k-Nearest Shapelets for Time Series Classification

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. Which we hypothesise contributes to low classification accuracy among more complex time series.
This research aims to investigate and produce a k-Nearest Shapelet classification method for time series which aims to increase classification accuracy among larger and more complex time series data.

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. 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, exemplified in cases where multiple shapelets may define a class to a greater extent than a single subsequence.

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.