Time Series Data Mining Challenges

Jose A. Lozano

Basque Center for Applied Mathematics (BCAM)
University of the Basque Country UPV/EHU

Time Series Days, Rennes, March 25-26, 2019

Outline of the presentation

1. Time Series Data Mining Activities
2. Clustering
3. (Early) Supervised Classification
4. Outlier/Anomaly Detection
5. Conclusions and Future Work

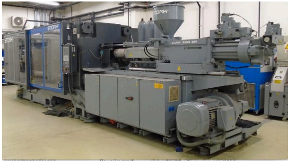
Outline of the presentation

1. Time Series Data Mining Activities
2. Clustering
3. (Early) Supervised Classification
4. Outlier/Anomaly Detection
5. Conclusions and Future Work

basque center for applied mathematics

UPV - EHU

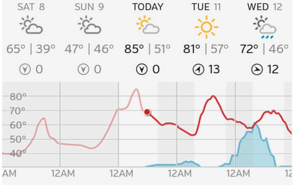
Time series all around



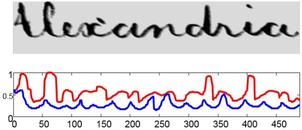
Industry 4.0

Bio Signals

- Temporal correlation
- High dimensionality
- Noisy

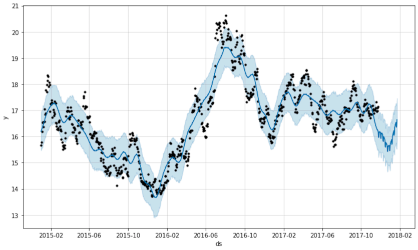


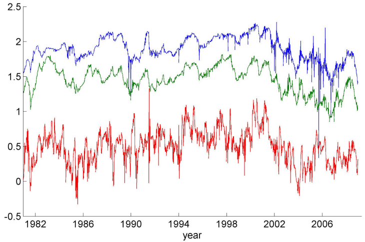
Weather Forecasting



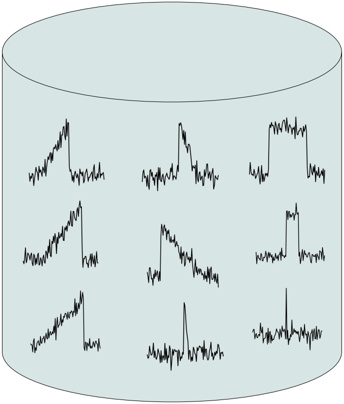
Shapes

Time series forecasting



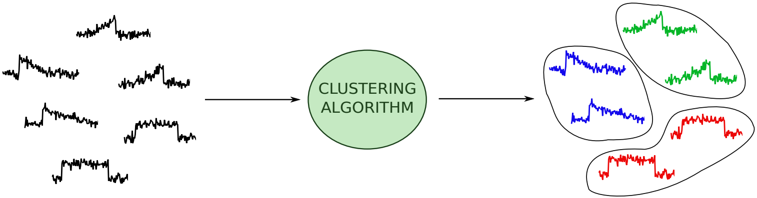


Time series data base: our object of study



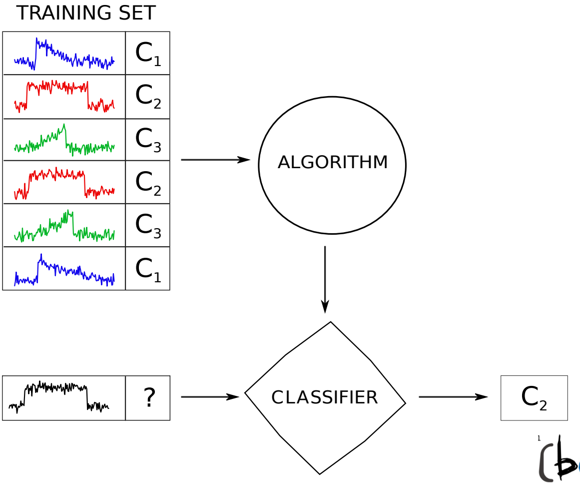
- A set of time series (usually big)
- Different lengths
- Multidimensional

Time series clustering. Examples

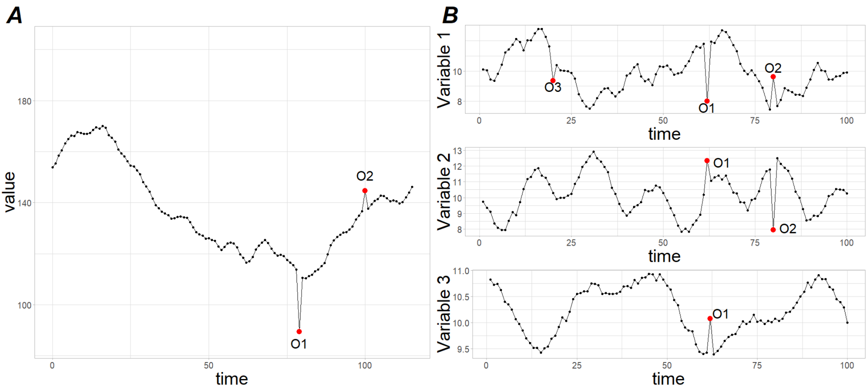


1

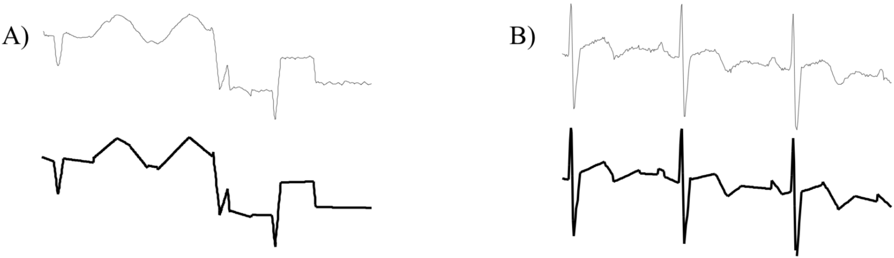
Supervised classification of time series



