Distributed and parallel time series feature extraction for industrial big data applications

Maximilian Christ\textsuperscript{a}, Andreas W. Kempa-Liehr\textsuperscript{b,c}, Michael Feindt\textsuperscript{a,1}

\textsuperscript{a}Blue Yonder GmbH, Karlsruhe, Germany
\textsuperscript{b}Department of Engineering Science, University of Auckland, Auckland, New Zealand
\textsuperscript{c}Freiburg Materials Research Center, University of Freiburg, Freiburg, Germany

Abstract

The all-relevant problem of feature selection is the identification of all strongly and weakly relevant attributes. This problem is especially hard to solve for time series classification and regression in industrial applications such as predictive maintenance or production line optimization, for which each label or regression target is associated with several time series and meta-information simultaneously.

Here, we are proposing an efficient, scalable feature extraction algorithm for time series, which filters the available features in an early stage of the machine learning pipeline with respect to their significance for the classification or regression task, while controlling the expected percentage of selected but irrelevant features.

The proposed algorithm combines established feature extraction methods with a feature importance filter. It has a low computational complexity, allows to start on a problem with only limited domain knowledge available, can be trivially parallelized, is highly scalable and based on well studied non-parametric hypothesis tests. We benchmark our proposed algorithm on all binary classification problems of the UCR time series classification archive as well as time series from a production line optimization project and simulated stochastic processes with underlying qualitative change of dynamics.

Keywords: Feature Engineering, Feature Selection, Time Series Feature Extraction

2010 MSC: 37M10

\textsuperscript{*}Corresponding author

Email addresses: maximilian.christ@blue-yonder.com (Maximilian Christ), a.kempa-liehr@auckland.ac.nz (Andreas W. Kempa-Liehr), michael.feindt@blue-yonder.com (Michael Feindt)

\textsuperscript{1}on leave of absence from Karlsruhe Institute of Technology
1. Introduction

Promising fields of application for machine learning are the Internet of Things (IoT) [1] and Industry 4.0 [2] environments. In these fields, machine learning models anticipate future device states by combining knowledge about device attributes with historic sensor time series. They permit the classification of devices (e.g. hard drives) into risk classes with respect to a specific defect [3]. Both fields are driven by the availability of cheap sensors and advancing connectivity between devices, which increases the need for machine learning on temporally annotated data.

In most cases the volume of the generated time series data forbids their transport to centralized databases [1]. Instead, algorithms for an efficient reduction of the data volume by means of feature extraction and feature selection are needed [1, p. 125–136]. Furthermore, for online applications of machine learning it is important to continuously select relevant features in order to deal with concept drifts caused by qualitative changes of the underlying dynamics [5].

Therefore, for industrial and other applications, one needs to combine distributed feature extraction methods with a scalable feature selection, especially for problems where several time series and meta-information have to be considered per label/target [6]. For time series classification, it proved to be efficient to apply comprehensive feature extraction algorithms and then filter the respective features [7].

Motivated by industrial applications for machine learning models [8], we are extending the approach of Fulcher and Jones [7] and propose Feature Extraction based on Scalable Hypothesis tests (FRESH). The algorithm characterizes time series with comprehensive and well-established feature mappings and considers additional features describing meta-information. In a second step, each feature vector is individually and independently evaluated with respect to its significance for predicting the target under investigation. The result of these tests is a vector of p-values, quantifying the significance of each feature for predicting the label/target. This vector is evaluated on basis of the Benjamini-Yekutieli procedure [9] in order to decide which features to keep.

The proposed algorithm is evaluated on all binary classification problems of the UCR time series classification archive [10] as well as time series data from a production line optimization project and simulated time series from a stochastic process with underlying qualitative change of dynamics [11, p. 164]. The results are benchmarked against well-established feature selection algorithms like linear discriminant analysis [7] and the Boruta algorithm [12], but also against Dynamic Time Warping [13]. The analysis shows that the proposed method outperforms Boruta based feature selection approaches as well as Dynamic Time Warping based approaches for problems with large feature sets and large time series samples. This contribution closes with a summary and an outlook on future work.

The FRESH algorithm and the comprehensive framework to extract the features have also been implemented in a Python package called tsfresh. Its
2. Time series feature extraction

Temporally annotated data come in three different variants [17]: Temporally invariant information (e.g., the manufacturer of a device), temporally variant information, which change irregularly (e.g., process states), and temporally variant information with regularly updated values (e.g., sensor measurements). The latter describe the continuously changing state $s_{i,j}(t)$ of a system or device $s_i$ with respect to a specific measurement of sensor $j$, which is repeated in intervals of length $\Delta t$. This sampling captures the state of the system or device under investigation as a sequence

\[ s_{i,j}(t_1) \rightarrow s_{i,j}(t_2) \rightarrow \ldots \rightarrow s_{i,j}(t_{\nu}) \rightarrow \ldots \rightarrow s_{i,j}(t_{n_i}) \]  

(1a)

with $t_{\nu+1} = t_{\nu} + \Delta t$. Such kind of sequences are called time series and are abbreviated by

\[ S_{i,j} = (s_{i,j}(t_1), s_{i,j}(t_2), \ldots, s_{i,j}(t_{\nu}), \ldots, s_{i,j}(t_{n_i}))^T \]

\[ = (s_{i,j,1}, s_{i,j,2}, \ldots, s_{i,j,\nu}, \ldots, s_{i,j,n_i})^T. \]  

(1b)

Here, we are considering $i = 1, \ldots, m$ devices with $j = 1, \ldots, n$ different time series per device. Therefore, we are dealing with $n \cdot m \cdot n_i$ values describing the ensemble under investigation. In order to characterize a time series with respect to its dynamics and reduce the data volume, a mapping $\theta_k : \mathbb{R}^{n_i} \rightarrow \mathbb{R}$ is introduced, which captures a specific aspect $k$ of the time series. One example for such mapping might be the maximum operator

\[ \theta_{\max}(S_{i,j}) = \max\{s_{i,j,1}, s_{i,j,2}, \ldots, s_{i,j,\nu}, \ldots, s_{i,j,n_i}\}, \]

which quantifies the maximal value ever recorded for time series $s_{i,j}$. This kind of lower dimensional representation is called a feature, which is a measurable characteristic of the considered time series. Other examples for feature mappings $\theta_k$ of time series might be their mean, the number of peaks with a certain steepness, their periodicity, a global trend, etc. Comprehensive collections of time series feature mappings are discussed by Fulcher and Jones [7] and Nune et al. [18]. Fulcher and Jones [7] propose to use more than 9000 features from 1000 different feature generating algorithms. A comprehensive list of feature mappings used for this evaluation is given in Appendix A.

Now, consider $n_f$ different time series feature mappings, which are applied to all $m \cdot n$ time series recorded from $n$ sensors of $m$ devices (Fig. 1). The resulting feature matrix $X \in \mathbb{R}^{m \times n_\phi}$ has $m$ rows (one for each device) and $n_\phi = n \cdot n_f + n_i$. 
columns with $n_i$ denoting the number of features generated from device specific meta-information. Each column of $X$ comprises a vector $X \in \mathbb{R}^m$ capturing a specific characteristic of all considered devices.

### 3. Feature filtering

Typically, time series are noisy and contain redundancies. Therefore, one should keep the balance between extracting meaningful but probably fragile features and robust but probably non-significant features. Some features such as the median will not be heavily influenced by outliers, others such as $\max$ will be intrinsically fragile. The choice of the right time series feature mappings is crucial to capture the right characteristics for the task at hand.

