Recent advances in Time Series Classification

Simon Malinowski, LinkMedia Research Team

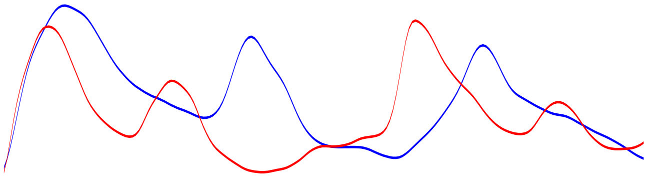
Classification day #3

Time Series Classification

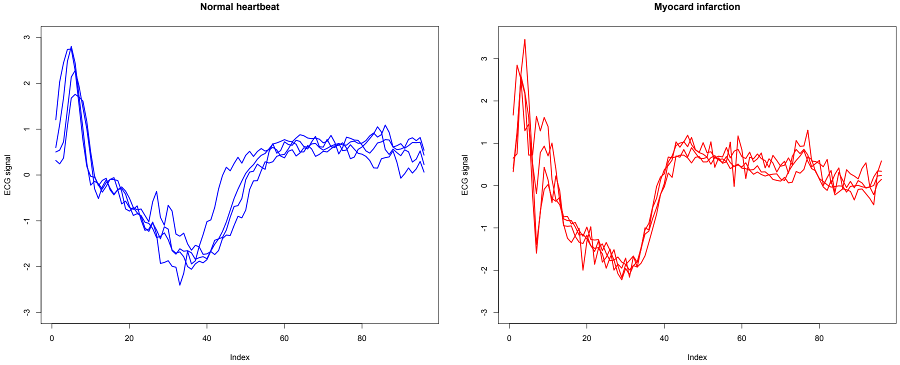
Time series...
A time series x is a set of values $x(1), x(2), \ldots, x(L), x(i) \in \mathbb{R}$, **ordered** in time.

...Classification

- Supervised task :
  We are given a training set $\mathcal{T} = \{(x_k, y_k), 1 \leq k \leq N\}$, where
  - x_k is a time series $\in \mathbb{R}^L$
  - y_k is a label associated to x_k, $y_k \in \{1, \ldots, C\}$
- Aim : Learn a relationship between the time series in $\mathcal{T}$ and their labels to predict unknown labels of testing time series.



Example of time series classification



Many other applications:
Satellite Image Time Series, gesture recognition, finance, music ...

Main approaches for time series classification

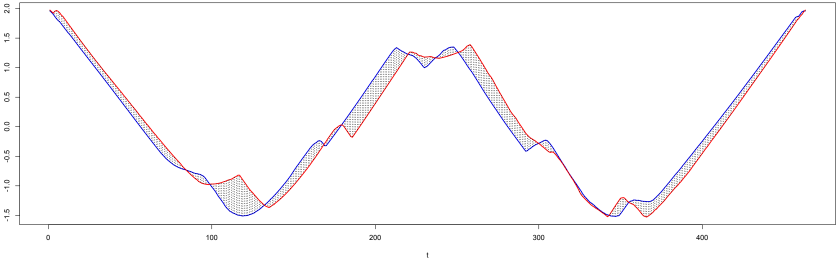
1. Distance-based approaches
   - rely on distance measures between raw time series
2. Shapelet-based approaches
   - rely on shapelets
3. Dictionary-based approaches
   - Bag-of-Words
4. Enhanced Bag-of-Words
5. Ensemble approaches

You can find a good recent review about TSC in [Bagnall et al., 2016]
Website : `timeseriesclassification.com`, with source codes and more than 80 datasets

Euclidean Distance

$X = x_1, \dots, x_n$ and $Y = y_1, \dots, y_n$ two time series.

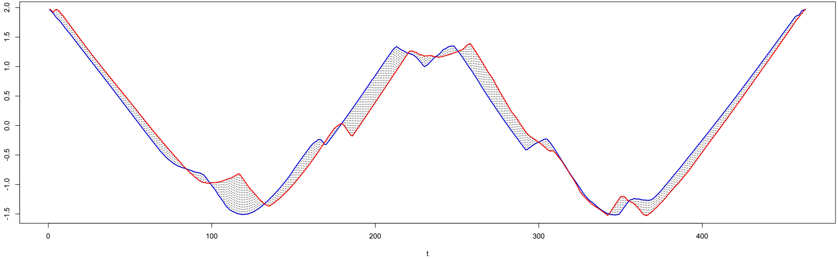
$$d(X, Y)^2 = \sum_{i=1}^{n} (x_i - y_i)^2$$



Euclidean Distance

$X = x_1, \ldots, x_n$ and $Y = y_1, \ldots, y_n$ two time series.

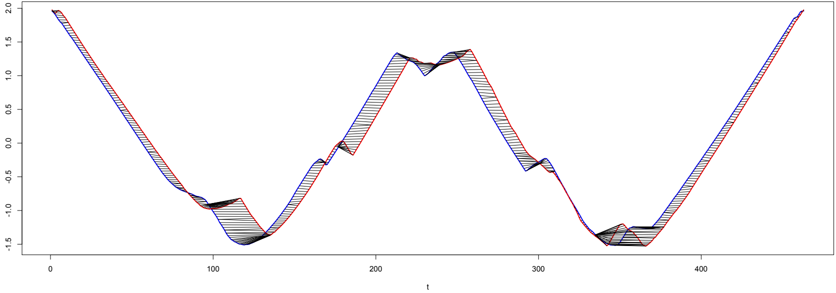
$$d(X, Y)^2 = \sum_{i=1}^{n} (x_i - y_i)^2$$



- ✓ Easy to compute
- ✓ Compatible with kernel machines
- × Same lengths
- × Shifts, warpings

Dynamic Time Warping [Sakoe and Chiba, 1978]