Anomaly/outlier detection



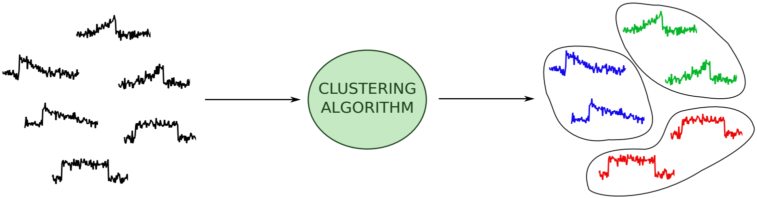
Segmentation



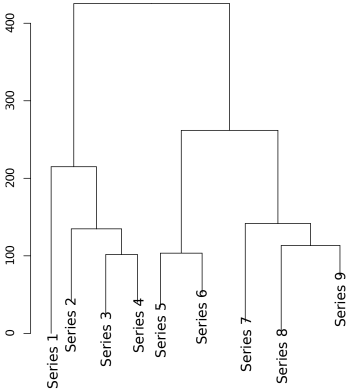
Outline of the presentation

1. Time Series Data Mining Activities
2. Clustering
3. (Early) Supervised Classification
4. Outlier/Anomaly Detection
5. Conclusions and Future Work

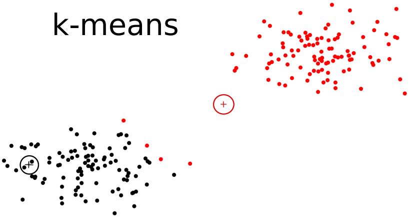
Time series clustering. Examples



Time series clustering: hierarchical, partitional



k-means

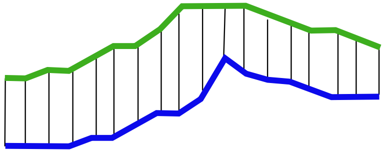


we need a

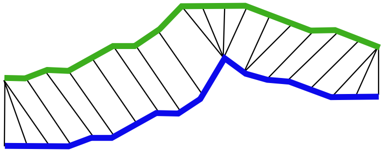
DISTANCE

Distance between time series

Rigid Distance

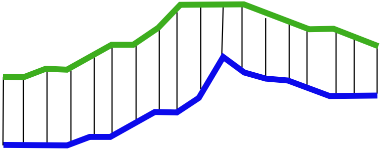


Flexible Distance



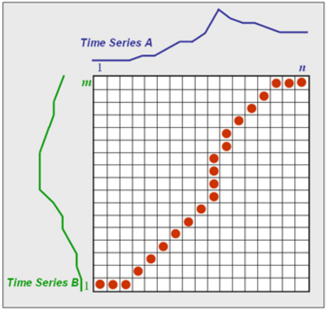
Euclidean Distance (ED)

$$D(X, Y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

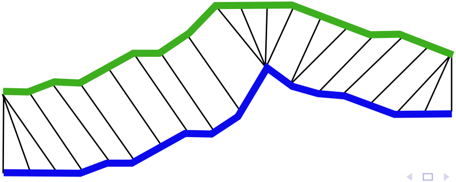


- Easy to compute ✓
- Only for series with the same distance ✓
- Does not consider the time ✓
- Sensitivity to noise ✓

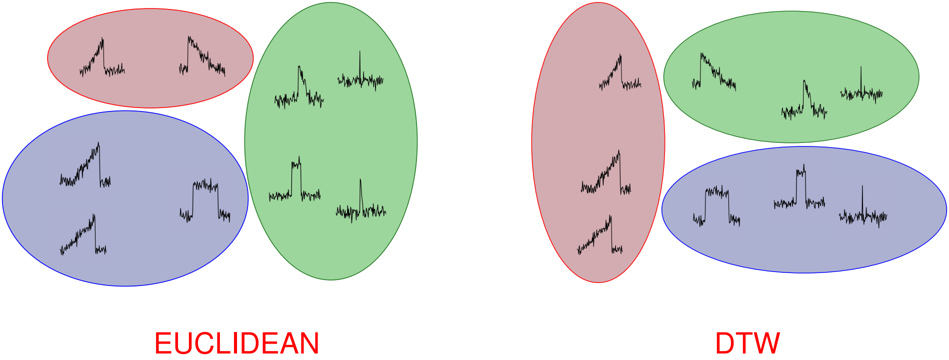
Dynamic Time Warping (DTW)



- Takes into account the ordered sequence (time) ✓
- It can deal with series of different sizes ✓
- Computationally expensive $O(min\{m,n\}^2)$ ✓



Euclidean Distance vs Dynamic Time Warping



Remark on distances

More on elastic distances

- Cheap versions of dynamic time warping (Sakoe-chiba, bounding)
- Edit distance for real sequences (EDR)
- Mori et al 2016, R journal
- On-line versions (Oregi et al 2019, PR)

Alternatives to calculate distances

- Represent each series by means of a set of features and calculate the distance between the features
- Learn a parametric model for each series and calculate the distance between the parameters