#### 3.1. Relevance of features

A meaningless feature describes a characteristic of the time series that is not useful for the classification or regression task at hand. Radivojac et al. [19] considered a binary target $Y$, stating that the relevance of feature $X$ is measured as the difference between the class conditional distributions $f_{X|Y=y_1}$ and $f_{X|Y=y_2}$. We adopt this definition and consider a feature $X$ being relevant for the classification of the binary target $Y$ if those distributions are not equal. In general, a feature $X$ is relevant for predicting target $Y$ if and only if

$$\exists \ y_1, y_2 \text{ with } f_Y(y_1) > 0, f_Y(y_2) > 0 : f_{X|Y=y_1} \neq f_{X|Y=y_2}. \quad (2)$$

---

Figure 1: Data processing tiers of the filtered feature extraction algorithm. All but the Benjamini-Yekutieli procedure can be computed in parallel.
The condition from Eq. (2) is equivalent to

\[ X \text{ is not relevant for target } Y \iff \forall y_1, y_2 \text{ with } f_Y(y_1) > 0, f_Y(y_2) > 0 : f_{X|Y=y_1} = f_{X|Y=y_2} \]
\[ \iff \forall y_1 \text{ with } f_Y(y_1) > 0 : f_{X|Y=y_1} = f_X \iff f_{X,Y} = f_X f_Y \]
\[ \iff X, Y \text{ are statistically independent} \tag{3} \]

We will use the statistical independence to derive a shorter definition of a relevant feature:

**Definition 1 (A relevant feature).** A feature \( X_\phi \) is relevant or meaningful for the prediction of \( Y \) if and only if \( X_\phi \) and \( Y \) are not statistically independent.

### 3.2. Hypothesis tests

For every extracted feature \( X_1, \ldots, X_\phi, \ldots, X_n \) we will deploy a singular statistical test checking the hypotheses

\[ H_{\phi}^0 = \{ X_\phi \text{ is irrelevant for predicting } Y \}, \]
\[ H_{\phi}^1 = \{ X_\phi \text{ is relevant for predicting } Y \}. \tag{4} \]

The result of each hypothesis test \( H_{\phi}^0 \) is a so-called p-value \( p_\phi \), which quantifies the probability that feature \( X_\phi \) is not relevant for predicting \( Y \). Small p-values indicate features, which are relevant for predicting the target.

Based on the vector \( (p_1, \ldots, p_{n_\phi})^T \) of all hypothesis tests, a multiple testing approach will select the relevant features (Sec. 3.3). We propose to treat every feature uniquely by a different statistical test, depending on whether the codomains of target and feature are binary or not. The usage of one general feature test for all constellations is not recommended. Specialized hypothesis tests yield a higher statistical power due to more assumptions about the codomains that can be used during the construction of those tests. The proposed feature significance tests are based on nonparametric hypothesis tests, which do not make any assumptions about the distribution of the variables, thus ensuring robustness of the procedure.

**Exact Fisher test of independence:** This feature significance test can be used if both the target and the inspected feature are binary. Fisher’s exact test [20] is based on the contingency table formed by \( X_\phi \) and \( Y \). It inspects if both variables are statistically independent, which corresponds to the hypotheses from Eq. (1). Fisher’s test belongs to the class of exact tests. For such tests, the significance of the deviation from a null hypothesis (e.g., the p-value) can be calculated exactly, rather than relying on asymptotic results.

**Kolmogorov-Smirnov test (binary feature):** This feature significance test assumes the feature to be binary and the target to be continuous. In general, the Kolmogorov-Smirnov (KS) test is a non-parametric and stable goodness-of-fit test, which checks if two random variables \( A \) and \( B \) follow the same distribution [21]:

\[ H_0 = \{ f_A = f_B \}, \]
\[ H_1 = \{ f_A \neq f_B \}. \]
By conditionally modeling the distribution function of target $Y$ on the two possible values $x_1, x_2$ of the feature $X_\phi$ we can use the KS test to check if the distribution of $Y$ differs given different values of $X_\phi$. Setting $A = Y|X_\phi = x_1$ and $B = Y|X_\phi = x_2$ results in

$$H_0^\phi = \{f_{Y|X_\phi=x_1} = f_{Y|X_\phi=x_2}\}, \quad H_1^\phi = \{f_{Y|X_\phi=x_1} \neq f_{Y|X_\phi=x_2}\}. \quad (5)$$

The hypotheses from Eq. (4) and (5) are equivalent as demonstrated in the chain of Eqns. (3). Hence, the KS test can address the feature relevance of $X_\phi$.

**Kolmogorov-Smirnov test (binary target):** When the target is binary and the feature non-binary, we can deploy the Kolmogorov-Smirnov test again. We have to switch roles of target and feature variable, resulting in the testing of the following hypothesis:

$$H_0^\phi = \{f_{X_\phi|Y=y_1} = f_{X_\phi|Y=y_2}\}, \quad H_1^\phi = \{f_{X_\phi|Y=y_1} \neq f_{X_\phi|Y=y_2}\}.$$

This time $y_1$ and $y_2$ are the two possible values of $Y$ and $f_{X_\phi|Y=y_j}$ is the conditional density function of $X_\phi$ given $Y$. This hypothesis is also equivalent to the one in Eq. (4).

**Kendall rank test:** This filter can be deployed if neither target nor feature are binary. Kendall’s rank test [22] checks if two continuous variables may be regarded as statistically dependent, hence naturally fitting our hypotheses from Eq. (4). It is a non-parametric test based on Kendall’s rank statistic $\tau$, measuring the strength of monotonic association between $X_\phi$ and $Y$. The calculation of the rank statistic is more complex when ties are involved [23], i.e. feature or target are categorical.

### 3.3. Feature significance testing

In the context of time series feature extraction, a wrongly added feature is a feature $X_\phi$ for which the null hypothesis $H_0^\phi$ has been rejected by the respective feature significance test, even though $H_0^\phi$ is true. The risk of such a false positive result is by construction of the hypothesis tests only controlled for individual features. However, when comparing multiple hypotheses and features
simultaneously, errors in the inference tend to accumulate \[24\]. In multiple testing, the expected proportion of erroneous rejections among all rejections is called false discovery rate (FDR).

The FDR as a measure of the accumulated statistical error was suggested by Benjamini and Hochberg \[25\]. Later the non-parametric Benjamini-Yekutieli procedure was proposed. Based on the p-values it tells which hypotheses to reject while still controlling the FDR under any dependency structure between those hypotheses \[9\]. It will be the last component of our filtered feature extraction algorithm.

The procedure searches for the first intersection between the ordered sequence of p-values \(p(\phi)\) (blue dots in Fig. 2) with a linear sequence (green lines in Fig. 2)

\[
r_\phi = \frac{\phi q}{n_\phi \sum_{\mu=1}^{\phi} \frac{1}{\mu}}.
\]

Here, \(n_\phi\) is the number of all tested null hypotheses and \(q\) is the FDR level that the procedure controls. It will reject all hypotheses belonging to p-values, which have a lower value than the p-value at the intersection (Fig. 2(b)).

By deploying the Benjamini-Yekutieli procedure, a global inference error measure, the False Extraction Rate (FER) is controlled:

\[
\text{FER} = \mathbb{E} \left[ \frac{\text{number of irrelevant extracted features}}{\text{number of all extracted features}} \right]
\]

The following algorithm controls the FER for all distributions of features and target as well as for every dependency structure asymptotically.

### 3.4. The proposed feature extraction algorithm

We propose \textbf{F}eature \textbf{E}xtraction based on \textbf{S}calable \textbf{H}ypothesis tests (FRESH) for parameter \(q \in [0, 1]\), given by the following three steps:

1. Perform a set of \(n_\phi\) univariate feature mappings as introduced in Sec. 2 on \(m \cdot n\) different time series to create the feature vectors \(X_{\phi}\) with \(\phi = 1, \ldots, n_\phi\).

2. For each generated feature vector \(X_1, \ldots, X_{n_\phi}\) perform exactly one hypothesis test for the hypothesis \(H_{0\phi}\) from Eq. (4). To do so, take the corresponding feature significance test from Sec. 3.2. Calculate the p-values \(p_1, \ldots, p_{n_\phi}\) of the tests.