Measure based on the best alignment between two time series



Dynamic Time Warping

x_4					$d_{4,5}$
x_3					
x_2	$d_{2,1}$				
x_1	$d_{1,1}$	$d_{1,2}$			
	y_1	y_2	y_3	y_4	y_5

$$d_{i,j} = (x_i - y_j)^2$$

Dynamic Time Warping

x_4					$d_{4,5}$
x_3					
x_2	$d_{2,1}$		π_1		
x_1	$d_{1,1}$	$d_{1,2}$			
	y_1	y_2	y_3	y_4	y_5

$$S_1^2 = \sum_{i,j\in\pi_1} d_{i,j}$$

Dynamic Time Warping

x_4					$d_{4,5}$
x_3			π_2		
x_2	$d_{2,1}$		π_1		
x_1	$d_{1,1}$	$d_{1,2}$			
	y_1	y_2	y_3	y_4	y_5

$$S_1 = \sum_{i,j \in \pi_1} d_{i,j}$$

$$S_2^2 = \sum_{i,j \in \pi_2} d_{i,j}$$

Dynamic Time Warping

x_4					$d_{4,5}$
x_3			π_2		
x_2	$d_{2,1}$		π_1		
x_1	$d_{1,1}$	$d_{1,2}$			
	y_1	y_2	y_3	y_4	y_5

$$S_1 = \sum_{i,j \in \pi_1} d_{i,j}$$

$$S_2^2 = \sum_{i,j \in \pi_2} d_{i,j}$$

The best alignment is the one leading to the minimum score, this score being the DTW

Dynamic Time Warping

x_4					$d_{4,5}$
x_3			π_2		
x_2	$d_{2,1}$		π_1		
x_1	$d_{1,1}$	$d_{1,2}$			
	y_1	y_2	y_3	y_4	y_5

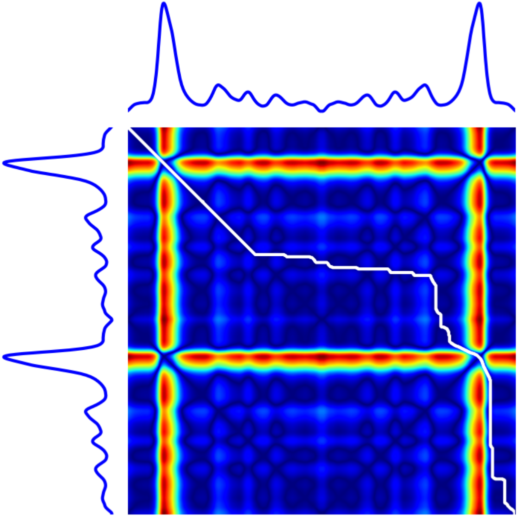
$$S_1 = \sum_{i,j \in \pi_1} d_{i,j}$$

$$S_2^2 = \sum_{i,j \in \pi_2} d_{i,j}$$

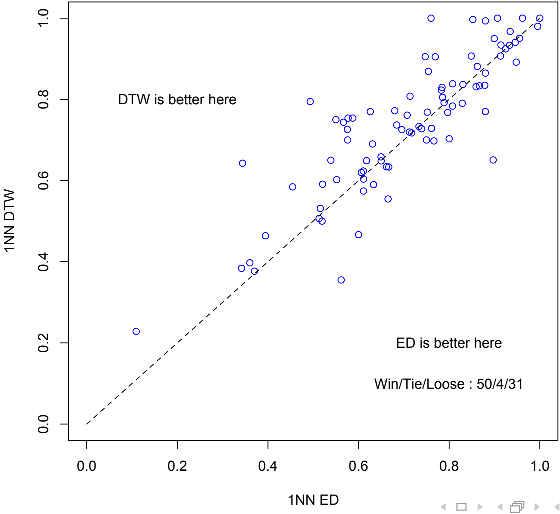
The best alignment is the one leading to the minimum score, this score being the DTW

- ✓ Shifts, warpings
- ✓ Good performance
- × Complexity
- × Not a proper distance

Dynamic Time Warping



ED versus DTW



Global Alignment Kernel - GAK [Cuturi, 2011]

x_4					$k_{4,5}$
x_3					
x_2	$k_{2,1}$		π_1		
x_1	$k_{1,1}$	$k_{1,2}$			
	y_1	y_2	y_3	y_4	y_5

$$k_{i,j} = e^{-(x_i - y_j)^2}$$

Global Alignment Kernel - GAK [Cuturi, 2011]

x_4					$k_{4,5}$
x_3					
x_2	$k_{2,1}$		π_1		
x_1	$k_{1,1}$	$k_{1,2}$			
	y_1	y_2	y_3	y_4	y_5

$$k_{i,j} = e^{-(x_i - y_j)^2}$$

DTW is score of the best path

Global Alignment Kernel - GAK [Cuturi, 2011]

x_4					$k_{4,5}$
x_3					
x_2	$k_{2,1}$		π_1		
x_1	$k_{1,1}$	$k_{1,2}$			
	y_1	y_2	y_3	y_4	y_5

$$k_{i,j} = e^{-(x_i - y_j)^2}$$

DTW is score of the best path
GAK considers all possible paths :

$$GAK(X, Y) = \sum_{\pi \in \Pi} \prod_{(i,j) \in \pi} k_{i,j}$$

Global Alignment Kernel - GAK [Cuturi, 2011]

x_4					$k_{4,5}$
x_3					
x_2	$k_{2,1}$		π_1		
x_1	$k_{1,1}$	$k_{1,2}$			
	y_1	y_2	y_3	y_4	y_5

$$k_{i,j} = e^{-(x_i - y_j)^2}$$

DTW is score of the best path

GAK considers all possible paths :

$$GAK(X, Y) = \sum_{\pi \in \Pi} \prod_{(i,j) \in \pi} k_{i,j}$$

✓ Shifts, warpings

✓ Good performance

✓ Can be used in kernel machines

× Complexity

1 Distance-Based Time Series Classification

2 Shapelet-Based Time Series Classification
- Shapelets
- Shapelet transform
- Learning Shapelets

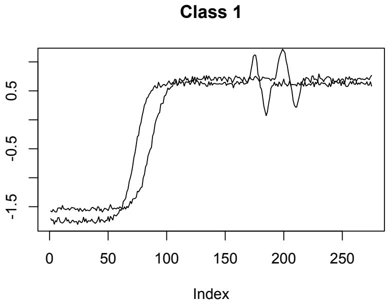
3 Dictionnary-based approaches

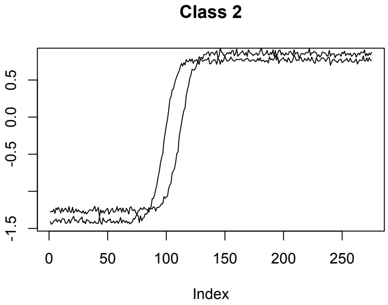
4 Efficient kernels for time series classification