Distances between series

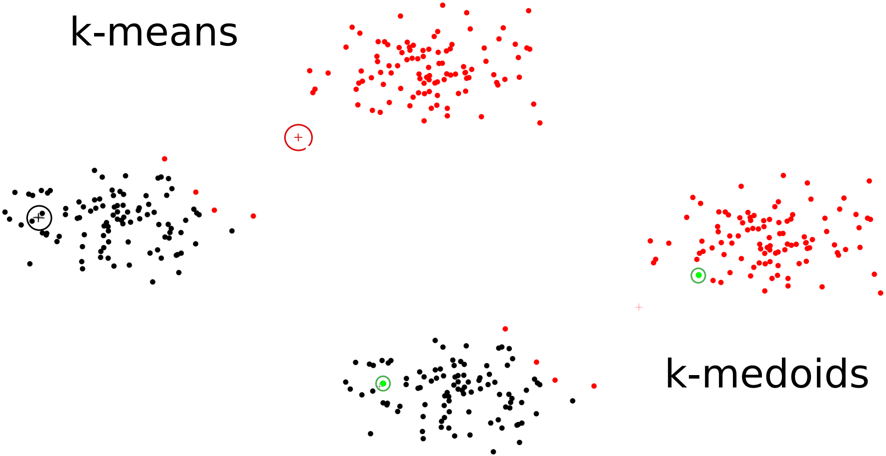
Remarks

- There is not best distance (no free lunch)
- Each problem requires a different distance
- The distance to be used needs to be in agreement with our knowledge about what is far and what is close
- Hint: try with several distances

Challenge:

Design a method to the (semi)automatic selection of a distance
(e.g. Mori et al. 2016, TKDE)

...Coming back to clustering: K-means



Remarks on clustering

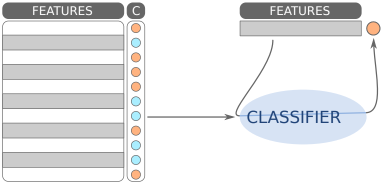
- Recent papers on the computation of a *mean* series
- Alternate clustering methods: graph-based, spectral, model-based,...
- Multivariate time series clustering

Outline of the presentation

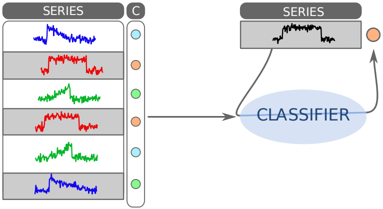
1. Time Series Data Mining Activities
2. Clustering
3. (Early) Supervised Classification
4. Outlier/Anomaly Detection
5. Conclusions and Future Work

Supervised Classification of Time Series

General-purpose classifiers



Specific TS classifiers



General-purpose classifiers

- Each series is considered an instance
- Each time stamp is considered a feature

t_1	t_2	t_3	$\dots$	t_n	C
x_{11}	x_{12}	x_{13}	$\dots$	x_{1n}	c_1
x_{21}	x_{22}	x_{23}	$\dots$	x_{2n}	c_2
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
x_{m1}	x_{m2}	x_{m3}	$\dots$	x_{mn}	c_2

General-purpose classifiers

- Each series is considered an instance
- Each time stamp is considered a feature

t_2	t_1	t_3	$\ldots$	t_n	C
x_{12}	x_{11}	x_{13}	$\ldots$	x_{1n}	c_1
x_{22}	x_{21}	x_{23}	$\ldots$	x_{2n}	c_2
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
x_{m2}	x_{m1}	x_{m3}	$\ldots$	x_{mn}	c_2

General-purpose classifiers

- Each series is considered an instance
- Each time stamp is considered a feature

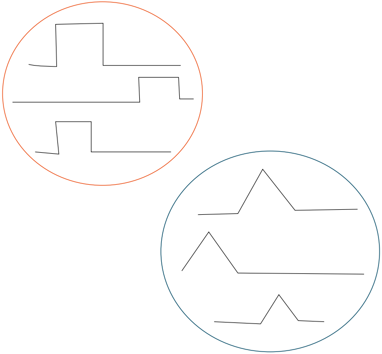
t_2	t_1	t_3	$\dots$	t_n	C
x_{12}	x_{11}	x_{13}	$\dots$	x_{1n}	c_1
x_{22}	x_{21}	x_{23}	$\dots$	x_{2n}	c_2
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
x_{m2}	x_{m1}	x_{m3}	$\dots$	x_{mn}	c_2

CHALLENGE

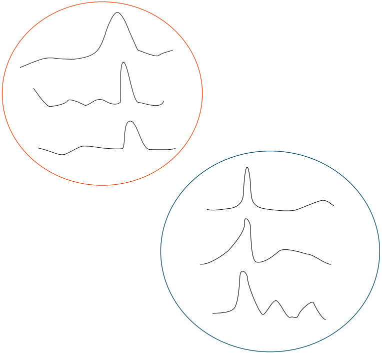
When to use general-purpose and when time-series specific?

What is relevant in TSC?

PROBLEM I

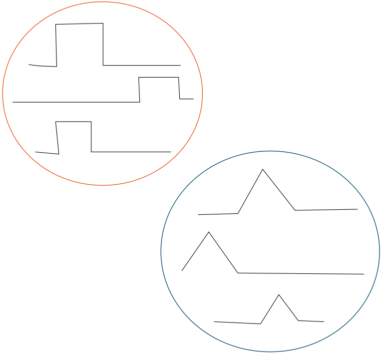


PROBLEM II

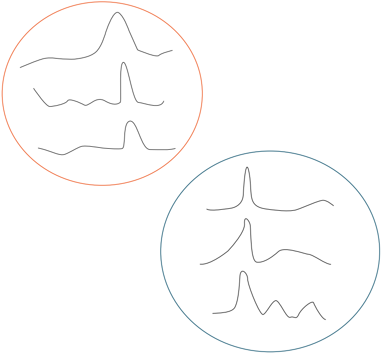


What is relevant in TSC?