3. Perform the Benjamini-Yekutieli procedure under correction for dependent hypotheses \[9\] for a FDR level of \(q\) on the collected p-values \(p_1, \ldots, p_{n_\phi}\) in order to decide which null hypothesis \(H_{0\phi}\) to be rejected (c.f. Sec. 3.3). Only return features vectors for which the respective hypothesis \(H_{0\phi}\) was rejected by the procedure.
3.5. Variants of FRESH

A problem of many filter feature methods such as the one we utilize in steps 2 and 3 of FRESH is the redundancy in the feature selection. As long as features are considered associated with the target, they will all be selected by the filter even though many of them are highly correlated to each other [26]. For example, median and mean are highly correlated in the absence of outliers and therefore we expect FRESH to either select or drop both median and mean at the same time. To avoid generating a group of highly correlated features we propose to add another step to FRESH:

* Normalize the features and perform a principal performance analysis (PCA).
  Keep the principal components with highest eigenvalue describing $p$ percent of the variance.

This step will reduce the number of features and the obtained principal components are de-correlated, orthogonal variables [27].

One could perform step * between steps 1 and 2 of FRESH to get rid of the correlations between the created variables early. Then the feature significance tests in step 2 of FRESH will take principal components instead of the original features as input. We will denote this variant of FRESH as FRESH_PCAb(efore). Also, one could perform step * after the FRESH algorithm, directly after step 3. This means that the PCA will only process those features, which are found relevant by the FRESH algorithm instead of processing all features. This variant of FRESH is called FRESH_PCAa(fter).

4. Evaluation

In the following, the performance of FRESH, its two variants from Sec. 3.5 and other time series feature extraction methods are compared. The evaluation is done with respect to both the meaningfulness of the extracted features as well as the time it takes to extract the features. While doing so, all feature extraction methods operate on the same feature mappings, the differences lay only in the used feature selection process.

4.1. Setup

FRESH is parameterized with $q = 10\%$, cf. Eq. (6). Its variants, which apply a PCA, are using $p = 95\%$ (Sec. 3.5). Full_X uses all the features, which are created during step 1 of FRESH (Sec. 3.4) without any subsequent filtering. Features contained in Full_X will be filtered by applying the Boruta feature selection algorithm [12], or by a forward selection with a linear discriminant analysis classifier denoted LDA. Further, the direct classifier DTW_NN, a nearest neighbor search under the Dynamic Time Warping distance, is considered [28]. Those six different extraction methods and DTW_NN were each picked for a reason. DTW_NN is reported to reach the highest accuracy rates among other time series classifiers. LDA was the first proposed algorithm to automatically extract features
from time series [7] and **Boruta** is a promising feature selection algorithm that incorporates interactions between features.

To guarantee reproducibility we use all 31 time series data sets from the UCR time series archive [10] containing a binary classification problem. In order to compare the runtime of the different methods, time series of flexible length and sample number belonging to two classes are generated by simulating the stochastic dynamics of a dissipative soliton [11, p. 164]. The last data source originates from the production of steel billets, extracted during the German research project iPRODICT. The project demonstrates a typical application of industrial time series analysis, aiming to predict the passing or failing of product specification testings based on temporally annotated data. It contains 26 univariate meta-variables forming the baseline feature set extended by 20 different sensor time series having up to 44 data points for each sample. The data set contained 1554 samples of two classes "broken" and "not broken" each, in total 3108 samples.

Due to limitations regarding the size of this paper, we do not address the regression performance of **FRESH** and variants yet. In a future work, we will catch up on this.

4.2. Accuracy

For the data sets from the UCR time series repository as well as the iPRODICT data, the underlying structure and therefore the relevant features are unknown. We cannot compare the different methods on their ability to extract meaningful features because we do not know which features are meaningful and which are not. Also, we cannot compare the extracted features to direct classifiers such as **DTW_NN**. Therefore, we evaluate the performance of the feature extraction algorithms by comparing the performance of a classification algorithm on the extracted features. Hereby, we assume that more meaningful features will result in a better classification result.

To investigate the feature extraction methods under different conditions and to make sure that the feature filtering is not tuned to a specific classifier, the classification task is solved by a collection of a one layer neural network/perceptron (NN), a logistic regression model (LR), a support vector machine (SVM), a random forest classifier (RFC) and an adaptive boosting classifier (ABC). The hyperparameters for those methods are not optimized to get an unbiased view on the meaningfulness of the extracted features, instead the default values from the Python package scikitlearn version 0.17.1 were used [14].

For every data set, all available samples will be used to perform the feature extraction itself. One third of the samples are randomly picked for a test set, the remaining two thirds are used to train the classifiers. We define the index set for the classification algorithms as $\mathcal{M} := \{\text{NN, SVM, RFC, ABC, LR}\}$ and for the inspected feature extraction methods we define $\mathcal{A} := \{\text{FRESH, LDA, FULL_X, FRESH_PCAa, FRESH_PCab, Boruta}\}$. Then, the accuracy of a classifier $m \in \mathcal{M}$ on the test part of the data set $d$ for the features selected by method $a \in \mathcal{A}$ is denoted as $acc_m^d(a)$. 
Further, we calculate the mean of the accuracy of the five classification algorithms, which is denoted by $\text{acc}^d(a) := \frac{1}{5} \sum_{m \in M} \text{acc}^d_m(a)$. DTW\_NN itself does not perform any feature extraction and its accuracy is directly calculated by predicting on the test set. It is denoted by $\text{acc}^d(\text{DTW\_NN})$. Now, we are able to denote the average accuracy over all 31 UCR data sets with binary classification tasks \cite{UCRArchive} for each method $a \in A \cup \{\text{DTW\_NN}\}$ by $\text{acc}^{UCR}(a) = \frac{1}{31} \sum_{i=1,...,31} \text{acc}^d_i(a)$, cf. third columns of both Tab. 1(a) and Tab. 1(b).

From column $\text{acc}^{UCR}$ of Tab. 1(a) we can observe that FRESH\_PCAa dominated the feature based approaches on the UCR data sets but it was not able to beat DTW\_NN. On the iPRODICT data, DTW\_NN could only operate on one type of time series without the univariate features. This seems to be the reasons why Boruta and FRESH\_PCAa beat the accuracy of DTW\_NN as shown in column $t_{iPRODICT}$ of Tab. 1(b). Here FRESH\_PCAa again achieved the highest accuracy among all feature based approaches.

Further, we count how often a feature extraction method $a^* \in A$ reaches the highest accuracy for the 155 classifier/ data set combinations on the 31 UCR data sets among all six feature extraction methods in $A$ while a draw counts for both methods:

$$n_{\text{UCR}}^{\text{best}}(a^*) := \left| \left\{ (m,d_i) \mid m \in M, i = 1, \ldots, 31 : \text{acc}^d_m(a^*) = \max\{\text{acc}^d_m(a) \mid a \in A\} \right\} \right|.$$  

The evaluation metric $n_{\text{UCR}}^{\text{best}}$ is reported in the 5th column of Tab. 1(a). Again FRESH\_PCAa achieved the best result, for over half of the 155 inspected combinations it had the highest reported accuracy and again Boruta came in second with 74.3 combinations on average. Regarding both accuracy metrics, FRESH\_PCAa seems to be favorable over the other considered feature based approaches, with Boruta coming close.

4.3. Runtime

The second columns of both Tab. 1(a) and Tab. 1(b) contain the average pipeline runtime, which is the combined runtime of feature extraction, training of a classifier and predicting. For DTW\_NN the pipeline runtime captures just the fitting and predicting steps as no features are extracted. There are tradeoffs between the number of extracted features and the time spent on feature extraction, which force us to consider the pipeline runtime instead of the extraction runtime. E.g., time spent in the extraction process can be compensated by a lower number of extracted features which reduces the time spent for training the final classifier.