5 Conclusion

Shapelets

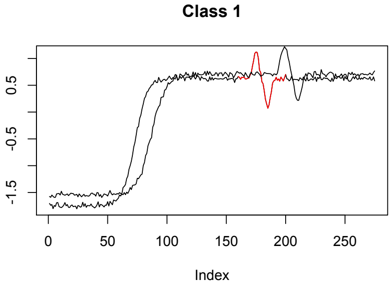
- First introduced in 2009 in [Ye and Keogh, 2009]
- A shapelet is a subsequence (of a time series) that is representative of a class

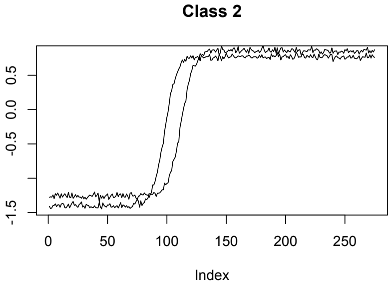




Shapelets

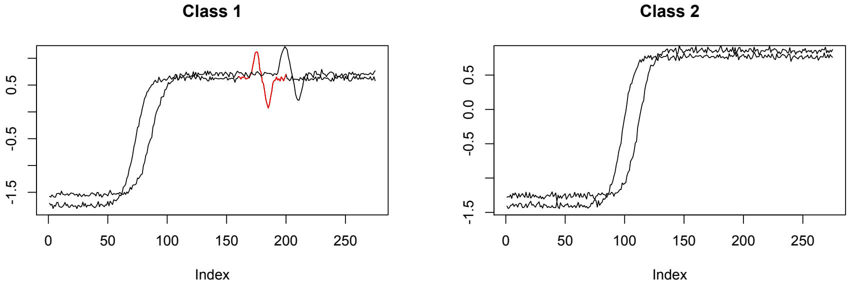
- First introduced in 2009 in [Ye and Keogh, 2009]
- A shapelet is a subsequence (of a time series) that is representative of a class





Shapelets

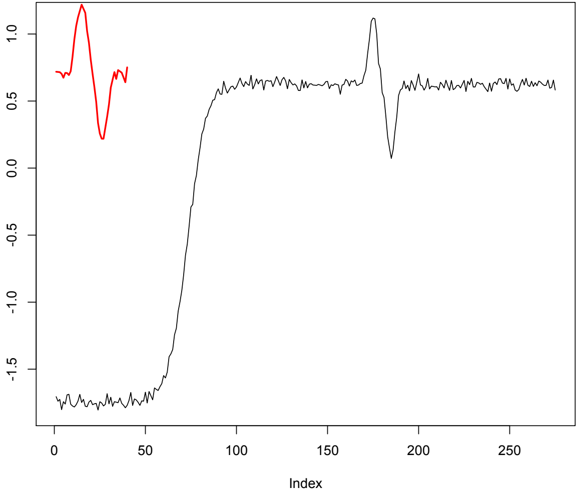
- First introduced in 2009 in [Ye and Keogh, 2009]
- A shapelet is a subsequence (of a time series) that is representative of a class



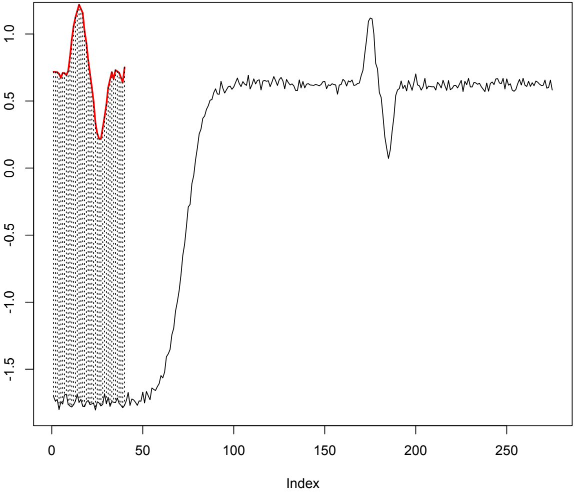
Shapelet s_1 is a representative of class 1 if

- distances between s_1 and time series of class 1 are low
- distances between s_1 and time series of other classes are high

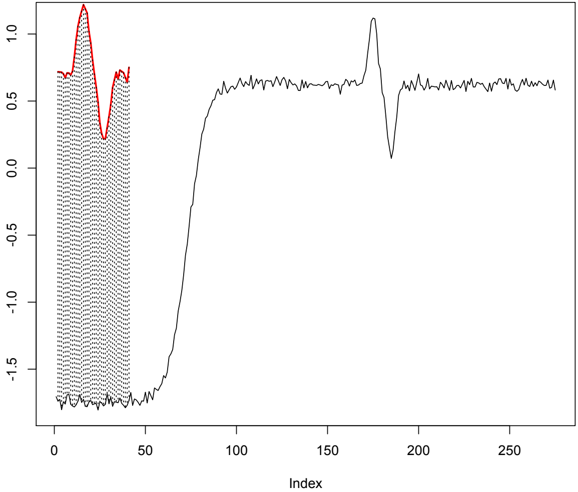
Distance between a shapelet and a time series



