Application of HSIC-Lasso for high-dimensional feature selection in shapelet-based decomposition

JULIEN PELAMATTI
EDF R&D, Chatou, France

IBRAHIM SEYDI
Sorbonne Université, Paris, France

NICOLAS BOUSQUET
EDF R&D, Chatou, France

In this work, we present a novel approach for selecting an optimal projection basis for time series objects when using shapelet decomposition in a classification framework, leveraging a Lasso-like high-dimensional non-linear feature selection method based on the Hilbert-Schmidt Independence Criterion (HSIC)[1].

Time series analysis frequently faces challenges such as high dimensionality, autocorrelation, and the difficulty of identifying key features that capture essential dynamics across different temporal scales and phase offsets. To address these issues, we employ shapelet decomposition[2], a technique designed to extract shape-based features from time series, preserving both temporal and frequency information. The core idea is to represent a time series dataset through minimal distances to specific representative patterns. Shapelet decomposition not only performs well compared to other state-of-the-art methods for time series learning but also provides enhanced interpretability by identifying patterns most relevant to the algorithm decisions, often offering insights into the physical meaning behind these decisions.

Given the large number of possible shapes that can be extracted from a dataset, selecting the most relevant shapes-i.e., the most informative projection basis-is crucial for numerical tractability. In most prior works, this selection is achieved by iteratively optimizing the information gain among a set of candidate patterns[3]. However, this approach has two main drawbacks: its high computational cost due to two nested optimization problems and the potential for selecting interdependent features, leading to redundant information.

As an alternative, we propose selecting the optimal shapelet decomposition basis using HSIC Lasso, a Lasso-like non-linear high-dimensional feature selection method that uses HSIC to identify a sparse subset of the most informative and mutually independent features[4]. This approach requires only a single optimization loop over the Lasso weights, making it a significantly more computationally efficient alternative to the standard method. Additionally, as already mentioned, the set of informative features selected by this approach is so that its components do not present strong interdependence.

We validate our approach on both synthetic and real-world datasets, demonstrating its potential to improve performance, scalability, and interpretability of time series classification models. Our method offers a powerful tool for a wide range of application domains.

References:

[1] Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. (2005, October). Measuring statistical dependence with Hilbert-Schmidt norms. In International conference on algorithmic learning theory (pp. 63-77). Berlin, Heidelberg: Springer Berlin Heidelberg.

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

[3] Hills, J., Lines, J., Baranauskas, E., Mapp, J., and Bagnall, A. (2014). Classification of time series by shapelet transformation. Data mining and knowledge discovery, 28, 851-881.

[4] Yamada, M., Jitkrittum, W., Sigal, L., Xing, E. P., and Sugiyama, M. (2014). High-dimensional feature selection by feature-wise kernelized lasso. Neural computation, 26(1), 185-207.

[Julien Pelamatti; EDF R&D]
[`julien.pelamatti@edf.fr` –]