PROBLEM I



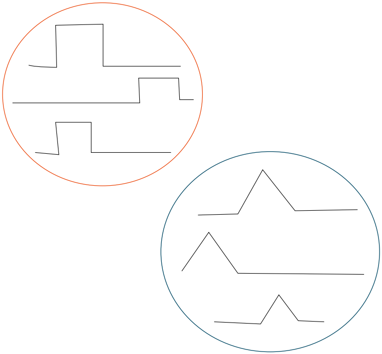
PROBLEM II



SHAPE

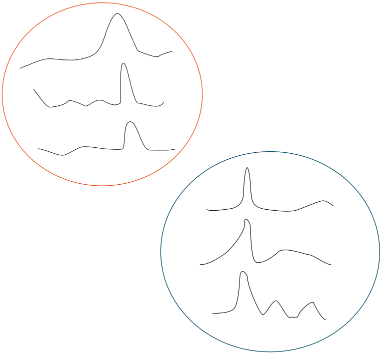
What is relevant in TSC?

PROBLEM I



SHAPE

PROBLEM II



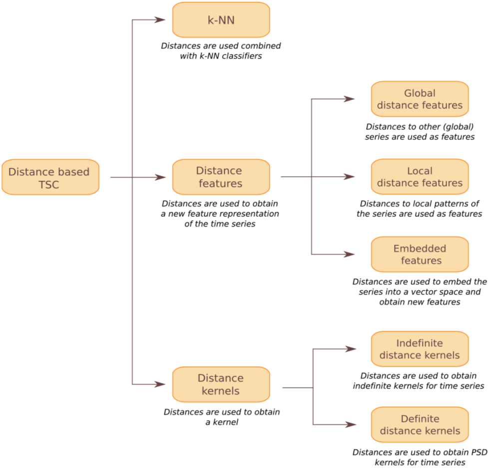
LOCATION

A taxonomy of time series classification methods

Taxonomy

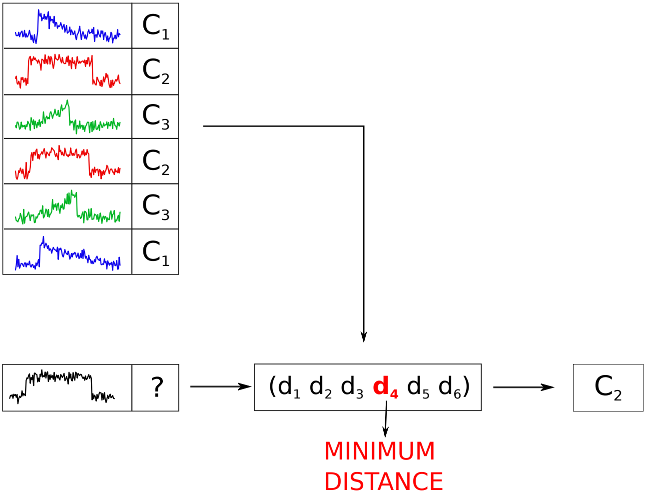
- Distance-based classifiers
- Model-based classifiers
- Feature-based classifiers

Taxonomy of distance-based TSC (Abanda et al. 2019, DAMI)



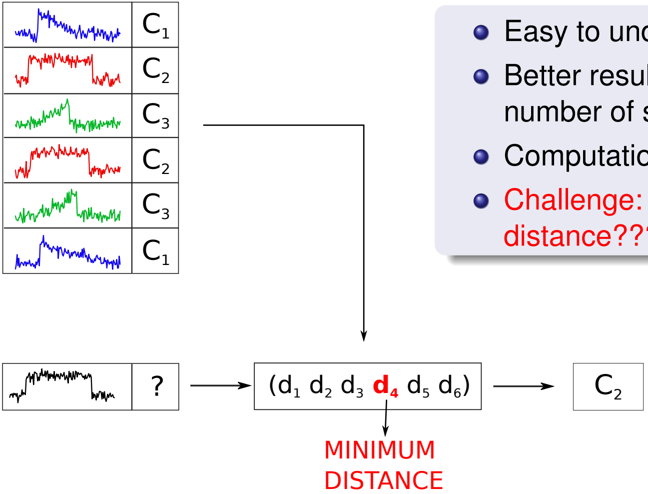
1

1-Nearest Neighbour (1-NN)

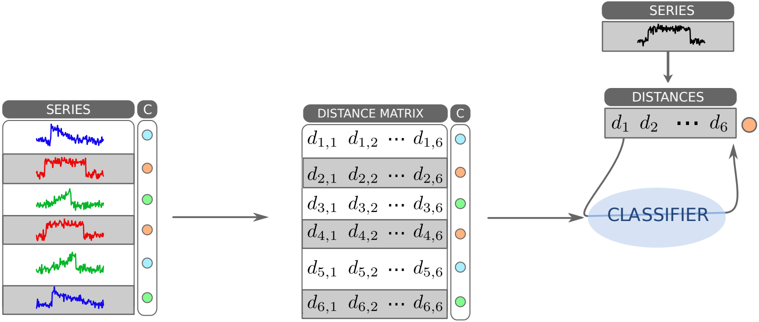


1-Nearest Neighbour (1-NN)

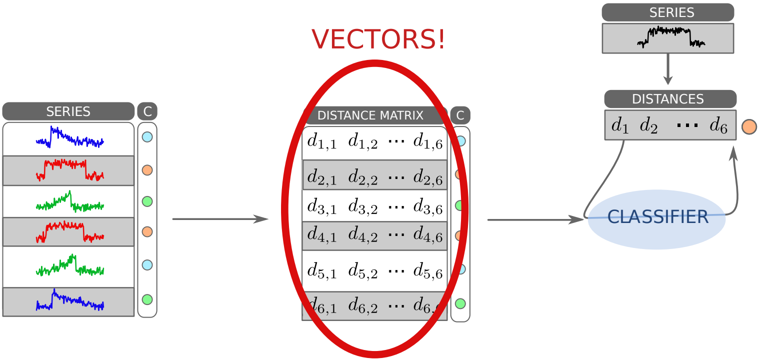
- Easy to understand ✓
- Better results with higher number of series ✓
- Computational cost ✓
- Challenge: What distance???