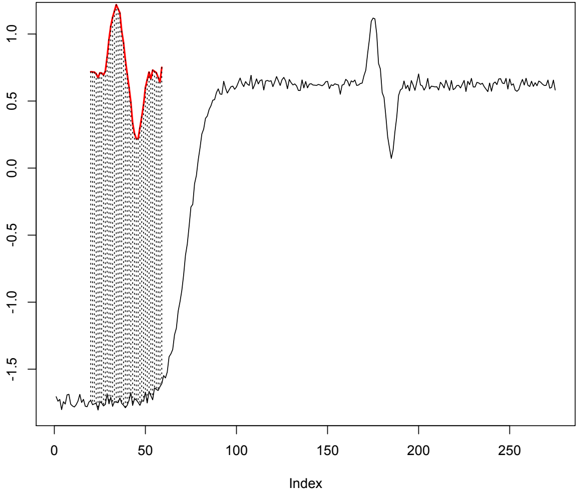
Distance between a shapelet and a time series



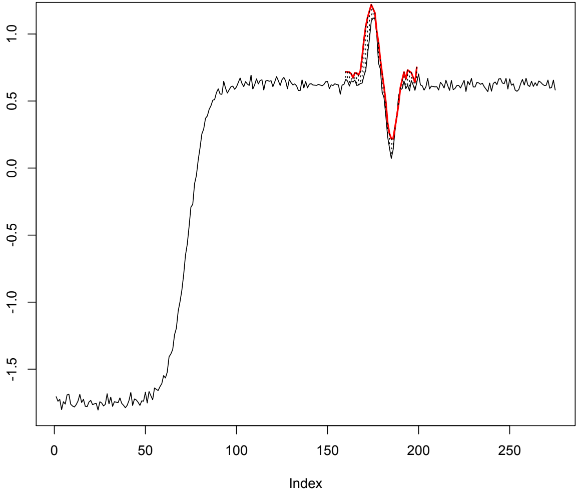
Distance between a shapelet and a time series



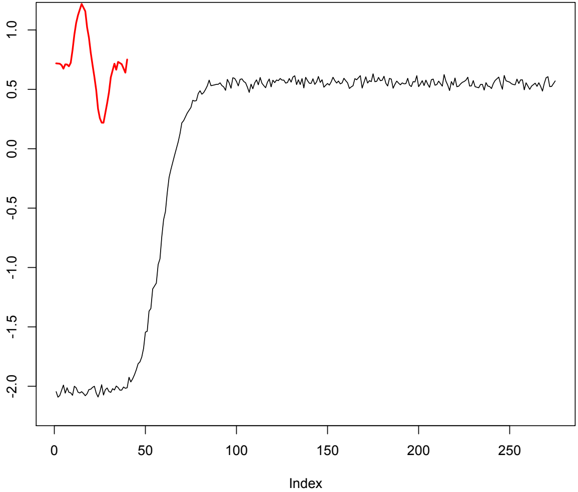
Distance between a shapelet and a time series



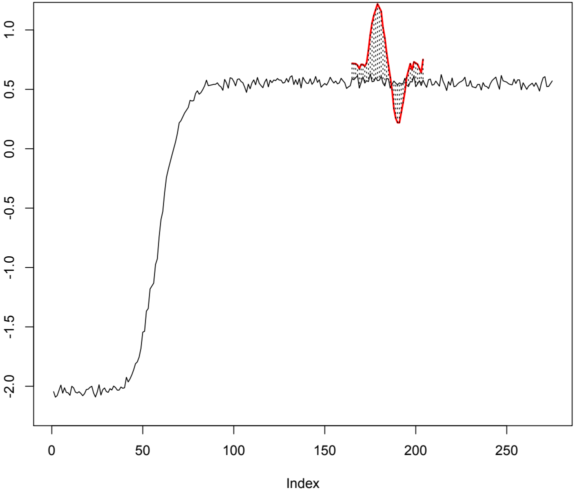
Distance between a shapelet and a time series



Distance between a shapelet and a time series

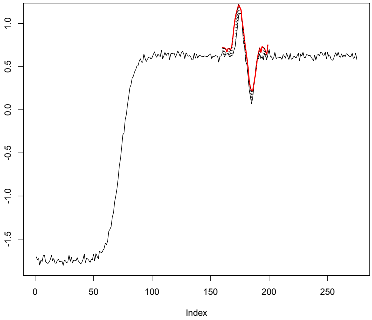


Distance between a shapelet and a time series

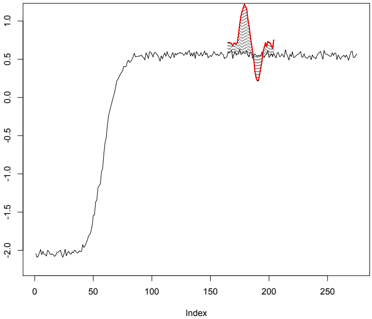


Distance between a shapelet and a time series

This shapelet represents class 1:



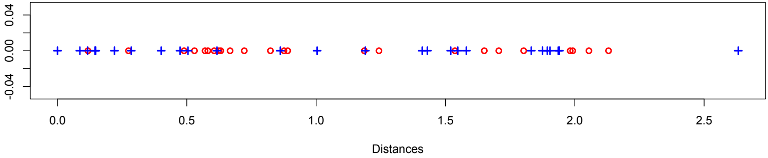
Low distance with class 1 time series



High distance with class 2 time series

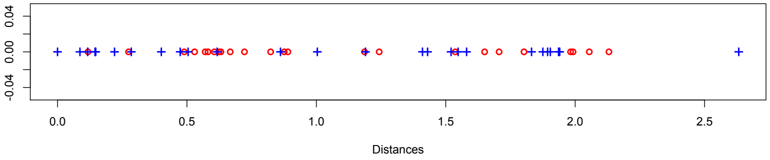
Bad or good shapelet ?

A bad shapelet ?

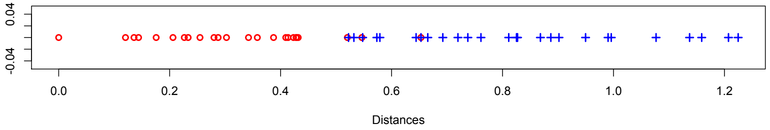


Bad or good shapelet ?

