Package: oddwater (via r-universe)

November 21, 2024

Type Package

Title Outlier Detection in Data from Water-Quality Sensors

Version 0.7.0

Depends R (>= 3.4.0)

Maintainer Priyanga Dilini Talagala <pritalagala@gmail.com>

Description We propose a framework to detect technical outliers in water quality data from in situ sensors.

License GPL-3

Encoding UTF-8

LazyData true

RoxygenNote 6.1.1

Imports lubridate, ggplot2, tidyr, scales, GGally, tsibble, shiny, dbscan

Suggests magrittr

Config/pak/sysreqs make libicu-dev libssl-dev zlib1g-dev

Repository https://robjhyndman.r-universe.dev

RemoteUrl https://github.com/pridiltal/oddwater

RemoteRef HEAD

RemoteSha 51831c6a1d014e67d409c1733f592c28fff311b7

Contents

`calc_MSE` *Compute mean squared error*

Description

Compute mean squared error

Usage

```
calc_MSE(y_truth, y_pred)
```

Arguments

`y_truth`	A numeric vector containing the gorund truth
`y_pred`	A numeric vector containing the estimated values

Author(s)

Priyanga Dilini Talagala

`calc_performance_metrics`
Compute performance metrics

Description

Computes various measure to evaluate the performance of an algorithm

Usage

```
calc_performance_metrics(y_truth, y_output, pos_label, neg_label,
  print_out = TRUE)
```

Arguments

`y_truth`	A character vector containing the gorund truth
`y_output`	a character vector containing the predicted labels from the algorithm
`pos_label`	A character string. Label used to indicate the outliers in the original dataframe.
`neg_label`	A character string. Label used to indicate the typical values in the original dataframe.
`print_out`	If TRUE, output will be printed to console.

Value

A list with the following elements:

TN	True negatives
FN	False negatives
FP	False positives
TP	True positives
Accuracy	Accuracy
Error_Rate	Error Rate
Sensitivity	Sensitivity
Specificity	Specificity
Precision	Precision
Recall	Recall
F_Measure	F Measure
Optimised_Precision	Optimised Precision
PPV	Positive Predictive Value
NPV	Negative Predictive Value

Author(s)

Priyanga Dilini Talagala

Examples

```
true_labels <- c("out", "out", "normal", "out", "normal", "normal",
                 "normal", "normal", "normal", "normal")
output <- c("out", "normal", "normal", "normal", "out", "out",
            "normal", "normal", "normal", "normal")
out<- calc_performance_metrics(y_truth = true_labels, y_output = output,
                               pos_label = "out", neg_label = "normal")
```

---

data_pioneer_anom *Water Quality Sensor data - Pioneer*

Description

A multivariate dataset containing the variables obtained using water quality sensors from Pioneer. The characteristics of the different types of anomalies are presented in detail in Leigh, et al. (2019). The anomaly types can be further grouped into three general classes. Class 1 included anomalies described by a sudden change in value from the previous observation (types A, D, I, and J). Class 2 included those anomaly types that should be detectable by simple, hard-coded classification rules, such as measurements outside the detectable range of the sensor (types F, G and K). Class 3 anomalies may require user intervention post hoc (i.e. after data collection rather than in real time) to confirm observations as anomalous or otherwise in combination with automated detection (types B, C, E, H and L).

Usage

```
data_pioneer_anom
```

Format

A data frame with 6303 rows and 10 variables:

Timestamp Time Stamps

Level Level

Cond Conductivity

Tur Turbidity

label_Level Whether individual data points are anomalous or not in the level series. 1 - outlier, 0 - typical

label_Cond Whether individual data points are anomalous or not in the conductivity series. 1 - outlier, 0 - typical

label_Tur Whether individual data points are anomalous or not in the tubidity series. 1 - outlier, 0 - typical

type_Level Type of the anomaly in the level series. A - sudden large spikes, B - low variability including persistent values, C- constant offsets, D - sudden shifts, E - high variability, F - impossible values, G - out-of-sensor-range values, H - drift, I- clusters of spikes, J - sudden small spikes, K - missing values and L - other untrustworthy (not described by types A-K)

type_Cond Type of the anomaly in the conductivity series. A - sudden large spikes, B - low variability including persistent values, C- constant offsets, D - sudden shifts, E - high variability, F - impossible values, G - out-of-sensor-range values, H - drift, I- clusters of spikes, J - sudden small spikes, K - missing values and L - other untrustworthy (not described by types A-K)

type_Tur Type of the anomaly in the Turbidity series. A - sudden large spikes, B - low variability including persistent values, C- constant offsets, D - sudden shifts, E - high variability, F - impossible values, G - out-of-sensor-range values, H - drift, I- clusters of spikes, J - sudden small spikes, K - missing values and L - other untrustworthy (not described by types A-K)

References

Leigh, C, O Alsibai, RJ Hyndman, S Kandanaarachchi, OC King, JM McGree, C Neelamraju, J Strauss, PD Talagala, RD Turner, K Mengersen & EE Peterson (2019). A framework for automated anomaly detection in high frequency water-quality data from in situ sensors. Science of the Total Environment 664, 885–898.

`data_sandy_anom` *Water Quality Sensor data - Sandy_Creek*

Description

A multivariate dataset containing the variables obtained using water quality sensors from Sandy Creek. The characteristics of the different types of anomalies are presented in detail in Leigh, et al. (2019). The anomaly types can be further grouped into three general classes. Class 1 included anomalies described by a sudden change in value from the previous observation (types A, D, I, and J). Class 2 included those anomaly types that should be detectable by simple, hard-coded classification rules, such as measurements outside the detectable range of the sensor (types F, G and K). Class 3 anomalies may require user intervention post hoc (i.e. after data collection rather than in real time) to confirm observations as anomalous or otherwise in combination with automated detection (types B, C, E, H and L).