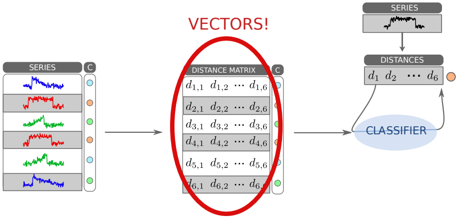
Distance-based. Distance features



Distance-based. Distance features. Global

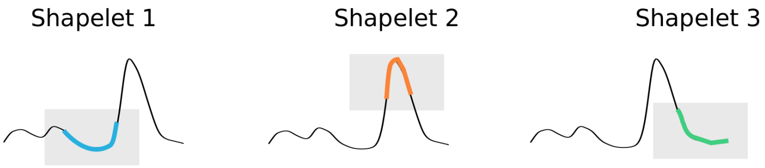


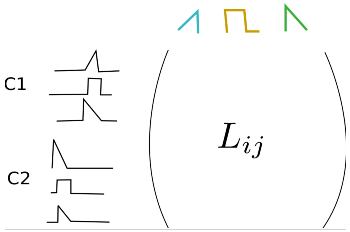
Distance-based. Distance features. Global



- Any general-purpose algorithm could be applied ✓
- It depends on the number of series in training ✓
- Computationally expensive ✓
- Difficult to transfer to the on-line setting ✓

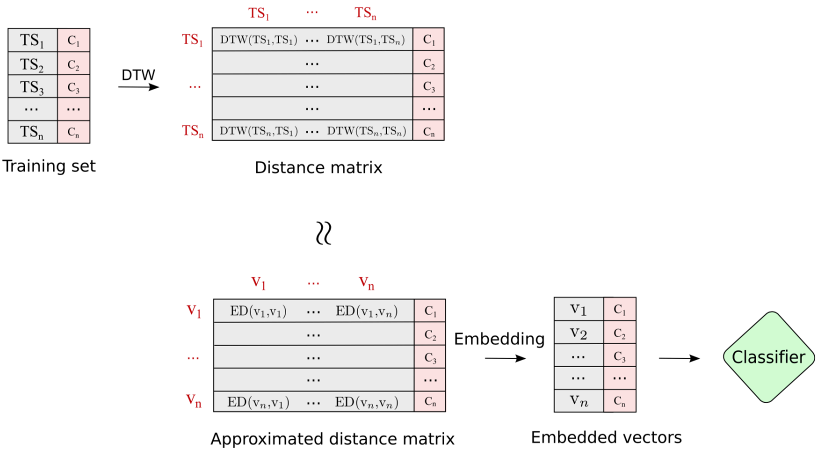
Distance-based. Distance features. Local





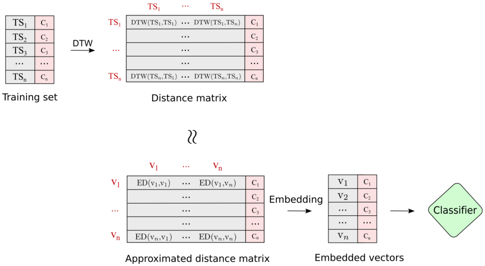
- L_{ij} could be distance or presence
- Computationally expensive ✗
- When the shapelets are relevant extremely good results ✓
- Easy to interpret ✓

Distance-based. Embedding



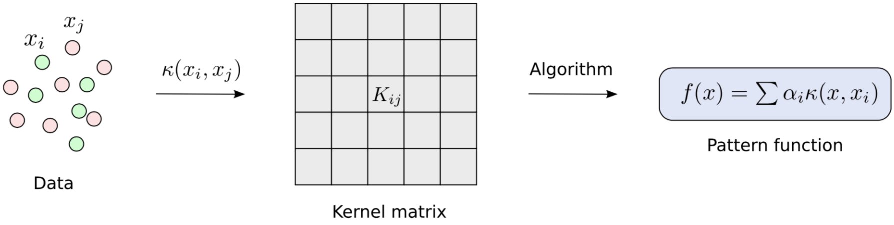
[1]

Distance-based. Embedding



- Many classifiers defined in Euclidean spaces ✓
- Computational complexity ✓
- Prediction ✓

Distance Kernels



Definite (PSD) Kernel

- All the SVM machinery works ✓
- Difficult to define/check ✓

Indefinite

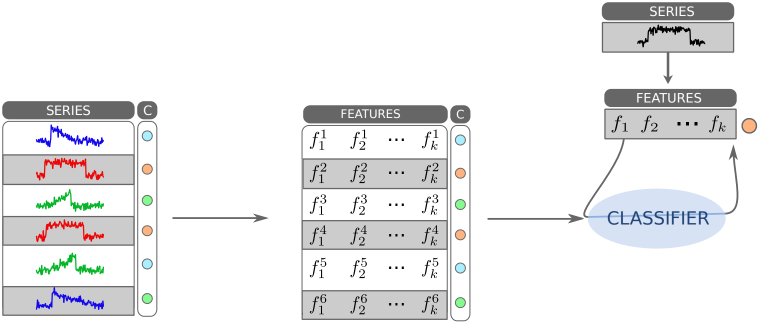
- Theoretical properties are lost ✓
- Easy to define ✓
- Some methods can not be applied ✓

Distance Kernels. Indefinite

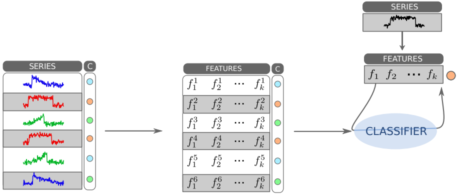
Gaussian Distance Substitution Kernels

$$GDS_d(x, x') = exp\left(-\frac{d(x, x')^2}{\sigma^2}\right) \text{ where } d = DTW, ..$$

