Business Intelligence
WS 2014/15

Classification Methods for Temporal Data

W. Grossmann

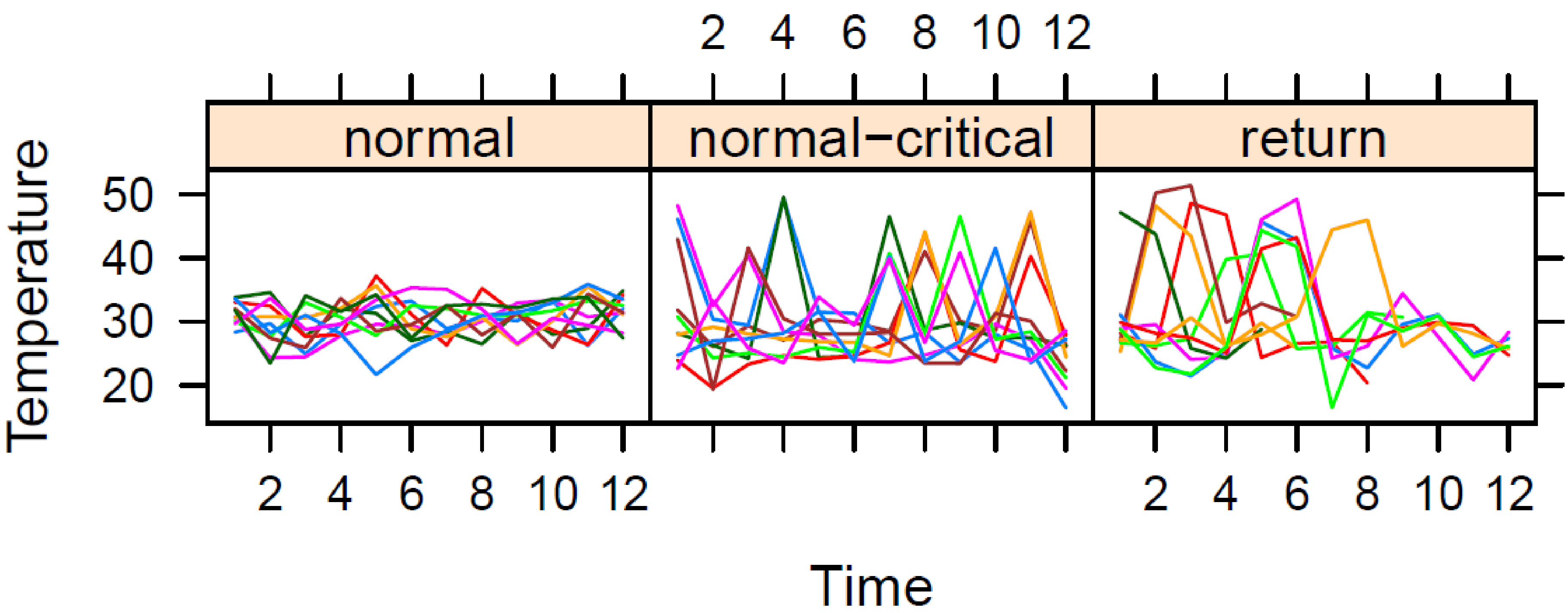
Content

- Problem Formulation
- Classification Based on Similarity Measures
- Classification Based on Response Features
- Clustering of Time Sequenece

Problem Formulation

- Example: Temperature monitoring of a container during transport
  - In case of “abnormal” temperature behavior the cargo is damaged and the transport has to be interrupted and returns to the origin
  - We distinguish three different scenarios as shown in the graphic

Problem Formulation



Problem Formulation

- General Problem formulation:
- Given are data of customer behavior represented as time sequences for process instances
- These data are classified into different groups
- Task: Find a classification rule which allows the assignment of a time sequence to one of the classes

Problem Formulation

- Strategies for solving this task:
  - Strategy 1: Define a similarity measure for the time sequences and apply nearest neighbor classification
  - Strategy 2: Extract from the time sequences features which allow the application of the classification methods for cross-sectional data
- In general, the first strategy is recommended if no additional knowledge about the time sequence is known

Classification Based on Similarity Measures

- Problem with the definition of similarity:
  - Time sequences may have different length
  - Similarity may be blurred by some temporal transformations like stretching or squeezing some parts of the time sequence (see example)
- We have to define the similarity by matching the observed values of two time sequences in such a way that the above defined effects are compensated

Classification Based on Similarity Measures

- Dynamic time warping allows the calculation of similarity
- Basic is the definition of a warping path:

Given two sequences $(x_1, x_2, \ldots, x_N)$ and $(y_1, y_2, \ldots, y_M)$

Define a sequence $(p_1, p_2, \ldots, p_L)$ of matching indices pairs (i_ℓ, j_ℓ) such that

$$p_1 = (1,1) \quad p_L = (N, M)$$

$$(i_1 \le i_2 \le \ldots \le i_L) \text{ and } (j_1 \le j_2 \le \ldots \le j_L)$$

$$p_{\ell+1} - p_\ell \in \{(1,0), (0,1), (1,1)\}$$

Classification Based on Similarity Measures

The costs of a warping path is defined by

$$D_P = \sum_{\ell=1}^{L} d(i_\ell, j_\ell) = \sum_{\ell=1}^{L} | x_{i_\ell} - y_{j_\ell} |$$

- The dynamic time warping algorithm finds a warping path for two sequences with minimal costs
  - The word "dynamic" indicates that the algorithm is based on dynamic programming

Classification Based on Similarity Measures

- Application of the dynamic warping algorithm for all pairs of sequences defines a distance matrix for the observed time sequences
- We can apply now k-nearest neighbor classification for obtaining the classification rule
- Application for the example with 1-nearest neighbor is shown on CEWebS in *Klassifikation_NearestNeighbor*

Classification Based on Response Features

- In that case we extract from the time sequence a number of time independent characteristic features
- Examples of features:
  - Maximum and minimum of the time sequence
  - Temporal location of maximum and minimum
  - Breakpoints in the time sequence
  - Largest difference between two sequenced values
  - Length of the sequence
  - Area under the polygon defined by the sequence

Classification Based on Response Features

- More theoretically motivated features:
  - Transformation to frequencies and looking at the maximum frequency (Time sequence is sound or light)
  - Definition of a regression model for the time sequence
  - Definition of a representation language
- Based on these attributes one can apply methods of the classification of cross sectional data

Clustering of Time Sequences

- Clustering of time sequences can be done using the same principles as in the case of classification
- The definition of time warping defines a distance for the sequences which cam be used as input for cluster analysis (hierarchical or k-means)
- In the case of response features the distance between the time sequences is based on the response features