Machine Learning on temporal data

Learning Dissimilarities on Time Series

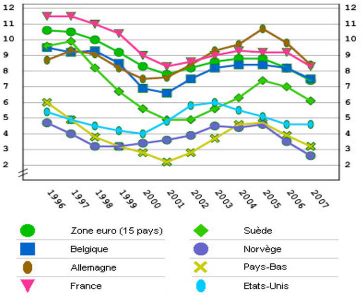
Ahlame Douzal (Ahlame.Douzal@imag.fr)
AMA, LIG, Université Joseph Fourier
Master 2R - MOSIG (2011)

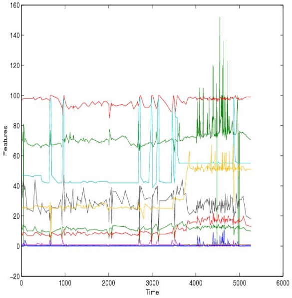
Plan

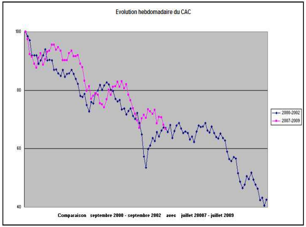
- Time series structure and sources
- Motivations of Learning from time series
- Values-based dissimilarities
- Behavior-based dissimilarities
- Adaptive dissimilarities
- Unified formalism for time series dissimilarities

Time series structure

- A set of observations made sequentially in time:







- Multidimensional time series

Patients	Time	Pressure	...	Temperature
pat_1	t_1	x_{11}	...	x_{1p}
	...			
	t_{n_1}	x_{n1}	...	x_{np}
pat_2	t_1	x_{11}	...	x_{1p}
	...			
	t_{n_2}	x_{n1}	...	x_{np}

Time series sources

- Physicians, biologists, physicists, computer scientists, ...
- generate, measure, observe, ...
- parameters, indicators, statistics, characteristics,...
- progress over times

Motivations for mining time series

Objectives

- Clustering time series: building groups from unlabeled time series *prototype extraction, dimensionality reduction, ...*
- Time series classification: time series assignments to known classes
- discriminating times series: finding models (functions, rules,...) to differentiate time series classes, localizing discriminant sub-sequences
- Multidimensional exploration of time series datasets *estimating relative proximities, optimizing a given criterion in a lower dimensionality space*

General approaches

- Multidimensional exploration: MDS, factorial approaches,...
- Supervised approaches, semi-supervised (time series totally, partially) labeled,...
- Unsupervised approaches (unlabeled time series)

Need of appropriate time series proximity measures !!

Internships on machine learning from temporal data

- Schneider Electric, the global specialist in Energy Management has launched an e-learning website Energy University to provide the latest information and professional training on Energy Efficiency concepts and best practice.
  - HOMES project experiments a method to know the effective performance of buildings:
    - Installation of sensors (temperature, humidity, presence detection) in buildings
    - Integration of sensors in the electrical panel
    - Establishment of a point of communication (phone transmitter box or ADSL box) to transfer information and sensor measurements
    - Transformation of electrical panels to add controllers
    - Integration of display systems in the building (screen display in the rooms)
  - Homes project objective: Machine learning and data analysis of consumption and conditions of the building (comfort, weather, structure and orientation of the building) to evaluate the energy savings potential.
- CEA (Research Center on Atomic Energy): on biological data

Categories of proximity measures

- Values-based proximity measures
  - without time warping: all L_p norm
  - with time warping: dynamic time warping, String edits, Fréchet distance (L_∞ norm or Chybechev-norm)
- Behavior-based proximity measures
- Behavior and values based proximity measures

Notations

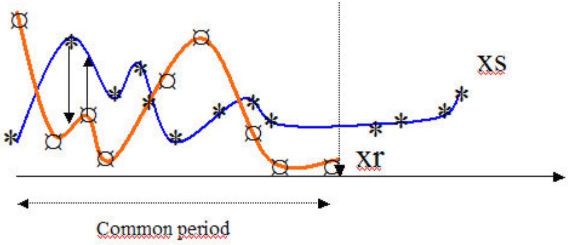
$$\mathbf{S}_1 = (u_1, \ldots, u_p) \text{ and } \mathbf{t}_1 = (t_{11}, \ldots, t_{p1})$$

$$\mathbf{S}_2 = (v_1, \ldots, v_q) \text{ and } \mathbf{t}_2 = (t_{12}, \ldots, t_{q2})$$

$$\delta_2 = (t_{q2} - t_{12}) \text{ the duration of } \mathbf{S}_2 \text{ and } q \text{ its length}$$