Feature-based time series classification



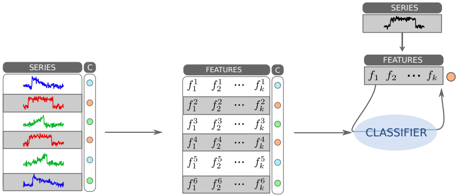
Feature-based time series classification



Features

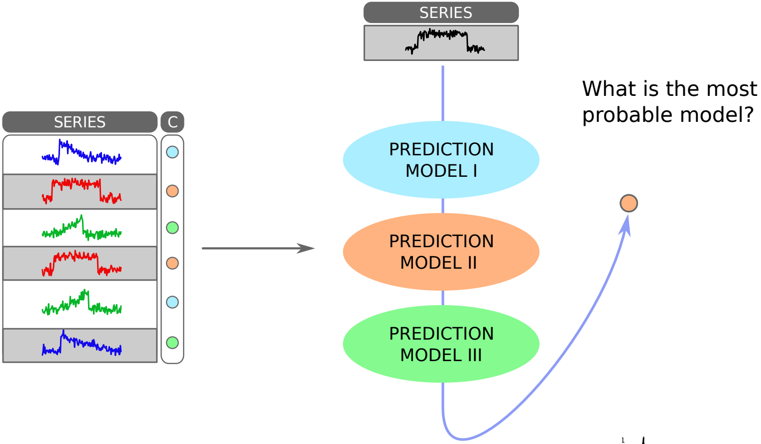
- Statistics: mean, variance
- Autorregresive coefficients
- Fourier coefficients
- Shift, trend, ...

Feature-based time series classification

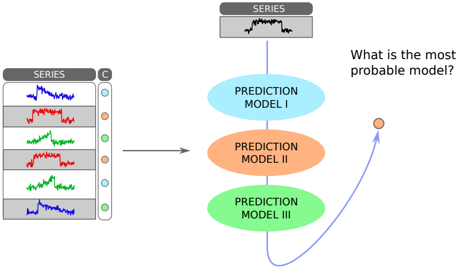


- Representation independent on the number of series ✓
- Interpretable representation ✓
- Challenge: what features to use?

Model-based time series classification



Model-based time series classification



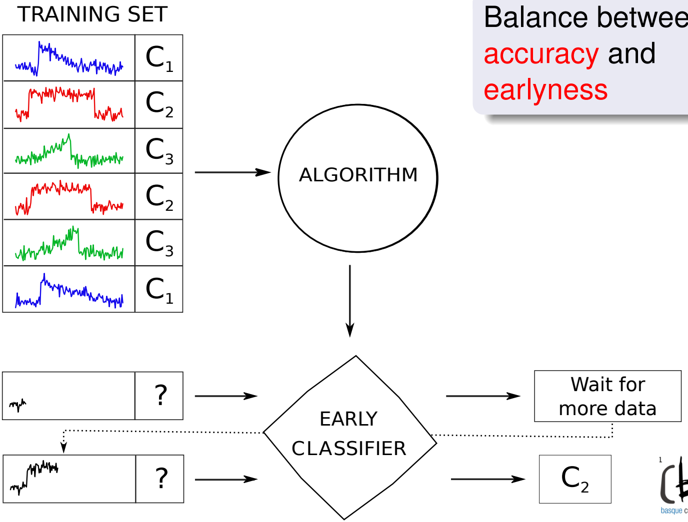
- Good results with an appropriate model ✓
- Choice of model ✓
- Existence of model ✓

Early time series classification

Examples

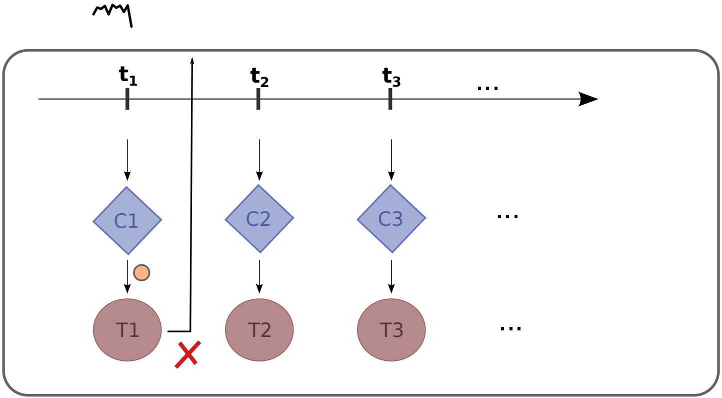
- Early activity recognition
- Early disease recognition in electrocardiograms
- Early detection of sepsis in newborn
- Early detection of failures in machines (predictive maintenance)

