Data Mining, Lecture 11: Mining Time Series

S. Nõmm

[1] Department of Computer Science, Tallinn University of Technology

April 18, 2016

Introduction

- Preparation
- Forecasting
- Motifs
- Time series to Sequences Data Mining
- Periodic Patterns
- Clustering
- Outlier Detection
- Classification

Preparation

- Handling missing, unequally spaced, or unsynchronized values. Linear interpolation: generates estimated values at the desired time stamps. Let y_i and y_j are two values of the time series at times t_i and t_j respectively, where $j > i$. Let t denote desired time stamp from the interval $(t_i, \dots, t_j)$. Then the interpolated value corresponding to the time t is as follows:

$$y = y_i + \left(\frac{t - t_i}{t_j - t_i}\right)(y_j - y_i).$$

Preparation

- Noise removal
  - Binning. Assumption: the timestamps are equally spaced apart. Divide the series into k equal intervals. The average value of the data points in each interval are reported as the smoothed values.
  - Moving-Average Smoothing.
  - Exponential Smoothing: smoothed value is defined as the linear combination of the current value of the time series, and the previously smoothed value.

$$y'_i = \alpha y_i + (1 - \alpha) y'_{i-1}.$$

Preparation

- Normalization.
    - Range-based normalization.
    - Standardization.
- Data Transformation and Reduction.
    - Discrete Wavelet Transform.
    - Discrete Fourier Transform.
    - Symbolic Aggregate Approximation (SAX).

Time Series Similarity Measures

- DTW
- Edit Distance
- Longest Common Subsequence

Time Series Forecasting

- Stationary and non stationary time series.

Definition

Time series is said to be stationary if the probabilistic distribution of the values in any time interval $[t_i, t_j]$ is identical to that in the shifted interval $[t_i + h, t_j + h]$ for any value of the time shift h.

- Differencing is the common approach used to convert time series into the stationary form.

$$y'_i = y_i - y_{i-1}$$

- Second order differencing:

$$y''_i = y_i - 2t_{i-1} + y_{i-2}.$$

- Seasonal differencing:

$$y_i = y_i - y_{i-m}$$

Time Series Forecasting

- Autoregressive Models: Univariate time series contain a single variable that may be predicted by means of autocorrelation. Autocorrelations: the correlations between adjacently located time stamps in the time series. The autocorrelations in a time series are defined with respect to a particular value of the lag L.

$$A(L) = \frac{C_t(y_t, y_{t+L})}{V_t(y_t)}$$

- Autoregressive model:

$$y_t = \sum_{i=1}^{p} \alpha_i y_{t_i} + c + \epsilon t$$

- The model can be used effectively for forecasting future values, only if the key properties of the time series, such as the mean, variance, and autocorrelation do not change significantly with time.

Time Series Forecasting

- One of the possible goodness parameters:

$$R^2 = 1 - \frac{\mu_t(\epsilon_t^2)}{V_t(y_t)}$$

Time Series Forecasting

- Autoregressive Moving Average Models. Autocorrelation does not always explain all the variations. The unexpected component of the variations (shocks) may be captured with the use of moving average.

$$y_t = \sum_{i=1}^{q} b_i \epsilon_{t-i} + c + \epsilon_t$$

- Autoregressive Moving Average Model:

$$y_t = \sum_{i=1}^{p} a_i y_{t-1} + \sum_{i=1}^{q} b_i \epsilon_{t-1} + c + \epsilon_t$$

- Multivariate Forecasting with Hidden Variables

Motifs

- A motif is a frequently occurring pattern or shape in the time series.
- Single series versus database of many series.
- Contiguous versus noncontiguous motifs.
- Multigranularity motifs.

When does a motif belong to a time series?

- Distance-based support.
- Transformation to sequential pattern mining.
- Periodic patterns.

Distance Based Motifs

- Distance-based motifs are always defined on contiguous segments of the time series.

- Definition

  A sequence (or motif) $S = s_1, \dots s_w$ of real values is said to approximately match a contiguous subsequence of length w in the time series $(y_1, \dots y_n)$ $(w \leq n)$ starting at position i, if the distance between $(s_1, \dots, s_w)$ and $(y_i, \dots y_{i+w1})$ is at most ϵ.

  - Euclidean distance is a common choice in this case.
  - Frequency of the motif:

  Definition

  The number of matches of a time series window $S = s_1 \dots s_w$ to the time series $(y_1 \dots y_n)$ at threshold level ϵ, is equal to the number of windows of length w in $(y_1 \dots y_n)$, for which the distance between the corresponding subsequences is at most ϵ.

Clustering

- Real-time clustering.
  - Online Clustering of Coevolving Series:based on determining correlations across the series, in online fashion.
- Database of time series is available. Shape-Based Clustering.
  - k-means.
  - k-medoids.
  - Hierarchical methods.
  - Graph based methods.x

Outlier detection

- Point outliers: A point outlier is a sudden change in a time series value at a given timestamp.
- Shape outliers: In this case, a consecutive pattern of data points in a contiguous window may be defined as an anomaly.

Point outliers

- Determine the forecasted values of the time series at each timestamp. Let the forecasted value of the of the rth timestamp be dentoed $\bar{W}_r$
- Compute the time series of deviations $\bar{\Delta}_1, \ldots, \bar{\Delta}_r$

$$\bar{\Delta}_r = \bar{W}_r - \bar{Y}_r.$$

- Let $\bar{\Delta}_r = \{\delta_r^1, \ldots \delta_r^d\}$. Let the mean and standard deviation of the i th series of deviations be denoted by μ_i and σi.
- Compute the normalized deviations:

$$\delta z_r^i = \frac{\delta_r^i - \mu_i}{\sigma_i}$$

- The unified alarm level U_r at timestamp r can be reported as:

$$U_r = \max_{i \in \{1,\ldots,d\} \sigma z_r^i}$$

Classification

- Point labels: In this case, the class labels are associated with individual timestamps. (Supervised event detection.)
- Whole-series labels: In this case, the class labels are associated with the full series. (Whole series classification.)
    - Wavelet-Based Rules.
    - Nearest Neighbor Classifier.
    - Graph-Based Methods.