Analog to the accuracy evaluation metrics, $t^d_m(a)$ denotes the pipeline runtime in seconds of classifier $m \in M$ and feature extraction method $a \in A$ on data set $d$. In the same way, $t^d(a) = \frac{1}{5} \sum_{m \in M} t^d_m(a)$ denotes the average pipeline runtime in seconds over all 5 classifiers and $t^d(\text{DTW\_NN})$ the runtime of fitting and predicting by DTW\_NN. The average pipeline runtimes over all UCR data sets for
Table 1: Performance metrics for the different feature extraction methods and DTW_NN on the 31 two-class data sets from the UCR time series archive as well as the data from the iPRODICT research project. All simulations have been conducted three times and the values of the performance metrics have been averaged to reduce the dependency on random elements such as the test and train data split.

(a) Results on UCR data sets

| Method     | t_{UCR}    | acc_{UCR} | n_{UCR} | n_{best} |
|------------|------------|-----------|---------|----------|
| DTW_NN     | 152.1 ± 0.3| 0.926 ± 0.011| –       | –        |
| Full_X     | 33.4 ± 0.5 | 0.752 ± 0.018| 161.00 ± 0.00 | 39.0 ± 3.6 |
| FRESH      | 33.9 ± 0.5 | 0.838 ± 0.020| 40.35 ± 0.05 | 61.0 ± 5.0 |
| FRESH_PCAa | 33.9 ± 0.5 | 0.877 ± 0.010| 8.22 ± 0.00  | 79.0 ± 3.0 |
| FRESH_PCAb | 33.5 ± 0.5 | 0.760 ± 0.010| 3.93 ± 0.00  | 19.0 ± 4.4 |
| LDA        | 37.5 ± 0.6 | 0.678 ± 0.020| 27.23 ± 0.02 | 13.0 ± 1.7 |
| Boruta     | 104.5 ± 0.3| 0.841 ± 0.014| 63.29 ± 0.28 | 74.3 ± 6.7 |

(b) Results on the iPRODICT data

| Method     | t_{iPRODICT} | acc_{iPRODICT} | n_{iPRODICT} |
|------------|--------------|-----------------|--------------|
| DTW_NN     | 0.04 ± 0.016 | 0.552 ± 0.013   | –            |
| Full_X     | 3.30 ± 0.149 | 0.520 ± 0.003   | 3246 ± 0     |
| FRESH      | 3.31 ± 0.148 | 0.546 ± 0.007   | 309 ± 52     |
| FRESH_PCAa | 3.31 ± 0.146 | 0.566 ± 0.017   | 33 ± 4       |
| FRESH_PCAb | 4.01 ± 0.276 | 0.515 ± 0.023   | 2 ± 1        |
| LDA        | 474 ± 390    | 0.525 ± 0.010   | 229 ± 114    |
| Boruta     | 4.64 ± 0.241 | 0.562 ± 0.006   | 104 ± 3      |

Each method $a \in \mathcal{A} \cup \{\text{DTW}\_NN\}$ are denoted by $t_{UCR}^a = \frac{1}{31} \sum_{i=1, \ldots, 31} t_i^a$. All calculations are executed on a single computational core in order to increase comparability.

The full feature matrix Full_X is the fastest extraction algorithm in our comparisons on both the UCR and the iPRODICT data, as seen in the 2nd columns of Tab. 1(a) and Tab. 1(b). However, the standard deviation of the runtime clarifies that this result is not significant with respect to the runtime of the FRESH variants. The saved time for selecting the features is compensated for the five classifiers having to be trained on more features. Accordingly, the PCA step of FRESH_PCAa saves so much time for the fitting of the classifiers that this step is basically “free”, while FRESH_PCAa has a mean pipeline runtime of 33.86 seconds. On average, FRESH takes 33.87 seconds. The same can be observed on the iPRODICT data where Full_X, FRESH and FRESH_PCAa all had similar pipeline runtimes even though the number of extracted features varied greatly. The low average pipeline runtime of just $t_{iPRODICT} = 40.17$ seconds for DTW_NN in the 2th column of Tab. 1(b) is due to the classification algorithm only considering one type of time series while the extraction methods operate on 20 different time series.

Apart from the observed runtimes, we are interested in the feature extraction
method’s ability to scale with an increasing number of feature mappings, time series length and device numbers. As expected, Fig. 4.3 shows that all considered feature extraction methods – in contrast to DTW_NN – scale linearly with an increasing length of the time series or increasing number of samples. This is due to the considered feature mapping having a linear runtime with respect to the length of the time series. However, Fig. 4.3 shows that, among the feature based approaches, only FRESH and FRESH_PCAa scale linear with an increasing number of features (e.g. due to more devices, feature mappings or types of time series).

4.4. Selected features

The number of features extracted by algorithm $a \in \mathcal{A}$ on the data set $d$ are denoted by $n_f^d(a)$. Again this can be averaged over the UCR data sets resulting in $n_f^{UCR}(a) = \frac{1}{31} \sum_{i=1,..,31} n_f^d(a)$, cf. rightmost column of Tab. 1(a).
During our simulations, \texttt{FRESH\_PCAa} was able to reduce the number of features drastically. For the UCR data, it reduces 161 considered feature mappings to an average number of 8.2 features while on the iPRODICT data it reduces the 3246 calculated features to 33. \texttt{FRESH\_PCAb} only selected four respective two features on average, which may explain its low accuracies. If one wishes to extract a minimal set of relevant features, we recommend to deploy \texttt{FRESH\_PCAa}, as it extracted the second lowest number of features but achieved the highest and second highest accuracies.

4.5. Resume

We proposed \texttt{FRESH} as a highly scalable feature extraction algorithm. Our simulations showed that, in contrast to other considered methods, \texttt{FRESH} is able to scale with the number of feature mappings and samples as well as with the amount of different types and length of the time series. While doing so, it is extracting meaningful features as demonstrated by competitive accuracies.

The relative bad performance of \texttt{FRESH\_PCAb} seems to originate in the PCA step selecting features only based on their ability to explain the variance in the input variables and not in their significance to predict the target variable. By this, relevant information for the classification or regression task can get lost.

On the other hand, the combination of \texttt{FRESH} with a subsequent PCA filtering to reduce the number of redundant and highly correlated features, denoted as \texttt{FRESH\_PCAA} was overall the most competitive feature based method in our evaluation. On the UCR data, it achieved the second highest and on the iPRODICT data it reached the highest accuracy. Further, it had the second lowest number of extracted features.

5. Discussion

5.1. \texttt{FRESH} assists the acquisition of domain knowledge

It is common knowledge that the quality of feature engineering is a crucial success factor for supervised machine learning in general \cite[p. 82]{29} and for time series analysis in particular \cite{30}. But comprehensive domain knowledge is needed in order to perform high quality feature engineering. Contrarily, it is quite common for machine learning projects that data scientists start with limited domain knowledge and improve their process understanding while continuously discussing their models with domain experts. This is basically the reason, why dedicated time series models are very hard to build from scratch.

Our experience with data science projects in the context of IoT and Industry 4.0 applications \cite{8} showed that it is very important to identify relevant time series features in an early stage of the project in order to engineer more specialized features in discussions with domain experts. The \texttt{FRESH} algorithm supports this approach by applying a huge variety of established time series feature mappings to different types of time series and meta-information simultaneously and identifies relevant features in a robust manner.
We observe that features extracted by FRESH contribute to a deeper understanding of the investigated problem, because each feature is intrinsically related to a distinct property of the investigated system and its dynamics. This fosters the interpretation of the extracted features by domain experts and allows for the engineering of more complex, domain specific features [8] including dedicated time series models, such that their predictions in return might become a future feature mapping for FRESH.

5.2. FRESH is operational

We have already mentioned that FRESH has been developed in the course of IoT and Industry 4.0 projects [8]. Especially for predictive maintenance applications with limited numbers of samples and high level of noise in e.g. sensor readings, it has been proven as crucial to filter irrelevant features in order to prevent overfitting. To ensure a robust and scalable filtering, we consider each feature importance individually. This causes several implications:

• FRESH is robust in the sense of classical statistics, because the hypothesis tests and the Benjamini-Yekutieli procedure do not make any assumptions about the probability distribution or dependence structure between the features. Here, robustness refers to the insensitivity of the estimator to outliers or violations in underlying assumptions [31].