Early time series classification

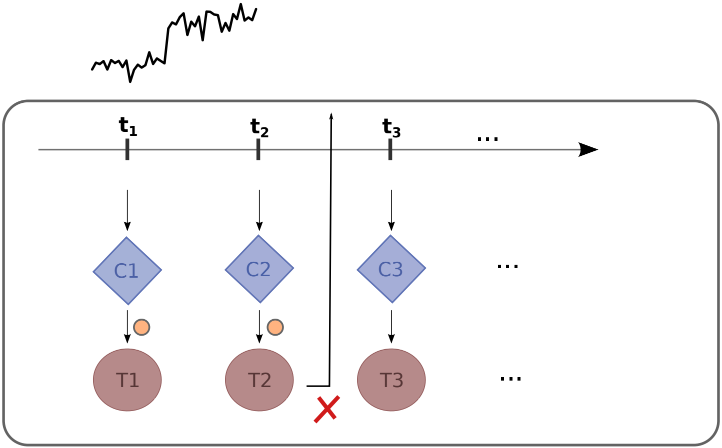


Balance between accuracy and earlyness

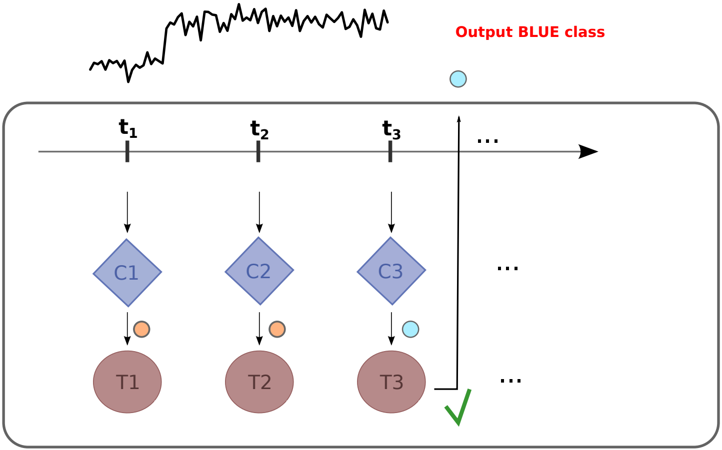
Early time series classificationc (Mori et al 2017, DAMI, TNNLS)



Early time series classification



Early time series classification



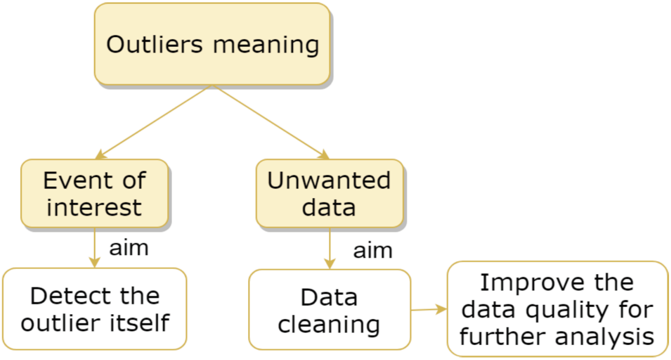
Multivariate time series classification

CHALLENGE

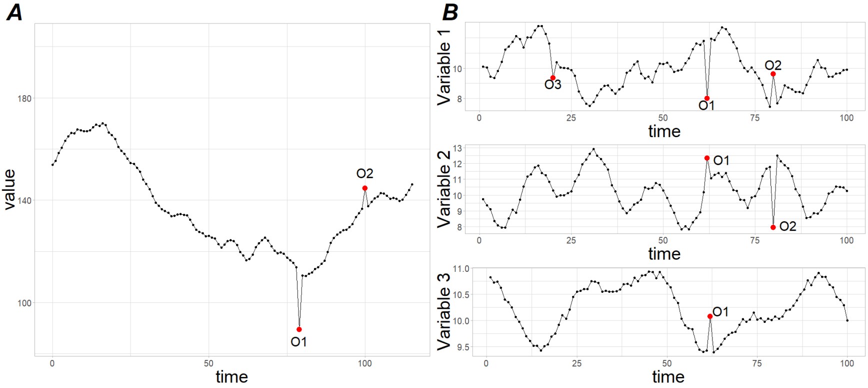
Outline of the presentation

1. Time Series Data Mining Activities
2. Clustering
3. (Early) Supervised Classification
4. Outlier/Anomaly Detection
5. Conclusions and Future Work

