Knowledge Discovery from Heterogeneous Dynamic Systems using Change-Point Correlations

IBM Research, Tokyo Research Lab
Tsuyoshi Ide

井手 剛
Data streams in the real world are full of intractable features. How can we discover knowledge on system’s mechanism?

**Example:**

- Time series data taken from various sensors in a car

**heterogeneity**

**sudden & unpredictable changes**  
**strong correlation**

Direct comparison between time-series leads to meaningless results due to the heterogeneity.
Tackling the heterogeneity using a nonlinear transformation — “Correlate by features, not by values”

original recordings

apparently different

synchronized features

potential relationship

change-point score
Singular spectrum transformation (SST) — transforms an original time series into a time-series of the change-point scores using SVD

1. For each of reference and test intervals, construct a matrix using subsequences as column vectors
2. Perform SVD on those matrices to find representative patterns
3. Compute dissimilarity between the patterns
Surprisingly, SST removes the heterogeneity with a single parameter set. No ad hoc parameter tuning is needed.

- recordings taken from a car
- quite heterogeneous
- comparable to each other
- existing clustering techniques are applicable
SST is robust over a very wide range of parameters
Dependence on the pattern length, $w$

change-point score for $5 < w < 38$, $l=3$

robust over a very wide range of $w$
Application to an automobile data set. SST effectively detects change-points irrespective of the heterogeneous behavior

**Data**
- powertrain control module
- sampling rate = 0.1 sec

**Variables**
- x1: fuel flow rate
- x2: engaged gear
- x3: vehicle speed
- x4: engine RPM
- x5: manifold absolute pressure

- raw data
- CP score
  (w = 25; 25 data points)
Application to an automobile data set.

SST is useful for discovering hidden structures among variables

- $x_2$ and $x_4$ behave quite differently in the original data
- However, after the SST, they form the nearest pair in the MDS plot.
  - Multidimensional scaling
    method for retrieving the coordinates from a distance matrix
  - This result can be naturally understood by the mechanism of the car

- engaged gear
- engine RPM
Backup
Looking for a change-point detection method applicable to heterogeneous systems *without any ad hoc parameter tuning*

- **Existing methods**
  - CUSUM (cumulated sum)
    - Suitable only for such data that distributes around a constant.
  - Autoregressive models
    - Cannot be used for streams with sudden & unpredictable changes.
  - Wavelet transforms
    - Essentially the same as differentiation. Applicable to a very limited class of variables. Otherwise, fine parameter tuning for individual variable is needed.
  - Gaussian mixture models
    - Cannot be used for streams with sudden & unpredictable changes.

A new “nonparametric” approach

Singular Spectrum Transformation

Note: The original inspiration for this approach was based on an aspect of [Moskvina-Zhigljavsky 2003]
SST needs only two parameters in a standard setting: the pattern length and the number of patterns.
Computing the distances between variables and visualize them using multidimensional scaling

- **Normalization policies and distance metrics**

<table>
<thead>
<tr>
<th>Normalization</th>
<th>original time series</th>
<th>change-point score</th>
</tr>
</thead>
<tbody>
<tr>
<td>normalization</td>
<td>( \int dt \ x_i(t)^2 = 1 )</td>
<td>( \int dt \ x_i(t) = 1 )</td>
</tr>
<tr>
<td>distance between xi &amp; xj</td>
<td>( \sqrt{\int dt \ (x_i - x_j)^2} )</td>
<td>( \int dt \</td>
</tr>
</tbody>
</table>

* SST series are nonnegative by definition, so they are naturally interpreted as the probability densities of changes

- **Multidimensional scaling**
  - retrieves the coordinates from a distance matrix
  - eigenvalue analysis shows \( d = 2 \) is sufficient
Summary and future work

- **Summary**
  - We proposed a new framework of data mining for heterogeneous dynamic systems
  - The SST is a robust and effective way to remove the heterogeneity
  - An experiment demonstrated the utility of SST in discovering hidden structures

- **Future work**
  - Speed up and sophisticate SST
  - Develop methodologies for causality analysis