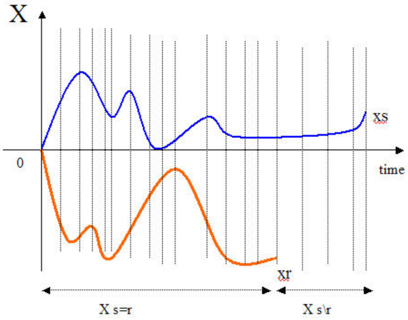
General assumptions

- $\delta_1 < \delta_2$
- $p \neq q$
- $\mathbf{t}_1$ and $\mathbf{t}_2$ are irregular



Preprocessing time series

- Centering
- re-sampling

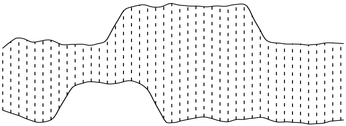


Values-based proximity measures

Mappings without time warping

$$\mathbf{S}_1 = (u_1, ..., u_p) \text{ and } \mathbf{t}_1 = (t_1, ..., t_p)$$
$$\mathbf{S}_2 = (v_1, .v_p, v_{p+1}, .., v_q) \text{ and } \mathbf{t}_2 = (t_1, ..., t_p, t_{p+1}, ..., t_q)$$

Example: Euclidean distance (L_2 norm) $\delta_E(S_1, S_2) = \left(\sum_{i=1}^{p} (u_i - v_i)^2 \right)^{\frac{1}{2}}$



- Invariance over time permutations
- Closeness on values
- Time series of the same duration and length

Values-based proximity measures

Mappings including time warping

L: a set of all possible mapping between S_1 and S_2, $r \in L$ defined by a sequence of m pairs:

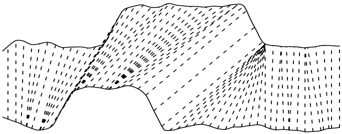
$$r = \left((u_1, v_1), \ldots, (u_{a_i}, v_{b_i}), (u_{a_{(i+1)}}, v_{b_{(i+1)}}), \ldots, (u_p, v_q)\right)$$

$$\text{with } a_i \in \{1, .., p\},\ b_i \in \{1, .., q\},$$

$$a_1 = b_1 = 1 \text{ and } a_m = p,\ b_m = q$$

and verifying for $i \in \{1, .., m-1\}$ (ordering constraint):

$$u_{a_{(i+1)}} = \begin{cases} u_{a_i} \\ \text{or} \\ u_{(a_i+1)} \end{cases} \qquad v_{b_{(i+1)}} = \begin{cases} v_{b_i} \\ \text{or} \\ v_{(b_i+1)} \end{cases}$$



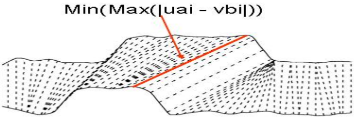
Values-based proximity measures

Fréchet distance

We note $|r|$ the mapping length representing the maximum span between coupled observations (L_∞ norm):

$$|r| = \max_{i=1,..,m} |u_{a_i} - v_{b_i}|$$

$$\delta_F(S_1, S_2) = \min_{r \in L} |r| = \min_{r \in L}(\max_{i=1,..,m} |u_{a_i} - v_{b_i}|)$$



Fréchet

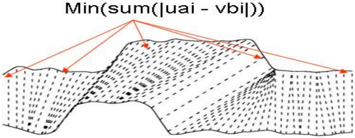
Values-based proximity measures

Dynamic Time Warping (DTW)

Let us consider a new definition of the mapping length:

$$|r| = \sum_{i=1,..,m} |u_{a_i} - v_{b_i}|$$

$$\delta_{\mathrm{DTW}}(S_1, S_2) = \min_{r \in L} |r| = \min_{r \in L}\Big(\sum_{i=1,..,m} |u_{a_i} - v_{b_i}|\Big)$$



