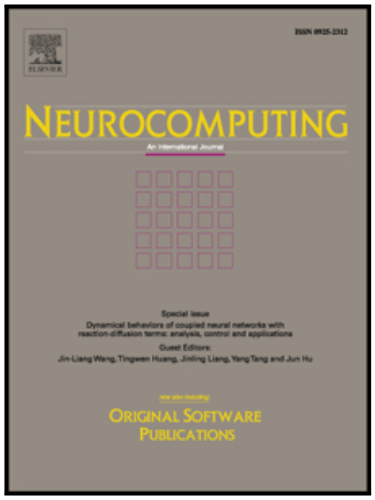
Communicated by Dr Francesco Dinuzzo

Accepted Manuscript



Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests (tsfresh – A Python package)

Maximilian Christ, Nils Braun, Julius Neuffer,
Andreas W. Kempa-Liehr

PII: S0925-2312(18)30484-3
DOI: 10.1016/j.neucom.2018.03.067
Reference: NEUCOM 19517

To appear in: *Neurocomputing*

Received date: 31 May 2017
Revised date: 22 March 2018
Accepted date: 23 March 2018

Please cite this article as: Maximilian Christ, Nils Braun, Julius Neuffer, Andreas W. Kempa-Liehr, Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests (tsfresh – A Python package), *Neurocomputing* (2018), doi: 10.1016/j.neucom.2018.03.067

Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests (tsfresh – A Python package)

Maximilian Christ[a], Nils Braun[b], Julius Neuffer[a], Andreas W. Kempa-Liehr[c,d,*]

[a] *Blue Yonder GmbH, Karlsruhe, Germany*
[b] *Institute of Experimental Particle Physics, Karlsruhe Institute of Technology, Germany*
[c] *Department of Engineering Science, University of Auckland, New Zealand*
[d] *Freiburg Materials Research Center, University of Freiburg, Germany*

Abstract

Time series feature engineering is a time-consuming process because scientists and engineers have to consider the multifarious algorithms of signal processing and time series analysis for identifying and extracting meaningful features from time series. The Python package `tsfresh` (Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests) accelerates this process by combining 63 time series characterization methods, which by default compute a total of 794 time series features, with feature selection on basis automatically configured hypothesis tests. By identifying statistically significant time series characteristics in an early stage of the data science process, `tsfresh` closes feedback loops with domain experts and fosters the development of domain specific features early on. The package implements standard APIs of time series and machine learning libraries (e.g. `pandas` and `scikit-learn`) and is designed for both exploratory analyses as well as straightforward integration into operational data science applications.

Keywords: Feature Engineering, Time series, Feature Extraction, Feature Selection, Machine Learning

1. Introduction

Trends such as the Internet of Things (IoT) [1], Industry 4.0 [2], and precision medicine [3] are driven by the availability of cheap sensors and advancing connectivity, which among others increases the availability of temporally annotated data. The resulting time series are the basis for machine learning applications like the classification of hard drives into risk classes concerning a specific defect

*Corresponding author
Email addresses: `max.christ@me.com` (Maximilian Christ), `nilslennartbraun@gmail.com` (Nils Braun), `julius.neuffer@gmail.com` (Julius Neuffer), `a.kempa-liehr@auckland.ac.nz` (Andreas W. Kempa-Liehr)