• FRESH is not considering the meaningfulness of interactions between features by design. Hence, in its discussed form it will not find meaningful feature combinations such as chessboard variables [32] Fig. 3a]. However, in our evaluation process the feature selection algorithm Boruta, which considers feature interactions, was not able to beat the performance of FRESH. Further, it is possible for FRESH to incorporate combinations of features and pre-defined interactions as new features themselves.

• FRESH is scalable due to the parallelity of the feature calculation and hypothesis tests (see the two topmost tiers in Fig. 1) and can be trivially parallelized and even distributed over several computational units. In addition, the feature filter process has computational costs compared to feature calculation and significance testing. Therefore, FRESH scales linearly with the number of extracted features, length of the time series, and number of considered time series.

• A side effect of ensuring robustness and testing features individually is that FRESH tends to extract highly correlated features, which could result in poor classification performance. We propose to combine FRESH with a subsequent PCA, which has been discussed as FRESH_PCA in Sec. 3.5 and indeed improved the performance significantly.

5.3. Feature selection of FRESH

Nilsson et al. [33] proposed to divide feature selection into two flavors: The minimal optimal problem is finding a set consisting of all strongly relevant attributes and a subset of weakly relevant attributes such that all remaining
weakly relevant attributes contain only redundant information. The all-relevant problem is finding all strongly and weakly relevant attributes. The first problem is way harder than the second, even asymptotically intractable for strictly positive distributions [33]. Accordingly, FRESH solves the second, easier problem as we extract every relevant feature, even though it might be a duplicate or highly correlated to another relevant feature [12].

Yu and Liu [34] separated feature selection algorithms into two categories, the wrapper model and the filter model. While the selection of wrapper models is based on the performance of a learning algorithm on the selected set of features, filter models use general characteristics to derive a decision about which features to keep. Filter models are further divided into feature weighting algorithms and subset search algorithms, which evaluate the goodness of features individually or through subsets. According to this definition, the feature selection part of FRESH is a filter model, more precisely, a feature weighting algorithm.

FRESH contains a feature selection part on basis of hypothesis tests and the Benjamini-Yekutieli procedure, which of course can be used as a feature selection algorithm itself. But, due to its systematic incorporation of scalable time series feature mappings and the proposed decomposition in computing tiers (Fig. 1) it is especially applicable to the needs of mass time series feature extraction and is considered as a feature extraction algorithm.

For a given set of attributes, feature selection techniques decide which attributes to delete and which to keep. While doing so, a possible evaluation metric is the expected ratio of deleted relevant features to all deleted features. This rate is the False Deletion Rate (FDR) and is formally defined by

\[
\text{FDR} = \mathbb{E}\left[\frac{\text{number of relevant but deleted features}}{\text{number of all deleted features}}\right]
\]

In contrast, FRESH controls the FER as defined in Eq. (7). A control of the FDR by FRESH would need a testing of the following hypotheses

\[
\begin{align*}
H_{0}^{\phi} &= \{X_{\phi} \text{ and } Y \text{ are dependent}\}, \\
H_{1}^{\phi} &= \{X_{\phi} \text{ and } Y \text{ are independent}\}.
\end{align*}
\]

Due to the topology of the hypotheses in Eq. (8) such a hypothesis testing is statistically not feasible. This means that FRESH can not be adapted to control the FDR. Hence, the feature filtering of FRESH is not suitable for feature selection jobs where the FDR has to be controlled.

Finally, by applying a multiple testing algorithm, FRESH avoids the “look-elsewhere effect” [35] which is a statistically significant observation arising by chance due to the high number of tested hypotheses. This effect triggered a recent discussions about the use of p-values in scientific publications [36].

5.4. Related work

There are both structural and statistical approaches to extract patterns from time series. Many statistical approaches rely on structures that allow the usage
of genetic algorithms. They express the feature pattern for example as a tree. While doing so, they aim for the best pattern and the most explaining features by alternating and optimizing the used feature mappings. In contrast, FRESH extracts the best fitting of a fixed set of patterns.

As an example for a structured pattern extraction, Olszewski detects six morphology types: constant, straight, exponential, sinusoidal, triangular, and rectangular phases. Those phases are detected by structure detectors which then output a new time series whose values stand for the identified structure. Based on this structure a domain-independent structural pattern recognition system is utilized to substitute the original time series signal by a known pattern. Due to its fixed patterns, FRESH can be considered to be a structured pattern extractor.

Of course, there are other promising approaches like the combination of nearest neighbor search with Dynamic Time Warping, which is specialized on considering an ensemble of exactly one dedicated time series type and cannot take meta-information into account. For binary classifications it scales with $O(n^2 \cdot m_{train} \cdot m_{test})$ with $m_{train}$ and $m_{test}$ being the number of devices in the train and test set, respectively. This approach also has the disadvantage that all data have to be transmitted to a central computing instance.

The extraction algorithm most similar to ours is presented by Fulcher and Jones. It applies a linear estimator with greedy search and a constant initial model to identify the most important features, which has been considered in this paper as LDA. The evaluation has shown, that FRESH outperforms the approach of Fulcher and Jones. Also, FRESH provides a more general approach to time series feature extraction, because it is able to extract features for regression tasks and not only for classification.

6. Summary and future work

In this work, FeatuRe Extraction based on Scalable Hypothesis tests (FRESH) for time series classification and regression is introduced. It combines well established feature extraction methods with a scalable feature selection based on non-parametric hypothesis tests and the Benjamini-Yekutieli procedure.

FRESH is highly parallel and suitable for distributed IoT and Industry 4.0 applications like predictive maintenance or process line optimization, because it allows to consider several different time series types per label and additionally takes meta-information into account. The latter has been demonstrated on basis of a steel billets process line optimization of project iPRODICT.

Our evaluation for UCR time series classification tasks has shown that FRESH in combination with a subsequent PCA outperforms all other feature extraction algorithms with respect to scalability and achieved accuracy. On the iPRODICT data set, it was even able to achieve a higher accuracy than a nearest neighbor search under Dynamic Time Warping.

The parallel nature of FRESH with respect to both feature extraction and filtering makes it highly applicable in situations where data is fragmented over
a widespread infrastructure and computations cannot be performed on centralized infrastructure. Due to its robustness and applicability to machine learning problems in the context of IoT and Industry 4.0, we are expecting that FRESH will find widespread application.

FRESH has been developed to extract meaningful features for classification and regression tasks. The considered data sets in this work only contained binary classification variables. We plan to close this gap by investigating the performance of FRESH for regression variables on suitable data sets in the future. Also, we are planning to investigate other hypothesis tests such as the Mann-Whitney-U-test to calculate the p-values.

Acknowledgement

The authors would like to thank Frank Kienle and Ulrich Kerzel for fruitful discussions as well as Nils Braun and Julius Neuffer for their most valuable contributions during the implementation of the tsfresh package. This research was funded in part by the German Federal Ministry of Education and Research under grant number 01IS14004 (project iPRODICT).

References

[1] J. Gubbi, R. Buyya, S. Marusic, M. Palaniswami, Internet of Things (IoT): A Vision, Architectural Elements, and Future Directions, Future Generation Computer Systems 29 (7) (2013) 1645–1660.

[2] M. Hermann, T. Pentek, B. Otto, Design Principles for Industrie 4.0 Scenarios, in: 49th Hawaii International Conference on System Sciences (HICSS), 3928–3937, 2016.

[3] R. K. Mobley, An Introduction to Predictive Maintenance, Elsevier Inc., 2 edn., 2002.

[4] V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos, Feature Selection for High-Dimensional Data, Artificial Intelligence: Foundations, Theory, and Algorithms, Springer International Publishing, 2015.

[5] H. Liu, R. Setiono, Some Issues on Scalable Feature Selection, Expert Systems with Applications 15 (3–4) (1998) 333 – 339.

[6] A. Kusiak, W. Li, The prediction and diagnosis of wind turbine faults, Renewable Energy 36 (1) (2011) 16–23.