DTW

The DTW implementation

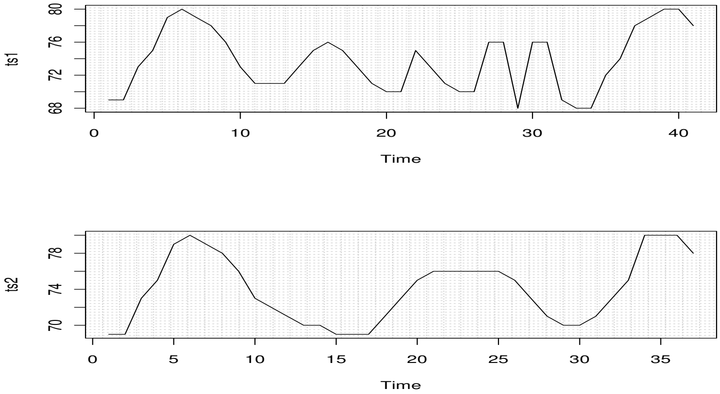
$$\mathbf{S}_1 = (u_1, ..., u_p),\ \mathbf{S}_2 = (v_1, ..., v_q) \text{ and } c(u_i; v_j) \text{ a cost function}$$

- Basic steps of the DP algorithm of the $DTW(S_1, S_2)$

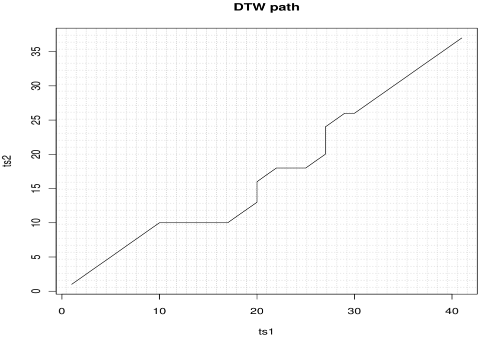
1. $D(1;j) = \sum_{k=1}^{j} c(u_1; v_k); \quad j \in [1;q]$
2. $D(i;1) = \sum_{k=1}^{i} c(u_k; v_1); \quad i \in [1;p]$
3. $D(i;j) = min\{D(i-1;j-1), D(i-1;j), D(i;j-1)\} + c(u_i; v_j)$

$i \in [1;p]; j \in [1;q]$.
The time cost of building this matrix is O(pq)

Illustration (1)



DTW path



Obtained alignment

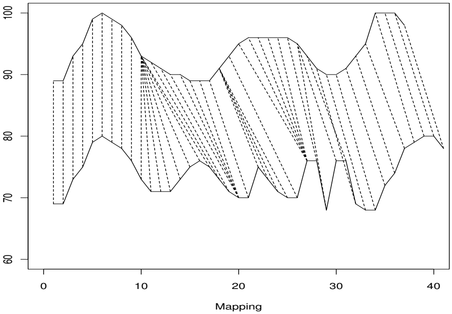


Illustration (2)

- 15 synthetic time series
- 3 classes: $F_1 = \{1, ..5\}$, $F_2 = \{6, ..10\}$ and $F_3 = \{11, ..15\}$

$$F_1 = \{f_1(t)/f_1(t) = g(t) + 2t + 3 + \epsilon\}$$
$$F_2 = \{f_2(t)/f_2(t) = \mu - g(t) + 2t + 3 + \epsilon\}$$
$$F_3 = \{f_3(t)/f_3(t) = 4g(t) - 3 + \epsilon\}$$

- g(t): a random discrete function,
- $\mu = E(g(t))$
- $\epsilon \rightsquigarrow N(0, 1)$,
- $2t + 3$: a linear trend effect.

