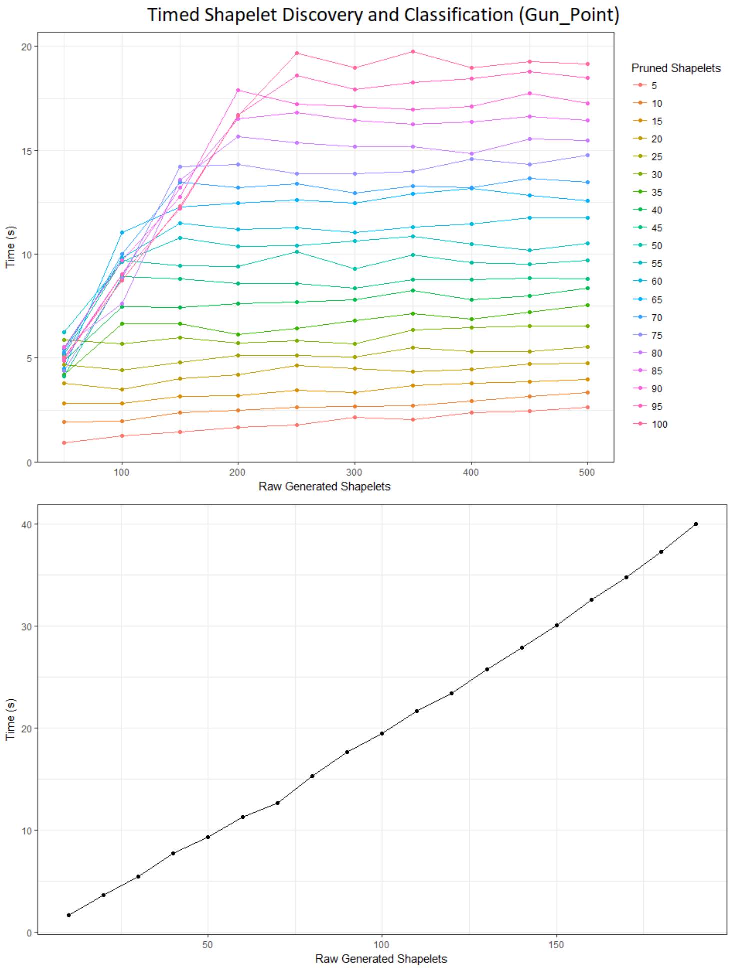

Figure 1: Plot of time taken for shapelet discovery and classification model training and testing. **Top**: Time taken using reduced SAX pruning methods. **Bottom**: Time taken using *Ultra-Fast Shapelets* algorithm.

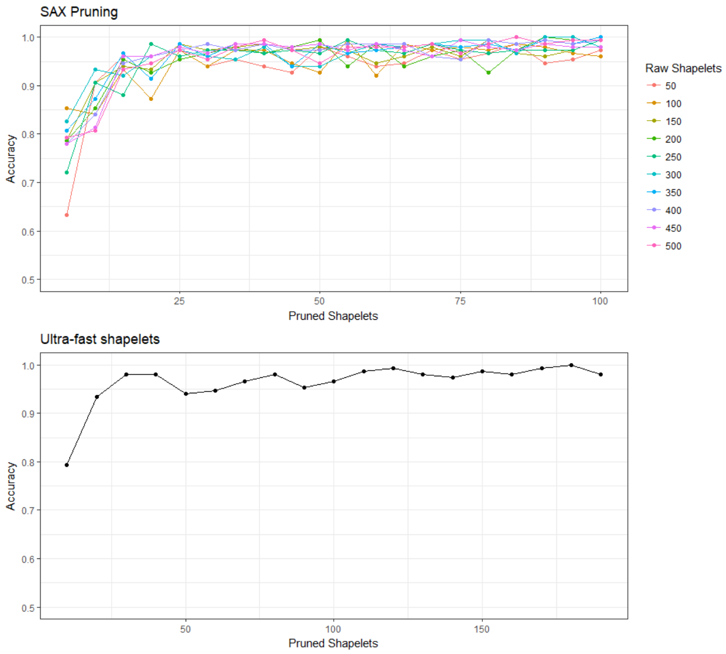

Figure 2: Plot of accuracy of classification models created from shapelets discovered using two methods. **Top**: Accuracy using reduced SAX pruning methods. **Bottom**: Accuracy using *Ultra-Fast Shapelets* algorithm.

5 Further Work

References

[1] Rakthanmanon, T., & Keogh, E. (2013, May). Fast shapelets: A scalable algorithm for discovering time series shapelets. In *Proceedings of the 2013 SIAM International Conference on Data Mining* (pp. 668-676). Society for Industrial and Applied Mathematics. Retrieved from http://epubs.siam.org/doi/pdf/10.1137/1.9781611972832.74

[2] Raza, A., & Kramer, S. (2017). Ensembles of Randomized Time Series Shapelets Provide Improved Accuracy while Reducing Computational Costs. *arXiv preprint arXiv*:1702.06712.

[3] Ye, L., & Keogh, E. (2011). Time series shapelets: a novel technique that allows accurate, interpretable and fast classification. *Data mining and knowledge discovery*, 22(1), 149-182.

[4] Shah, M., Grabocka, J., Schilling, N., Wistuba, M., & Schmidt-Thieme, L. (2016, March). Learning DTW-shapelets for time-series classification. In *Proceedings of the 3rd IKDD Conference on Data Science*, 2016 (p. 3). ACM.

[5] 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.

[6] Lines, J., Davis, L. M., Hills, J., & Bagnall, A. (2012, August). A shapelet transform for time series classification. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 289-297). ACM.
Available at **https://ueaeprints.uea.ac.uk/40201/1/LinesKDD2012.pdf**

[7] Wistuba, M., Grabocka, J., & Schmidt-Thieme, L. (2015). Ultra-fast shapelets for time series classification. *arXiv preprint* arXiv:1503.05018.