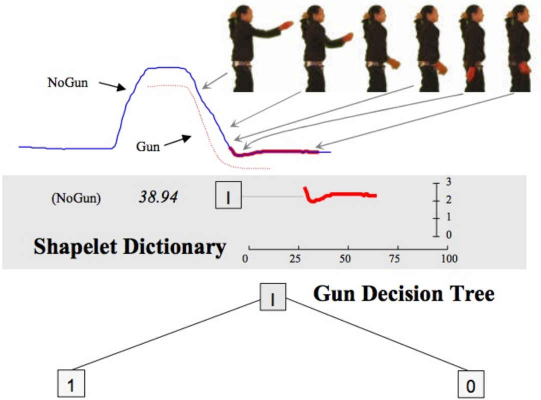
A bad shapelet ?



A good shapelet ?

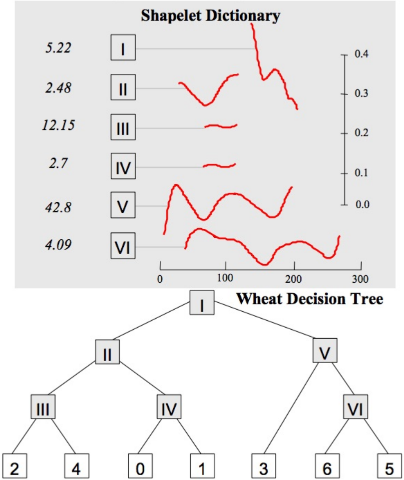


Shapelets



Images taken from [Ye and Keogh, 2009]

Shapelets

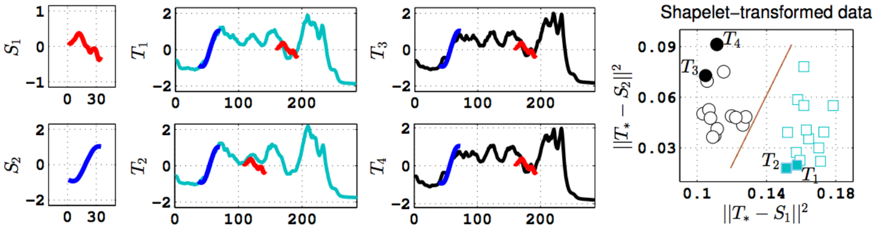


Images taken from [Ye and Keogh, 2009]

Shapelet transform [Hills et al., 2014]

This approach is based on 3 steps :

1. Extract the K best shapelets (information gain criterion)
2. Transform each training time series into a K-dimensional vector $M = M_1, \ldots, M_K$, where M_i is the distance between the time series and the i^{th} shapelet
3. Train a classifier on the transformed series



Picture taken from [Grabocka et al., 2014]

Learning shapelets [Grabocka et al., 2014]

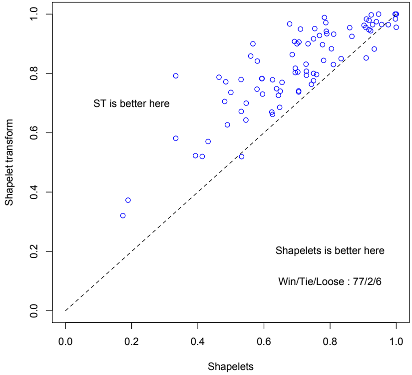
- Based on the shapelet transform principle
- BUT, shapelets are learned via logistic regression
- Given a set of weights $w_0, \ldots, w_K$ and a set of K shapelets

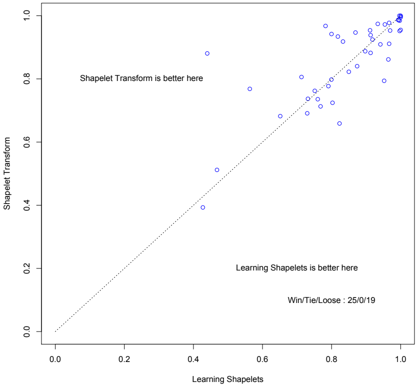
$$\hat{Y}_i = w_0 + \sum_{k=1}^{K} w_k M_{i,k}$$

$$\mathcal{L}(Y_i, \hat{Y}_i) = -Y_i ln(g(\hat{Y}_i)) - (1 - Y_i) ln(1 - g(\hat{Y}_i))$$

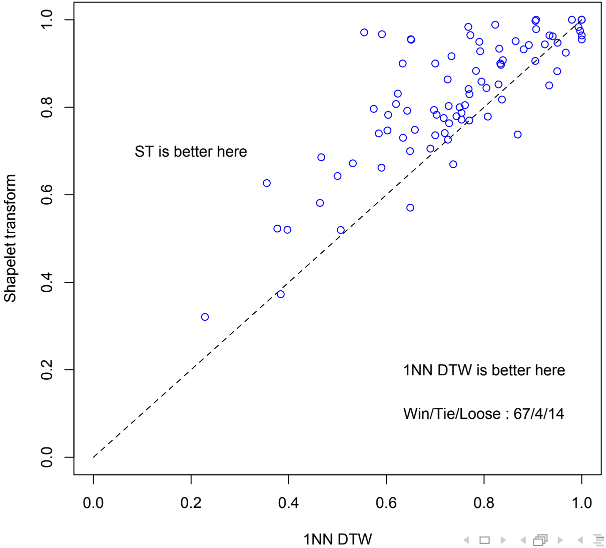
Objective : find K shapelets and K weights such that this loss is minimized
$\rightarrow$ gradient descent

Comparison between shapelet-based methods





Shapelet transform VS 1NN-DTW

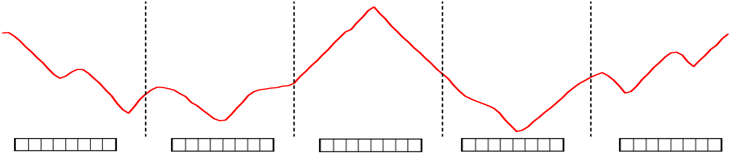


Bag-of-Words framework

The BoW framework is the following

$$
\begin{array}{ccccccccc}
 & & & & & & \textit{[codewords]} & & \\
 & & & & & & \downarrow & & \\
\text{Time Series} & \rightarrow & \begin{array}{c}\text{Selection of}\\ \text{Keypoints}\\ \text{(or windows)}\end{array} & \rightarrow & \begin{array}{c}\text{Description of}\\ \text{Keypoints}\\ \text{(or windows)}\end{array} & \rightarrow & \begin{array}{c}\text{Bag-of-Words}\\ \text{Representation}\end{array} & \rightarrow & \text{Classifier}
\end{array}
$$