[7] B. D. Fulcher, N. S. Jones, Highly Comparative Feature-Based Time-Series Classification, Knowledge and Data Engineering, IEEE Transactions on 26 (12) (2014) 3026–3037.
[8] M. Christ, F. Kienle, A. W. Kempa-Liehr, Time Series Analysis in Industrial Applications, in: Workshop on Extreme Value and Time Series Analysis, KIT Karlsruhe, doi:10.13140/RG.2.1.3130.7922, 2016.

[9] Y. Benjamini, D. Yekutieli, The Control of the False Discovery Rate in Multiple Testing under Dependency, Annals of statistics (2001) 1165–1188.

[10] Y. Chen, E. Keogh, B. Hu, N. Begum, A. Bagnall, A. Mueen, G. Batista, The UCR Time Series Classification Archive, www.cs.ucr.edu/~eamonn/time_series_data/, 2015.

[11] A. W. Liehr, Dissipative Solitons in Reaction Diffusion Systems, vol. 70 of Springer Series in Synergetics, Springer, 2013.

[12] M. B. Kursa, W. R. Rudnicki, The all relevant feature selection using random forest, arXiv preprint arXiv:1106.5112.

[13] X. Wang, A. Mueen, H. Ding, G. Trajcevski, P. Scheuermann, E. Keogh, Experimental Comparison of Representation Methods and Distance Measures for Time Series Data, Data Mining and Knowledge Discovery 26 (2) (2013) 275–309.

[14] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Pas- sos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.

[15] S. Van Der Walt, S. C. Colbert, G. Varoquaux, The NumPy Array: A Structure for Efficient Numerical Computation, Computing in Science & Engineering 13 (2) (2011) 22–30.

[16] W. McKinney, Data Structures for Statistical Computing in Python, in: S. van der Walt, J. Millman (Eds.), Proceedings of the 9th Python in Science Conference, 51 – 56, 2010.

[17] R. Elmasri, J. Y. Lee, Implementation Options for Time-Series Data, in: O. Etzion, S. Jajodia, S. Sripada (Eds.), Temporal Databases: Research and Practice, Springer Berlin Heidelberg, 115–128, 1998.

[18] I. Nun, P. Protopapas, B. Sim, M. Zhu, R. Dave, N. Castro, K. Pichara, FATS: Feature Analysis for Time Series, arXiv preprint arXiv:1506.00010.

[19] P. Radivojac, Z. Obradovic, A. K. Dunker, S. Vucetic, Feature Selection Filters Based on the Permutation Test, in: Machine Learning: ECML 2004, Springer, 334–346, 2004.

[20] R. A. Fisher, On the Interpretation of $\chi^2$ from Contingency Tables, and the Calculation of P, Journal of the Royal Statistical Society 85 (1) (1922-01) 87.
[21] F. J. Massey, The Kolmogorov-Smirnov Test for Goodness of Fit, Journal of the American Statistical Association 46 (253) (1951-03) 68.

[22] M. G. Kendall, A New Measure of Rank Correlation, Biometrika 30 (1/2) (1938-06) 81–93.

[23] L. M. Adler, A Modification of Kendall’s Tau for the Case of Arbitrary Ties in Both Rankings, Journal of the American Statistical Association 52 (277) (1957-03) 33–35.

[24] D. Curran-Everett, Multiple Comparisons: Philosophies and Illustrations, American Journal of Physiology - Regulatory, Integrative and Comparative Physiology 279 (1) (2000-07-01) R1–R8.

[25] Y. Benjamini, Y. Hochberg, Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing, Journal of the Royal Statistical Society. Series B (Methodological) (1995) 289–300.

[26] K. Kira, L. A. Rendell, A Practical Approach to Feature Selection, in: Proceedings of the Ninth International Workshop on Machine Learning, ML92, Morgan Kaufmann Publishers Inc., 249–256, 1992.

[27] H. Hotelling, Analysis of a Complex of Statistical Variables into Principal Components., Journal of educational psychology 24 (6) (1933) 417.

[28] X. Wang, K. Smith, R. Hyndman, Characteristic-Based Clustering for Time Series Data, Data Mining and Knowledge Discovery 13 (3) (2006-09-25) 335–364.

[29] P. Domingos, A Few Useful Things to Know about Machine Learning, Communications of the ACM 55 (10) (2012-10-01) 78–87.

[30] J. Timmer, C. Gantert, G. Deuschl, J. Honerkamp, Characteristics of Hand Tremor Time Series, Biological Cybernetics 70 (1) (1993) 75–80.

[31] J. M. M. John, F. Van Lishout, E. S. Gusareva, K. Van Steen, A Robustness Study of Parametric and Non-Parametric Tests in Model-Based Multifactor Dimensionality Reduction for Epistasis Detection, BioData mining 6 (1) (2013) 1.

[32] I. Guyon, A. Elisseeff, An Introduction to Variable and Feature Selection, The Journal of Machine Learning Research 3 (2003) 1157–1182.

[33] R. Nilsson, J. M. Peña, J. Björkegren, J. Tegnér, Consistent feature selection for pattern recognition in polynomial time, The Journal of Machine Learning Research 8 (2007) 589–612.

[34] L. Yu, H. Liu, Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution, in: ICML, vol. 3, 856–863, 2003.
[35] E. Gross, O. Vitells, Trial Factors for the Look Elsewhere Effect in High Energy Physics, The European Physical Journal C 70 (1-2) (2010) 525–530.

[36] R. L. Wasserstein, N. A. Lazar, The ASA’s Statement on P-Values: Context, Process, and Purpose, The American Statistician.

[37] I. Mierswa, K. Morik, Automatic Feature Extraction for Classifying Audio Data, Machine learning 58 (2-3) (2005) 127–149.

[38] P. Geurts, Pattern Extraction for Time Series Classification, in: Principles of Data Mining and Knowledge Discovery, Springer, 115–127, 2001.

[39] D. R. Eads, D. Hill, S. Davis, S. J. Perkins, J. Ma, R. B. Porter, J. P. Theiler, Genetic Algorithms and Support Vector Machines for Time Series Classification, in: B. Bosacchi, D. B. Fogel, J. C. Bezdek (Eds.), International Symposium on Optical Science and Technology, 74–85, 2002-12-01.

[40] R. T. Olszewski, Generalized Feature Extraction for Structural Pattern Recognition in Time-Series Data, DTIC Document CMU-CS-01-108, Carnegie Mellon University, 2001.

[41] L. Penserini, others, A Comparison of Two Machine-Learning Techniques to Focus the Diagnosis Task, Frontiers in Artificial Intelligence and Applications (2006) 265.

[42] D. Joanes, C. Gill, Comparing measures of sample skewness and kurtosis, Journal of the Royal Statistical Society: Series D (The Statistician) 47 (1) (1998) 183–189.

[43] E. W. Weisstein, CRC concise encyclopedia of mathematics, CRC press, 2002.

[44] W. A. Fuller, Introduction to statistical time series, vol. 428, John Wiley & Sons, 2009.

[45] J. G. MacKinnon, Approximate asymptotic distribution functions for unit-root and cointegration tests, Journal of Business & Economic Statistics 12 (2) (1994) 167–176.

[46] R. H. Jones, Maximum likelihood fitting of ARMA models to time series with missing observations, Technometrics 22 (3) (1980) 389–395.

[47] P. Du, W. A. Kibbe, S. M. Lin, Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching, Bioinformatics 22 (17) (2006) 2059–2065.

[48] N. Ricker, The form and laws of propagation of seismic wavelets, Geophysics 18 (1) (1953) 10–40.
[49] P. D. Welch, The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms, IEEE Transactions on audio and electroacoustics 15 (2) (1967) 70–73.

Vitae

Maximilian Christ received his M.S. degree in Mathematics and Statistics from Heinrich-Heine-Universität Düsseldorf, Germany, in 2014. In his daily work as a Data Science Consultant at Blue Yonder GmbH he optimizes business processes through data driven decisions. Beside his business work he is pursuing a Ph.D in collaboration with University of Kaiserslautern, Germany. His research interest relate on how to deliver business value through Machine Learning based optimizations.