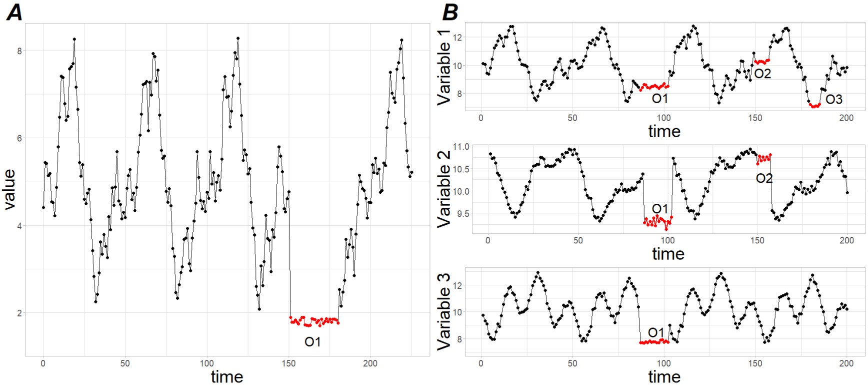
Outlier vs Anomaly



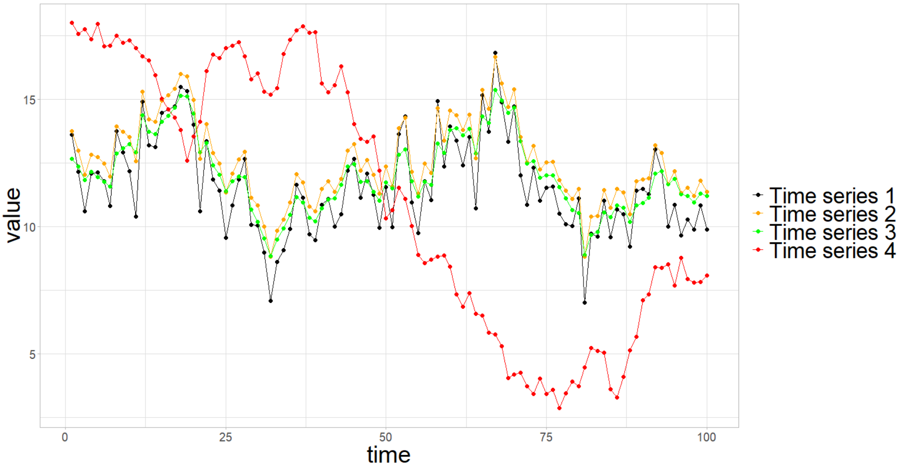
Type of outlier: point outlier



Type of outlier: subsequence outlier

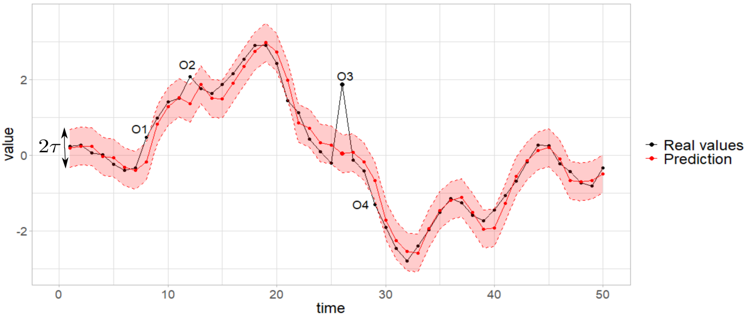


Type of outlier: series outlier



Outlier detection method: basic

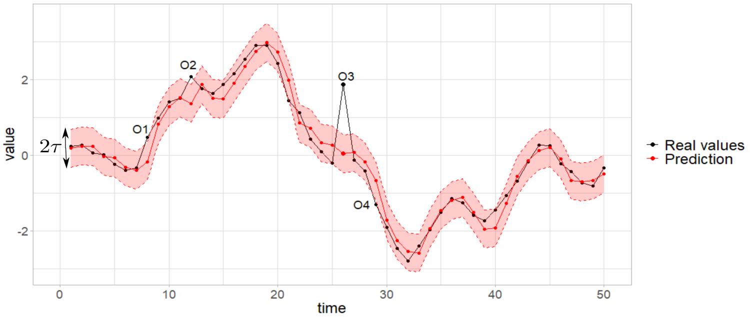
$$|x_t - \hat{x}_t| > \tau$$



Outlier detection method: basic

$$|x_t - \hat{x}_t| > \tau$$

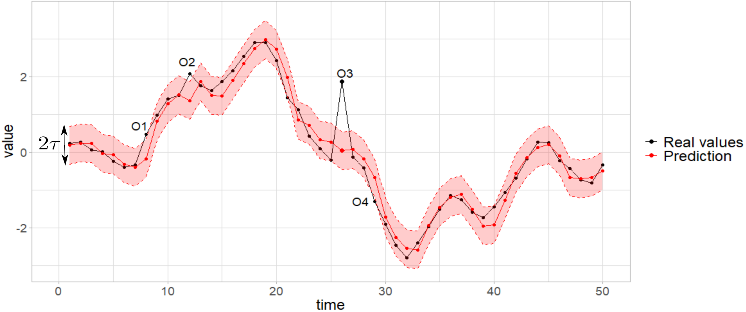
Median



Outlier detection method: basic

$$|x_t - \hat{x}_t| > \tau$$

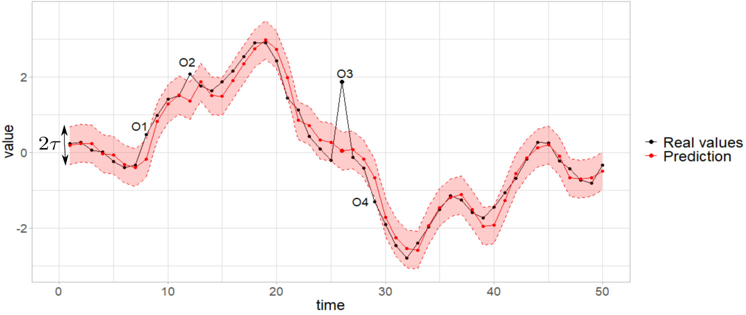
MAD



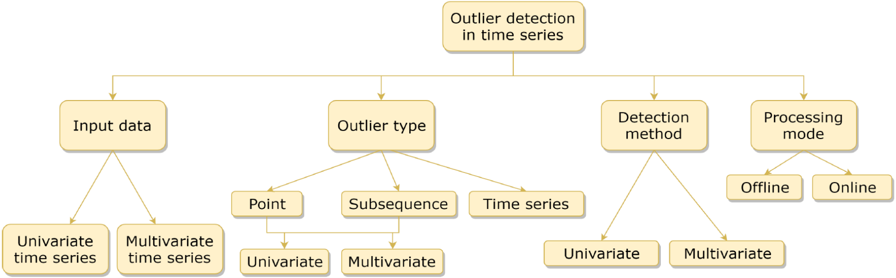
Outlier detection method: basic

$$|x_t - \hat{x}_t| > \tau$$

Model



An overview of outlier/anomaly detection



Outline of the presentation

1 Time Series Data Mining Activities

2 Clustering

3 (Early) Supervised Classification

4 Outlier/Anomaly Detection

5 Conclusions and Future Work

Not too explored lands

Challenges

- Time series subset selection
- Learning in weakly environments: semi-supervised, multi-label, crowd learning
- Theoretical bounds on learning: assumptions on the generating model

Collaboration

- Usue Mori (UPV/EHU), Amaia Abanda (BCAM)
- Ane Blazquez (Ikerlan), Angel Conde (Ikerlan)
- Aritz Perez (BCAM), Izaskun Oregui (Tecnalia), Javier del Ser (Tecnalia)
- Josu Ircio (Ikerlan), Aizea Lojo (Ikerlan)

Time Series Data Mining Challenges

Jose A. Lozano

Basque Center for Applied Mathematics (BCAM)
University of the Basque Country UPV/EHU

Time Series Days, Rennes, March 25-26, 2019