Feature extraction over windows



- Features that have been considered in the literature :
  - Mean, slope, variance, extrema, starting time [Baydogan et al., 2013]
  - SAX symbols (quantized mean) [Lin et al., 2012, Senin and Malinchik, 2013]
  - Fourier Coefficients [Schäfer, 2014]
  - Wavelet coefficients [Wang et al., 2013]
  - Temporal-SIFT descriptors [Bailly et al., 2015]
- These features are then used in a Bag-of-Word approaches

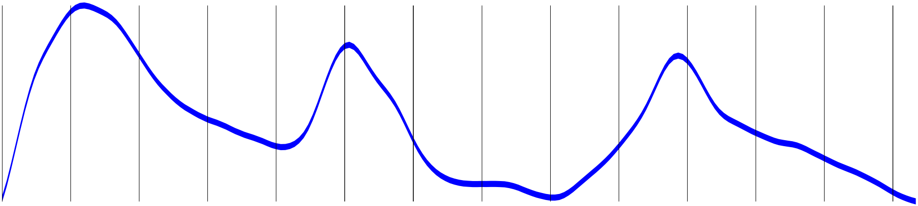
Bag-of-Temporal SIFT Words (BoTSW)

Joint work with Adeline Bailly, Romain Tavenard (Rennes 2) and Thomas Guyet (IRISA/AgroCampus)[Bailly et al., 2015]

BoTSW can be outlined by the following scheme:

						[codewords]		
						$\downarrow$		
Time Series	$\rightarrow$	Extraction of Keypoints	$\rightarrow$	Description of Keypoints	$\rightarrow$	Bag-of-Words Representation	$\rightarrow$	Classifier

Extraction of Keypoints - Dense Extraction



Keypoints are extracted densely in space.

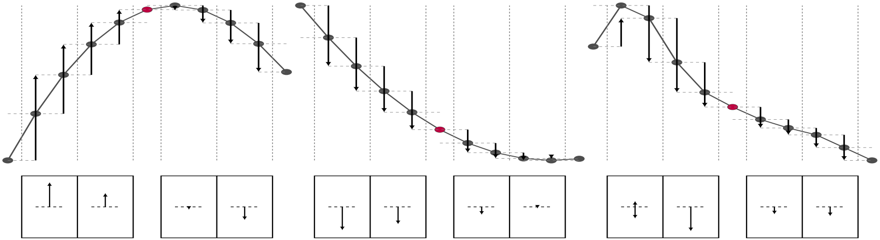
Outline

BoTSW can be outlined by the following scheme:

$$\text{Time Series} \rightarrow \begin{array}{c}\text{Extraction of}\\\text{Keypoints}\end{array} \rightarrow \begin{array}{c}\text{Description of}\\\text{Keypoints}\end{array} \rightarrow \begin{array}{c}\textit{[codewords]}\\\downarrow\\\text{Bag-of-Words}\\\text{Representation}\end{array} \rightarrow \text{Classifier}$$

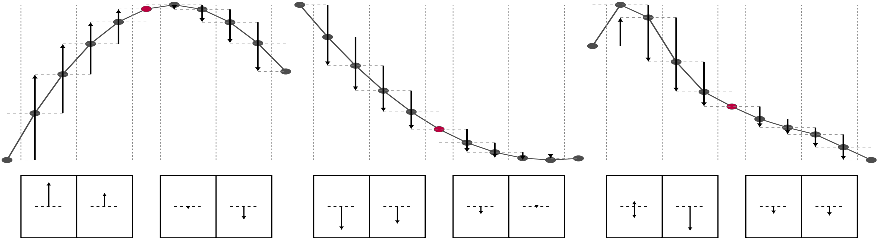
BoTSW - Feature vectors

Parameters: n_b, the number of blocks (4) and a, the block's size (2).



BoTSW - Feature vectors

Parameters: n_b, the number of blocks (4) and a, the block's size (2).



This description is done for the original time series $S(t)$ **and** for

$$L(t,\sigma) = G(t, k_0^s\ \sigma) * S(t), 0 \leq s \leq s_{max}$$

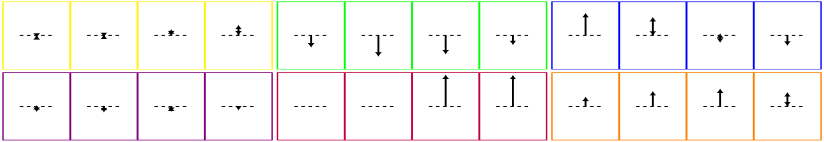
Outline

BoTSW can be outlined by the following scheme:

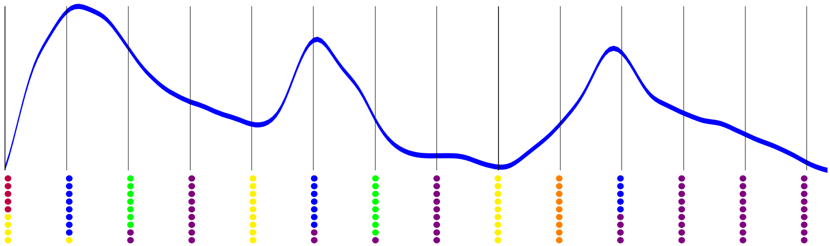
$$\text{Time Series} \rightarrow \begin{array}{c}\text{Extraction of}\\ \text{Keypoints}\end{array} \rightarrow \begin{array}{c}\text{Description of}\\ \text{Keypoints}\end{array} \rightarrow \begin{array}{c}\textit{[codewords]}\\ \downarrow \\ \text{Bag-of-Words}\\ \text{Representation}\end{array} \rightarrow \text{Classifier}$$