Dr. Andreas W. Kempa-Liehr is a Senior Lecturer at the Department of Engineering Science of the University of Auckland, New Zealand, and an Associate Member of the Freiburg Materials Research Center (FMF) at the University of Freiburg, Germany. Andreas received his doctorate from the University of Münster in 2004 and continued his research as head of service group Scientific Information Processing at FMF. From 2009 to 2016 he was working in different data science roles at EnBW Energie Baden-Württemberg AG and as Senior Data Scientist at Blue Yonder GmbH. His research is focussed on the synergy of data science and operations research for engineering applications.

Prof. Dr. Michael Feindt founded Blue Yonder GmbH, one of the world’s leading companies for Predictive Applications in the retail sector, in 2008 and serves as its Chief Scientific Advisor and Member of Advisory Board. He was a research fellow and staff member at the European Organization for Nuclear Research CERN from 1991 to 1997 and he also participated in the CDFII experiment at Fermilab in Chicago/USA as well as Belle and Belle II experiments at KEK in Japan. He is a professor at the Institute of Experimental Nuclear Physics at the Karlsruhe Institute of Technology (KIT), Germany. His main research interest lies in optimizing and automating research in many different areas by means of machine learning, especially combining Bayesian statistics, neural networks, conditional probability density estimation and causality.
Appendix A. Considered feature mappings

FRESH and the respective tsfresh package depend on feature mappings $\theta_k : \mathbb{R}^{n_t} \to \mathbb{R}$ for capturing relevant characteristics from the respective time series. Each of these features mappings $\theta(S)$ take a single time series

$$S = (s(t_1), s(t_2), \ldots, s(t_{n_t}), \ldots, s(t_{n_t}))^T = (s_1, \ldots, s_{n_t}, \ldots, s_{n_t})^T.$$  

as argument and return either a real valued feature (e.g. length), or a vector of features (e.g. binned_entropy). In the latter case, each element of the returned feature vector will be treated as separate feature. For the purpose of simplicity, the following descriptions are omitting the indices for different devices and sensors described in Eq. (1). Feature mappings with additional parameters are denoted by $\theta(S|\cdot)$.

Note, that the number of feature mappings provided by the tsfresh package is continuously increasing. However, the following subsections describe the feature mappings, which have been used for the evaluation in Section 4. The most recent list of feature mappings can be found at [http://tsfresh.readthedocs.io/en/latest/](http://tsfresh.readthedocs.io/en/latest/)

Appendix A.1. Features from summary statistics

- **maximum($S$):** Sample maximum of time series $S$
  
  $$\text{maximum}(S) = \max\{s_1, \ldots, s_{n_t}, \ldots, s_{n_t}\}.$$  

- **minimum($S$):** Sample minimum of time series $S$
  
  $$\text{minimum}(S) = \min\{s_1, \ldots, s_{n_t}, \ldots, s_{n_t}\}.$$  

- **mean($S$):** Arithmetic mean of time series $S$
  
  $$\text{mean}(S) = \bar{S} = \frac{1}{n_t} \sum_{\nu=1}^{n_t} s_{\nu}.$$  

- **var($S$):** Expectation of the squared deviation of time series $S$ from its mean $\bar{S}$ without bias correction
  
  $$\text{var}(S) = \frac{1}{n_t} \sum_{\nu=1}^{n_t} (s_{\nu} - \bar{S})^2.$$  

- **std($S$):** Uncorrected sample standard deviation of time series $S$
  
  $$\text{std}(S) = \sqrt{\text{var}(S)}$$
• **skewness**\((S)\): Sample skewness calculated with adjusted Fisher-Pearson standardized moment coefficient \[42\]:

\[
\text{skewness}(S) = \frac{n_t^2}{(n_t - 1)(n_t - 2)} \left( \frac{1}{n_t - 1} \sum_{\nu=1}^{n_t} (s_{\nu} - \bar{S})^3 \right)^{\frac{1}{2}}.
\]

• **kurtosis**\((S)\): Fourth central moment of time series \(S\) divided by the square of its variance

\[
\text{kurtosis}(S) = \frac{1}{n_t} \sum_{\nu=1}^{n_t} \left( \frac{s_{\nu} - \bar{S}}{\text{std}(S)} \right)^4 - 3
\]

as defined by Fisher \[43\]. The subtrahend of 3 ensures that the kurtosis is zero for normal distributed samples.

• **length**\((S)\): Number of samples \(n_t\) of time series \(S\)

\[
\text{length}(S) = n_t.
\]

• **median**\((S)\): For a time series \(S\) with an uneven number of samples \(n_t\), the median is the middle of the sorted time series values. If the time series has an even number of samples, the two middle values are averaged

\[
\text{median}(S) = \begin{cases} 
\frac{s(n_t+1)/2}{2} & : n_t \text{ is uneven,} \\
\frac{1}{2}(s(n_t/2) + s(n_t/2+1)) & : n_t \text{ is even.}
\end{cases}
\]

• **quantile_of_empiric_distribution_function**\((S|q)\): \(q\)-quantile \(Q_q(S)\) of empirical distribution function \(\hat{S}_{n_t}\) of time series sample \(S\)

\[
Q_q(S) = \inf \left\{ z \mid \hat{S}_{n_t}(z) = \frac{1}{n_t} \sum_{\nu=1}^{n_t} 1_{s_{\nu} \leq z} \geq q \right\}.
\]

Here, \(q\)% of the ordered values from \(S\) are lower or equal to this quantile.

**Appendix A.2. Additional characteristics of sample distribution**

• **absolute_energy**\((S)\): Interpreting the time series \(S\) as the velocity of a particle with unit mass 2, the observed energy computes to

\[
\text{absolute_energy}(S) = \sum_{\nu=1}^{n_t} s_{\nu}^2.
\]

• **augmented_dickey_fuller_test_statistic**\((S)\): The Augmented Dickey-Fuller test checks the hypothesis that a unit root is present in a time series sample \[44\]. This feature calculator returns the value of the respective test statistic calculated on \(S\) \[45\].

23
• **binned_entropy**(\(S|m\)): This feature calculator bins the values of the time series sample into \(m\) equidistant bins. The returned features is

\[
\min(m,n_t) \sum_{k=0}^{p_k \log(p_k) \cdot 1(p_k > 0)}
\]

where \(p_k\) is the percentage of samples in bin \(k\).

• **has_large_standard_deviation**(\(S\)): Boolean feature indicating that the standard deviation \(\text{std}(S)\) is bigger than half the difference between the maximal and minimal value

\[
\text{std}(S) > \frac{\text{maximum}(S) - \text{minimum}(S)}{2}.
\]

• **has_variance_larger_than_std**(\(S\)): Boolean feature indicating that the variance is greater than the standard deviation

\[
\text{var}(S) > \text{std}(S),
\]

which is equal to the variance of the sample being larger than 1.

• **is_symmetric_looking**(\(S\)): Boolean feature indicating

\[
|\text{mean}(S) - \text{median}(S)| < \frac{\text{maximum}(S) - \text{minimum}(S)}{2}.
\]

• **mass_quantile**(\(S|q\)): Relative index \(\nu\) where \(q \cdot 100\%\) of the mass of the time series \(S\) lie left of \(\nu\):

\[
\frac{i}{n_t} \text{ such that } i = \min \left\{ k \mid 1 \leq k \leq n_t, \frac{\sum_{\nu=1}^{k} s_{\nu}}{S} \geq q \right\}.
\]

For example for \(q = 50\%\) this feature calculator will return the mass center of the time series.

• **number_data_points_above_mean**(\(S\)): Number of data points, which are larger than the average value of the time series sample.

• **number_data_points_above_median**(\(S\)): Number of data points, which are larger than the median value of the time series sample.

• **number_data_points_below_mean**(\(S\)): Number of data points, which are lower than the average value of the time series sample.

