Continuous Imputation of Missing Values in Streams of Pattern-Determining Time Series

KEVIN WELLENZOHN† · MICHAEL H. BÖHLEN†
ANTON DIGNÖS† · JOHANN GAMPER† · HANNES MITTERRER‡

Problem

Problem. Streaming time series often have missing values, e.g. due to sensor failures or transmission delays!

Goal. Accurately impute (i.e. recover) the latest measurement by exploiting the correlation among streams.

Challenge. Streaming time series are often non-linearly correlated, e.g. due to phase shifts.

Approach (TKCM)

Intuition. Impute a missing value in time series \(s\) with past values from \(s\) when a set of correlated reference time series exhibited similar patterns.

Top-k Case Matching (TKCM). To impute a missing value in a time series \(s\) at time \(t_n\):

1. Define query pattern \(P(t_n)\), spanning the values of \(d\) reference time series over a time frame of \(l\) time points anchored at time \(t_n\).
2. Look for the \(k\) most similar non-overlapping patterns in a sliding window over the time series.
3. Impute the missing value \(s(t_n)\) as the average of the values of \(s\) at the anchor time points of the \(k\) previously found patterns

Parameter \(l\) (called the “pattern length”) enables TKCM to deal with non-linearly correlated time series, e.g. phase-shifted time series.

Phase Shifts

- Time series \(s\) and \(r\) are phase-shifted by 90 degrees.
- The scatterplot shows that \(s\) and \(r\) are non-linearly correlated.
- Their Pearson Correlation Coefficient is 0!
- For example, whenever \(r(t) = 0.5\), time series \(s\) has two different values, either \(s(t) = 0.86\) or \(s(t) = -0.86\)

TKCM & Non-Linear Correlations

With pattern length \(l > 1\),

- TKCM takes the temporal context into account and captures how time series change over time.
- There are less patterns with pattern dissimilarity 0

Application Scenario

Consider the set \(\{s, r_1, r_2\}\) of streaming time series obtained from a sensor network. Time series \(s\) has a missing value at current time \(t_n = 14:20\) that is imputed using the \(d = 2\) reference time series \(r_1\) and \(r_2\).

Time series \(s\) imputed by TKCM

Experiments

- The Chlorine dataset contains phase-shifted and hence non-linearly correlated time series
- A larger pattern length decreases the oscillation in the imputed time series
- TKCM is more accurate than its competitors for non-linearly correlated time series