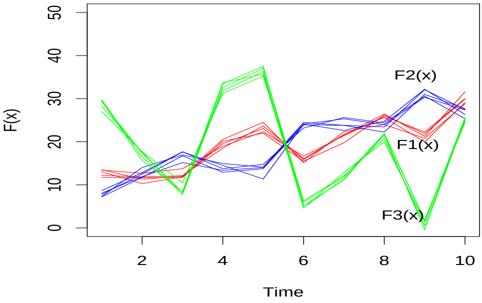
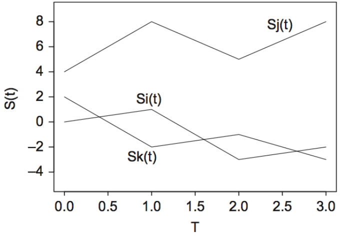


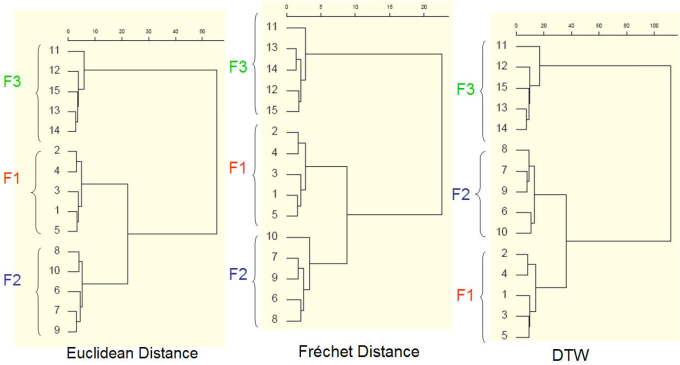
Illustration (2)



- Both the Euclidean distance and the dynamic time warping give S_i closer to S_k than to S_j,
- $d_E(S_i, S_k) = 4.24 < d_E(S_i, S_j) = 15.13 < d_E(S_j, S_k) = 16.15$
- $d_{dtw}(S_i, S_k) = 6 < d_{dtw}(S_i, S_j) = 29 < d_{dtw}(S_j, S_k) = 29$

Clustering time series

Hierarchical clustering



Behavior-based proximity measures

Behavior definition prior to the proximity measure proposition !

- S_1, S_2 are of similar behavior $\Leftrightarrow \forall\ [t_i, t_{i+1}]$ they increase or decrease simultaneously (monotonicity) with the same growth rate
- S_1, S_2 are of opposite behavior $\Leftrightarrow \forall\ [t_i, t_{i+1}]$ when S_1 increases, S_2 decreases and vice-versa with a same growth rates (in absolute value)

Main techniques to recover time series behaviors

- Slopes, derivatives, ... comparison
- Ranks comparison (Kendall, Spearman coefficients)
- Pearson correlation coefficient

Behavior based proximity measures

- slopes, derivatives, ... comparison

$$\mathbf{S}_1 = (u_1, ..., u_p),\ \mathbf{S}_2 = (v_1, ..., v_p)$$
$$\mathbf{t}_1 = (t_1, ..., t_p)$$

$$\delta_{deriv}(S_1, S_2) = \left(\sum_{i=1}^{p} \left(\frac{u_{i+1} - u_i}{t_{i+1} - t_i} - \frac{v_{i+1} - v_i}{t_{i+1} - t_i} \right)^2 \right)^{\frac{1}{2}}$$

Behavior-based proximity measures

- Kendall and Spearman conventional similarity between ordered variables

$$
\begin{aligned}
f(u_i, u_i') &= 1 \quad if \quad u_i < u_i' \\
&= -1 \quad if \quad u_i > u_i' \\
&= 0 \quad if \quad u_i = u_i' \\
S_1^* &= (f(u_1, u_2), f(u_1, u_3), ..., f(u_{p-1}, u_p)) \\
S_2^* &= (f(v_1, v_2), f(v_1, v_3), ..., f(v_{p-1}, v_p))
\end{aligned}
$$

$$
Kendall(S_1, S_2) = cor(S_1^*, S_2^*)
$$

Remark: $i = 1, ..., p$ assumed independent, overestimation of the behavior proximity, ignores growth intensity

Behavior-based proximity measures

Let $r(u_i)$ be the rank of u_i

$$\begin{aligned} S_1^* &= (r(u_1), ..., r(u_p)) \\ S_2^* &= (r(v_1), ..., r(v_p)) \end{aligned}$$

$$Spearman(S_1, S_2) = cor(S_1^*, S_2^*)$$

Remark: $i = 1, ..., p$ assumed independent,overestimation of the behavior proximity.