• **number_data_points_below_median**(\(S\)): Number of data points, which are lower than the median value of the time series sample.
Appendix A.3. Features derived from observed dynamics

- **arima_model_coefficients**\((S|i, k)\): This feature calculator fits the unconditional maximum likelihood of an autoregressive AR\((k)\) process \[46\]

\[
s_\nu = \varphi_0 + \sum_{j=1}^{k} \varphi_j s_{\nu-j} + \varepsilon_t
\]

on the time series \(S\). The parameter \(k\) is the maximum lag of the process and the calculated feature is the coefficient \(\varphi_i\) for index \(i\) of the fitted model.

- **continuous_wavelet_transformation_coefficients**\((S|a, b)\): Calculates a discretization of the continuous wavelet transformation \[47\] for the Ricker wavelet, also known as the *Mexican hat wavelet* \[48\] which is defined by

\[
\psi(\nu, a, b) = \frac{2}{\sqrt{3a\pi^4}} \left(1 - \frac{\nu - b}{a^2}\right) \exp\left(-\frac{(\nu - b)^2}{2a^2}\right).
\]

Here, \(a\) is the width and \(b\) is the location parameter. The continuous wavelet transformation performs a convolution of the time series \(S\) with the wavelet \(\psi(\nu, a, b)\)

\[
X_w(a, b) = \sum_{\nu=-\infty}^{\infty} s_\nu \overline{\psi}(\nu, a, b).
\]

The features are the coefficients \(X(a, b)\) for the wavelet of length \(a\) at position \(b\).

- **fast_fourier_transformation_coefficient**\((S|k)\): Calculates the one-dimensional discrete Fourier Transform for real input. The real part of the coefficients

\[
X_k = \sum_{\nu=1}^{n_t} s_\nu \cdot \exp\left(-\frac{2\pi ik(\nu - 1)}{n_t}\right), \quad k \in \mathbb{Z}
\]

are returned as features. For real valued input the real part of \(X_k\) is equal to the real part of \(X_{-k}\) which means that this feature calculator only take natural numbers as parameters.

- **first_index_max**\((S)\): This feature is the relative position

\[
\min \left(\arg\max_{n_t} S\right)
\]

for which the maximum value was observed for the first time in the time series sample.
• first_index_min$(S)$: This feature is the relative position
\[ \min \left( \arg \min_{n_t} S \right) \]
for which the minimal value was observed for the first time in the time series sample.

• lagged_autocorrelation$(S|l)$: Calculates the autocorrelation of the time series $S$ with its lagged version of lag $l$:
\[ \frac{1}{\text{var}(S)} \sum_{\nu=1}^{n_t-l} (s_{\nu} - \bar{S}) (s_{\nu+l} - \bar{S}) . \]

• large_number_of_peaks$(S|l, m)$: Boolean variable indicating wether the number of peaks of size $l$ as defined in number_peaks_of_size$(l)$ is greater than $m$.

• last_index_max$(S)$: Relative position
\[ \max \left( \arg \max_{n_t} S \right) \]
at which the maximum value was observed for the last time in the time series sample.

• last_index_min$(S)$: Relative position
\[ \max \left( \arg \min_{n_t} S \right) \]
at which the maximum value was observed for the last time in the time series sample.

• longest_strike_above_mean$(S)$: Length of the longest consecutive subsequence in $S$, which is larger or equal to the mean of $S$.

• longest_strike_above_median$(S)$: Length of the longest consecutive subsequence in $S$, which is larger or equal to the median of $S$.

• longest_strike_below_mean$(S)$: Length of the longest consecutive subsequence in $S$, which is smaller or equal to the mean of $S$.

• longest_strike_below_median$(S)$: Length of the longest consecutive subsequence in $S$, which is smaller or equal to the median of $S$.

• longest_strike_negative$(S)$: Length of the longest consecutive subsequence in $S$ that is smaller or equal to the median of $S$. 

26
• **longest_strike_zero**\((S)\): Length of the longest consecutive subsequence in \(S\) that is smaller or equal to the median of \(S\).

• **mean_absolute_change**\((S)\): Arithmetic mean of absolute differences between subsequent time series values

\[
\frac{1}{n_t} \sum_{\nu=1}^{n_t-1} |s_{\nu+1} - s_\nu|.
\]

• **mean_absolute_change_quantiles**\((S|q_{l}, q_{h})\): This feature calculator first fixes a corridor given by the quantiles \(Q_{q_{l}}\) and \(Q_{q_{h}}\) of the empirical distribution function of \(S\). Then it calculates the average absolute value of consecutive changes of the series \(S\) inside this corridor:

\[
\frac{\sum_{\nu=1}^{n_t-1} |s_{\nu+1} - s_\nu| 1_{Q_{q_{l}} \leq s_\nu \leq Q_{q_{h}}} 1_{Q_{q_{l}} \leq s_{\nu+1} \leq Q_{q_{h}}} }{\sum_{\nu=1}^{n_t-1} 1_{Q_{q_{l}} \leq s_\nu \leq Q_{q_{h}}} 1_{Q_{q_{l}} \leq s_{\nu+1} \leq Q_{q_{h}}} }.
\]

• **mean_autocorrelation**\((S)\): Average autocorrelation over possible lags \(l\) ranging from 1 to \(n_t - 1\):

\[
\frac{1}{(n_t - 1) \operatorname{var}(S)} \sum_{l=1}^{n_t} \sum_{\nu=1}^{n_t-l} (s_\nu - \bar{S})(s_{\nu+l} - \bar{S}).
\]

• **mean_second_derivate_central**\((S)\): Average value of the second derivate of the time series

\[
\frac{1}{n_t - 2} \sum_{\nu=2}^{n_t-2} \frac{1}{2} (s_{\nu-1} - 2 \cdot s_\nu + s_{\nu+1}).
\]

• **number_continuous_wavelet_transformation_peaks_of_size**\((S|l)\): This feature calculator detects peaks in the time series by inspecting the coefficients of a discretization of the continuous wavelet transformation of the time series \(S\). First, it calculates the wavelet transformation \(X_w(a, b)\) for widths \(a\) ranging from 1 to \(l\). Again the ricker wavelet is used. The peaks are then found by comparing the local signal-to-noise ratio (SNR), if a local maxima of the wavelet coefficients \(X_w(a, b)\) is observed for enough widths, it is counted as a peak \[47\].

• **number_peaks_of_size**\((S|l)\): A peak of support \(l\) is defined as a data point \(s_i\), which is larger than its \(l\) neighboring values to the left and to the right. This feature calculates the number of peaks with support of \(l\) in the time series \(S\).

• **spektral_welch_density**\((S|l)\): This feature calculator uses Welch’s method \[49\] to compute an estimate of the power spectral density of the time series sample \(S\). Welch’s methods divides the time series into overlapping
segments, computes a modified periodogram for each segment and then averages the periodograms. The frequencies $f$ are defined as follow

$$f = \left(0, \frac{1}{n_t}, \frac{2}{n_t}, \ldots, \frac{1}{2} - \frac{1}{n_t}, -\frac{1}{2} + \frac{1}{n_t}, \ldots, -\frac{2}{n_t}, -\frac{1}{n_t}\right)$$ if $n_t$ is even

$$f = \left(0, \frac{1}{n_t}, \frac{2}{n_t}, \ldots, \frac{n_t - 1}{2n_t}, -\frac{n_t - 1}{2n_t}, -\frac{1}{n_t} + \frac{1}{n_t}, \ldots, -\frac{2}{n_t}, -\frac{1}{n_t}\right)$$ if $n_t$ is odd

This feature calculator will return the calculated power spectrum of frequency $f_i$ as the feature.

• **time_reversal_asymmetry_statistic($S|l$):** Value of

$$\frac{1}{n_t - 2l} \sum_{\nu=1}^{n_t-2l} s_{\nu+2l}^2 \cdot s_{\nu+l} - s_{\nu+l} \cdot s_{\nu}^2$$

for lag $l$. It was proposed in [7] as promising feature to extract from time series.