Usage

```
data_sandy_anom
```

Format

A data frame with 5402 rows and 10 variables:

Timestamp Time Stamps

Level Level

Cond Conductivity

Tur Turbidity

label_Level Whether individual data points are anomalous or not in the level series. 1 - outlier, 0 - typical

label_Cond Whether individual data points are anomalous or not in the conductivity series. 1 - outlier, 0 - typical

label_Tur Whether individual data points are anomalous or not in the tubidity series. 1 - outlier, 0 - typical

type_Level Type of the anomaly in the level series. A - sudden large spikes, B - low variability including persistent values, C- constant offsets, D - sudden shifts, E - high variability, F - impossible values, G - out-of-sensor-range values, H - drift, I- clusters of spikes, J - sudden small spikes, K - missing values and L - other untrustworthy (not described by types A-K)

type_Cond Type of the anomaly in the conductivity series. A - sudden large spikes, B - low variability including persistent values, C- constant offsets, D - sudden shifts, E - high variability, F - impossible values, G - out-of-sensor-range values, H - drift, I- clusters of spikes, J - sudden small spikes, K - missing values and L - other untrustworthy (not described by types A-K)

type_Tur Type of the anomaly in the Turbidity series. A - sudden large spikes, B - low variability including persistent values, C- constant offsets, D - sudden shifts, E - high variability, F - impossible values, G - out-of-sensor-range values, H - drift, I- clusters of spikes, J - sudden small spikes, K - missing values and L - other untrustworthy (not described by types A-K)

References

Leigh, C, O Alsibai, RJ Hyndman, S Kandanaarachchi, OC King, JM McGree, C Neelamraju, J Strauss, PD Talagala, RD Turner, K Mengersen & EE Peterson (2019). A framework for automated

anomaly detection in high frequency water-quality data from in situ sensors. Science of the Total Environment 664, 885–898.

`explore_data` *Run Shiny Applications*

Description

Launch Shiny application

Usage

```
explore_data()
```

`NN_HD` *NN-HD algorithm*

Description

This algorithm is inspired by the HDoutliers algorithm which is an unsupervised outlier detection algorithm that searches for outliers in high dimensional data assuming there is a large distance between outliers and the typical data. Nearest neighbor distances between points are used to detect outliers. However, variables with large variance can bring disproportional influence on Euclidean distance calculation. Therefore, the columns of the data sets are first normalized such that the data are bounded by the unit hyper-cube. The nearest neighbor distances are then calculated for each observation. In contrast to the implementation of HDoutliers algorithm available in the `HDoutliers` package, NN_HD now generates outlier scores instead of labels for each observation.

Usage

```
NN_HD(dataset)
```

Arguments

`dataset`	A multivariate dataset containing numerical variables

Author(s)

Priyanga Dilini Talagala

References

Wilkinson, L. (2016). Visualizing outliers. https://www.cs.uic.edu/~wilkinson/Publications/outliers.pdf

oddwater *oddwater: A package for Outlier Detection in water quality sensor data*

Description

oddwater: A package for Outlier Detection in water quality sensor data

Author(s)

Priyanga Dilini Talagala, Rob J. Hyndman

See Also

The core functions in this package: `transform_data`, `plot_series`, `plot_pairs`, `calc_performance_metrics`, `calc_MSE`

plot_pairs *Make a matrix of plots with a given data set*

Description

provide a matrix plot and mark anomalous point in red colour and neighbouring points in green colour

Usage

```
plot_pairs(data)
```

Arguments

`data`	A dataframe. "Timestamp" column give the timestamps and "type" column gives types of the data points (outlier, neighbour, typical).

Value

A graphical representation of the matrix plot

Author(s)

Priyanga Dilini Talagala

plot_series *Plot Multivariate time series*

Description

Plot multivariate time series and mark anomalous point in red colour and neighbouring points in green colour

Usage

```
plot_series(data, title)
```

Arguments

data	A dataframe. "Timestamp" column give the timestamps and "type" column gives types of the data points (outlier, neighbour, typical).
title	A character string. This sis the main title of the plot

Value

A graphical representation of the multivariate time series

Author(s)

Priyanga Dilini Talagala

transform_data *Apply different transformations to the original series*

Description

This function apply different transformations to the original variables. This data preprocessing step was incorporated with the aim of highlighting different types of anomalies such as sudden isolated spikes, sudden isolated drops, sudden shifts, impossible values (negative values) and out of range values etc

Usage

```
transform_data(data, time_bound = 90, regular = FALSE,
  time_col = "Timestamp")
```

Arguments

data	A dataframe. This dataframe contains a seperate column for Timestamp, in addition to the variables that need to be transformed
time_bound	A positive constant. This is to reduce the effect coming from too small time gaps when calculating derivatives.
regular	Regular time interval (TRUE) or irregular (FALSE)
time_col	A quoted string to specify the column name of the timestamp

Value

A tsibble object with the original and the transformed series

Author(s)

Priyanga Dilini Talagala

Examples

```
data <- data_sandy_anom[,c("Timestamp", "Cond", "Tur", "Level")]
data <- tidyr::drop_na(data)
trans_data <- oddwater::transform_data(data)
```

Index