Behavior-based proximity measures

- Pearson correlation coefficient

$$Cor(S_1, S_2) = \frac{\sum_{i,i'}(u_i - u_{i'})(v_i - v_{i'})}{\sqrt{\sum_{i,i'}(u_i - u_{i'})^2}\sqrt{\sum_{i,i'}(v_i - v_{i'})^2}}$$

  - Overestimate the behavior proximity (involves all pairs of observations)
  - tendency effect sensitive
  - generally used for a mapping not involving time distortion, but easily generalized to mapping r including time distortion

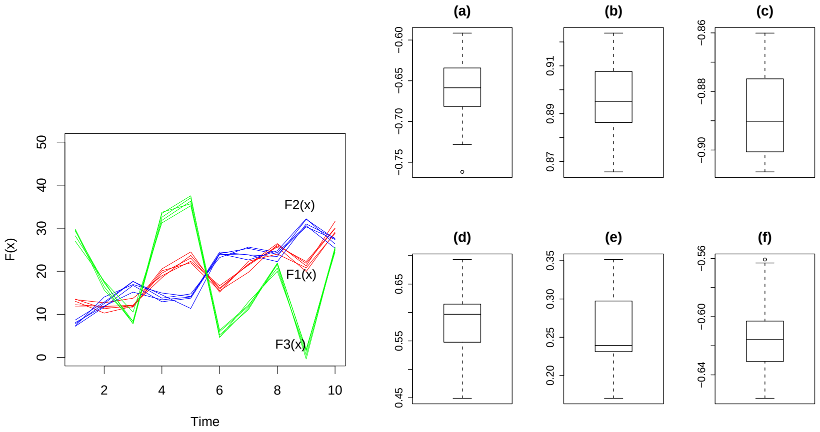
Behavior-based proximity measures

- Temporal correlation coefficient

$$cort(S_1, S_2) = \frac{\sum_{i=1}^{p-1}(u_{a_{(i+1)}} - u_{a_i})(v_{b_{(i+1)}} - v_{b_i})}{\sqrt{\sum_{i=1}^{p-1}(u_{a_{(i+1)}} - u_{a_i})^2}\sqrt{\sum_{i=1}^{p-1}(v_{b_{(i+1)}} - v_{b_i})^2}}$$

  - $cort = 1 \Leftrightarrow$ Similar behaviors,
  - $cort = -1 \Leftrightarrow$ Opposite behaviors,
  - $cort = 0 \Leftrightarrow$ Different behaviors
  - Noise sensitive

Illustration of *cor* vs *cort* distribution

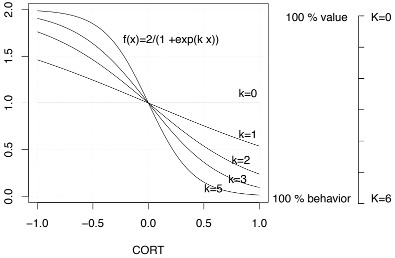


(a) $\text{CORT}(F_1, F_2)$, **(b)** $\text{CORT}(F_1, F_3)$) **(c)** $\text{CORT}(F_2, F_3)$,
(d)$\text{COR}(F_1, F_2)$, **(e)** $\text{COR}(F_1, F_3)$, **(f)** $\text{COR}(F_2, F_3)$

Behavior and values based proximity measures

- Basic behavior and values proximity measure: a weighted linear function combining values and behavior proximity measure
- Adaptive proximity measure

$$D_k(S_1, S_2) = f(B(S_1, S_2)) \cdot V(S_1, S_2) \text{ with } f(x) = \frac{2}{1 + \exp(k\, x)} \quad , \quad k \geq 0$$



k: the contribution of values and of behavior to D (to be learned)
B : the behavior based proximity
V : the values based proximity

Behavior and values based proximity measures

Example:
- Let $r = ((u_1, v_1), ..., (u_p, v_q))$ (without time warping)
- δ_E the proximity on values and *cort* the proximity on behavior

The adaptive proximity measure between S_1 and S_2:

$$D_k(S_1, S_2) = f(cort(S_1, S_2)) \cdot \delta_E(S_1, S_2)$$

Remark: V and B should be evaluated on the same mapping r

Unified formalism

Table 1
A unified formalism for time series metrics.