Generation of the dictionary

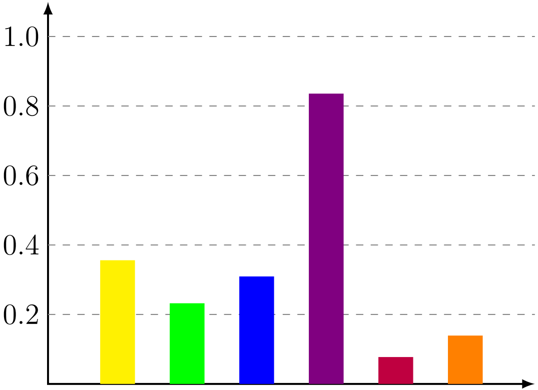
k-means to generate the codewords (here $k = 6$)



BoTSW - Feature vectors assignment



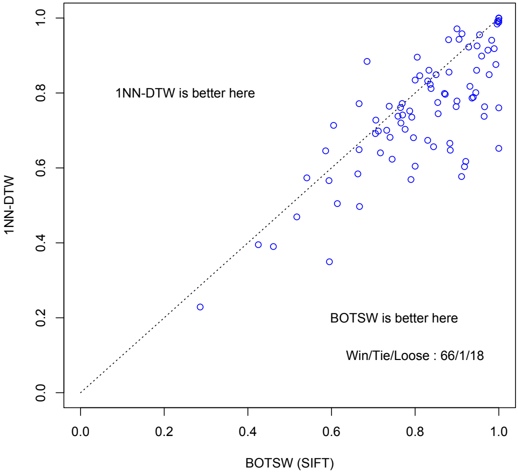
Bag-of-Words Representation



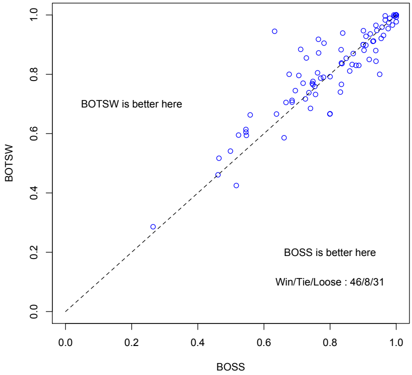
Different normalizations can be applied to this frequency vector :

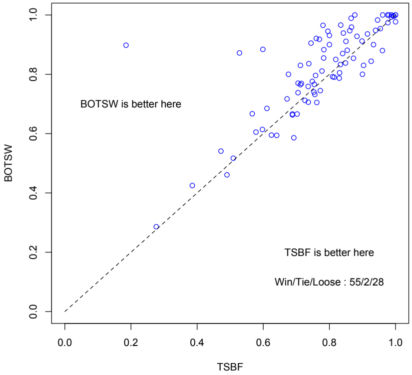
- Signed-Square-Root (SSR)
- TF-IDF

BOTSW versus DTW



BOTSW versus other BoW approaches





From BoW to kernels

Joint work with Romain Tavenard, Adeline Bailly (Rennes 2), Laetitia Chapel (IRISA), Benjamin Bustos (Univ. of Chile)[Tavenard et al., 2017]

Up to now,

- feature vectors are extracted from time series
- these vectors are **quantized** into words
- time series are represented as histogram of occurences of words

→ loss of information in the quantization

→ temporal information is lost

Design of kernels between sets of feature vectors

- no quantization of feature vectors
- integrate temporal information

Signature Quadratic Form Distance

Let $X = \{x_i, x_i \in \mathbb{R}^d\}_{i=1,\dots,n}$ be a time series (a set of feature vectors)
Let $Y = \{y_i, y_i \in \mathbb{R}^d\}_{i=1,\dots,m}$ be a time series (a set of feature vectors)
The SQFD between X and Y is defined as :

$$\mathrm{SQFD}(X,Y)^2 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} k(x_i,x_j) + \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m} k(y_i,y_j) - \frac{2}{n\cdot m}\sum_{i=1}^{n}\sum_{j=1}^{m} k(x_i,y_j),$$

where $k(.,.)$ is a local kernel between feature vectors (gaussian kernel for instance)

Signature Quadratic Form Distance

Let $X = \{x_i, x_i \in \mathbb{R}^d\}_{i=1,\dots,n}$ be a time series (a set of feature vectors)
Let $Y = \{y_i, y_i \in \mathbb{R}^d\}_{i=1,\dots,m}$ be a time series (a set of feature vectors)
The SQFD between X and Y is defined as :

$$\mathrm{SQFD}(X,Y)^2 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} k(x_i, x_j) + \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m} k(y_i, y_j) - \frac{2}{n \cdot m}\sum_{i=1}^{n}\sum_{j=1}^{m} k(x_i, y_j),$$

where $k(.,.)$ is a local kernel between feature vectors (gaussian kernel for instance)
It can be kernelized into

$$K_{\mathrm{SQFD}}(X,Y) = e^{-\gamma_f \, \mathrm{SQFD}(X,Y)^2}.$$

Incorporating time in the SQFD kernel

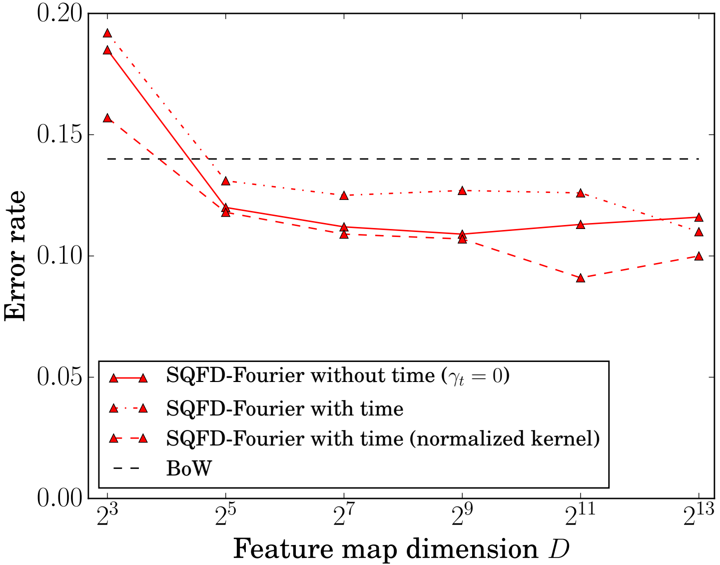
The more direct way is to integrate time in the local kernel