Type	R	$c(r)$	$Co(r)$	Metric
Values	$R \subset M$	$\sum_{i=1}^{m} \lvert u_{a_i} - v_{b_i} \rvert$	–	$d_{Dtw} = \min_{r \in R} \left(\sum_{i=1}^{m} \lvert u_{a_i} - v_{b_i} \rvert \right)$
	$R = \{r_0\}$	$\left(\sum_{i=1}^{m} (u_i - v_i)^2 \right)^{1/2}$	–	$d_E = c(r_0) = \left(\sum_{i=1}^{m} (u_i - v_i)^2 \right)^{1/2}$
Behavior	$R = \{r_0\}$	–	$Cor(r)$	$d_{Cor} = 1 - Cor(r_0)$
	$R = \{r_0\}$	–	$Cort(r)$	$d_{Cort} = 1 - Cort(r_0)$
	$R \subset M$	–	$Cor(r)$	$dtw_{Cor} = \min_{r \in R}(1 - Cor(r))$
	$R \subset M$	–	$Cort(r)$	$dtw_{Cort} = \min_{r \in R}(1 - Cort(r))$
Val. &	$R = \{r_0\}$	$\left(\sum_{i=1}^{m} (u_i - v_i)^2 \right)^{1/2}$	$Cor(r)$	$DE_k^{Cor} = \frac{2}{1+\exp(k\, Cor(r_0))} \left(\sum_{i=1}^{m} (u_i - v_i)^2 \right)^{1/2}$
	$R = \{r_0\}$	$\left(\sum_{i=1}^{m} (u_i - v_i)^2 \right)^{1/2}$	$Cort(r)$	$DE_k^{Cort} = \frac{2}{1+\exp(k\, Cort(r_0))} \left(\sum_{i=1}^{m} (u_i - v_i)^2 \right)^{1/2}$
Beh.	$R \subset M$	$\sum_{i=1}^{m} \lvert u_{a_i} - v_{b_i} \rvert$	$Cor(r)$	$DTW_k^{Cor} = \min_{r \in R} \left(\frac{2}{1+\exp(k\, Cor(r))} \sum_{i=1}^{m} \lvert u_{a_i} - v_{b_i} \rvert \right)$
	$R \subset M$	$\sum_{i=1}^{m} \lvert u_{a_i} - v_{b_i} \rvert$	$Cort(r)$	$DTW_k^{Cort} = \min_{r \in R} \left(\frac{2}{1+\exp(k\, Cort(r))} \sum_{i=1}^{m} \lvert u_{a_i} - v_{b_i} \rvert \right)$

References

[1] J. Kruskall, M. Liberman, The symmetric time warping algorithm: From continuous to discrete. In Time Warps, String Edits and Macromolecules., Addison-Wesley., 1983.
[2] G. Navarro, A guided tour to approximate string matching, ACM Com- puting Surveys 33 (1) (2001) 31–88.
[3] D. Sankoff, J. Kruskal, Time warps, string edits, and macromolecules: the theory and practice of sequence comparison, Addison-Wesley, 1983.
[4] D. Yu, X. Yu, Q. Hu, J. Liu, A. Wu, Dynamic time warping constraint learning for large margin nearest neighbor classification, Information Sci- ences 181 (2011) 27872796.
[5] H. Sakoe, S. Chiba, Dynamic programming algorithm optimization for spo- ken word recognition, IEEE Transactions on Acoustics, Speech, and Signal Processing 26 (1) (1978) 4349.
[6] C. A. Ratanamahatana, E. Keogh, Making time-series classification more accurate using learned constraints, in: SIAM International Conference on Data Mining, 2004, pp. 1122.
[7] R. Gaudin, N. Nicoloyannis, An adaptable time warping distance for time series learning, in: the 5th International Conference on Machine Learning and Applictions., 2006, pp. 213218.
[8] Y. Jeong, M. Jeong, O. Omitaomu, Weighted dynamic time warping for time series classification, Pattern Recognition 44 (2011) 22312240.