$$k(x_i, y_j) = e^{-\gamma_l \, \|x_i - x_j\|^2},$$

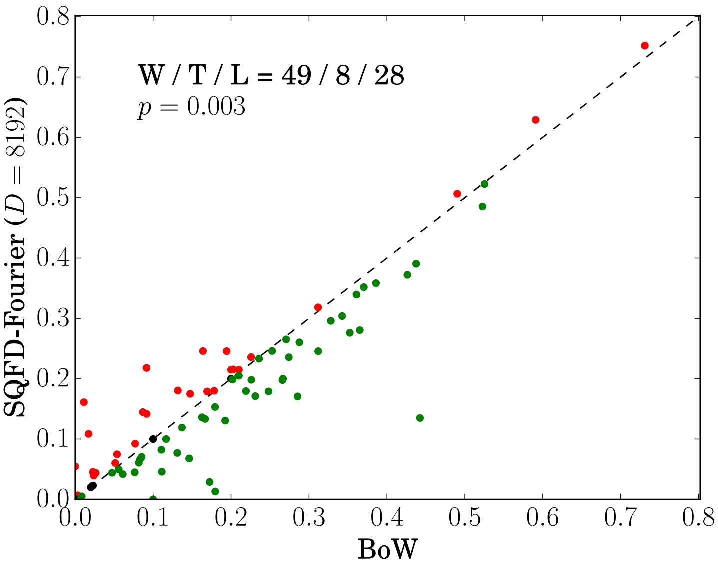
that now becomes :

$$k_{\mathrm{t}}(x_i, y_j) = e^{-\gamma_t \, (t_j - t_i)^2} \cdot k(x_i, y_j).$$

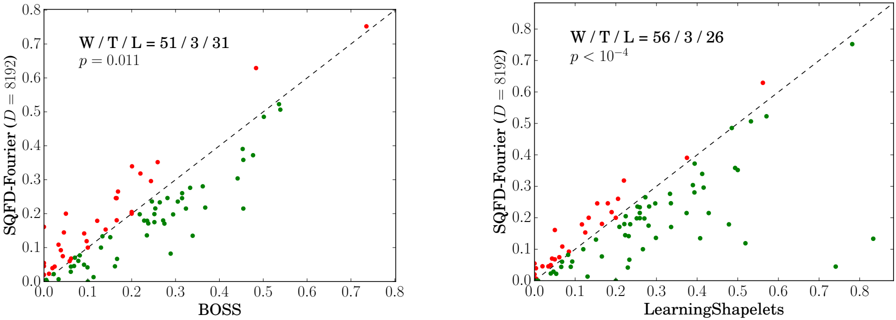
Classification performance



BoW VS Temporal kernel



Temporal kernel VS other approaches



Conclusion

Conclusion

- Three main familiy of approaches for TSC :
  - Distance-based
  - Shapelet-based
  - Dictionnary-based
- Shapelet and dictionnary-based are more accurate...
- ... but there is not A METHOD better than all others
- Dictionnary-based methods can be improved by temporal kernels

Bibliography I

Bagnall, A., Lines, J., Bostrom, A., Large, J., and Keogh, E. (2016).
The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances.
Data Mining and Knowledge Discovery, pages 1–55.

Bailly, A., Malinowski, S., Tavenard, R., Guyet, T., and Chapel, L. (2015).
Bag-of-Temporal-SIFT-Words for Time Series Classification.
In *Proceedings of the ECML-PKDD Workshop on Advanced Analytics and Learning on Temporal Data*.

Baydogan, M. G., Runger, G., and Tuv, E. (2013).
A Bag-of-Features Framework to Classify Time Series.
IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(11):2796–2802.

Bibliography II

Cuturi, M. (2011).
Fast global alignment kernels.
In *Proceedings of the International Conference on Machine Learning*, pages 929–936.

Grabocka, J., Schilling, N., Wistuba, M., and Schmidt-Thieme, L. (2014).
Learning time-series shapelets.
In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 392–401.

Hills, J., Lines, J., Baranauskas, E., Mapp, J., and Bagnall, A. (2014).
Classification of time series by shapelet transformation.
Data Mining and Knowledge Discovery, 28(4):851–881.

Lin, J., Khade, R., and Li, Y. (2012).
Rotation-invariant similarity in time series using bag-of-patterns representation.
International Journal of Information Systems, 39:287–315.

Bibliography III

Sakoe, H. and Chiba, S. (1978).
Dynamic programming algorithm optimization for spoken word recognition.
IEEE Transactions on Acoustics, Speech and Signal Processing, 26(1):43–49.

Schäfer, P. (2014).
The BOSS is concerned with time series classification in the presence of noise.
Data Mining and Knowledge Discovery, 29(6):1505–1530.

Senin, P. and Malinchik, S. (2013).
SAX-VSM: Interpretable Time Series Classification Using SAX and Vector Space Model.
Proceedings of the IEEE International Conference on Data Mining, pages 1175–1180.

Bibliography IV

Tavenard, R., Malinowski, S., Chapel, L., Bailly, A., Sanchez, H., and Bustos, B. (2017).
Efficient Temporal Kernels between Feature Sets for Time Series Classification.
In *Proceedings of ECML-PKDD*.

Wang, J., Liu, P., F.H. She, M., Nahavandi, S., and Kouzani, A. (2013).
Bag-of-Words Representation for Biomedical Time Series Classification.
Biomedical Signal Processing and Control, 8(6):634–644.

Ye, L. and Keogh, E. (2009).
Time series shapelets: a new primitive for data mining.
In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 947–956.

Questions?