Learning Metrics for Time Series (Part 2)

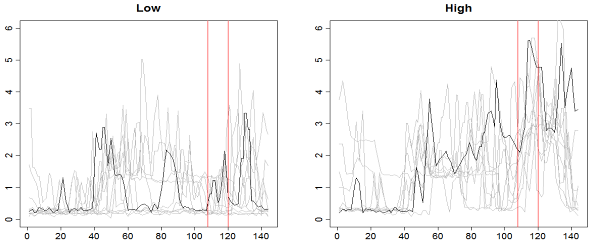
Ahlame Douzal

Univ. Grenoble Alpes
CNRS-LIG/AMA
France

Nov - 2013

Time Series: a complex data !

Real Data: UCI ML Household Electrical load consumption



Data characteristics:

- Each time series gives a daily load consumption
- In Low (vs. High) class the average consumption between 6-8 pm is lower (resp. higher) than the annual average consumption
- Consumption profiles are different within class

Objective and challenges

Objective

- The early classification (before 6 pm) of a load consumption to predict consumer demand on 6-8pm

Standard approaches

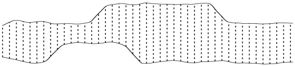
- Based on a standard time series metric (DTW)
- Assign a time series to the class of similar consumption profiles

Challenge

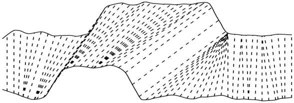
- Load consumption exhibit different global behaviors within classes or nearly similar ones between classes

Time series alignment

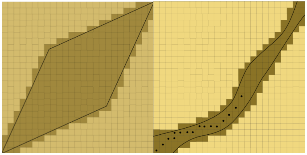
- Euclidean alignment



- Dynamic Time Warping DTW (Sankoff & Kruskal 1983, Kruskal & Liberman 1983).



- Alignment under global/local constraintes(Sakoe-Chiba 1975, Itakura 1971, Rabiner 1978)



Time series classification: DTW-based approaches

- **DTW Characteristics**
  - DTW metric, performed in light of a single pair of time series
  - Ignore the whole dynamics within and between classes
  - Achieved regardless of the analysis process (as clustering or classification)
  - Constrained alignments (temporal order, montonicity, no cross linkages,...)
  - Limited performances to classify/ cluster complex time series

- **Propositions to enhance DTW performances**
  - DTW constraint learning for large margin nearest neighbor classification Yu et al. 2011, Jeong et al. 2011
  - Probabilistic models to handle jointly time series analysis (clustering, classification) and matching processes Gaffney et al. 2005, Ramsay et al. 1998
  - Hierarchical bayesian model to detect slight differences between classes Listegarten et al. 2007

Problem: still assume similar global behaviors within classes !

Learning temporal matching for time series classification

Objective

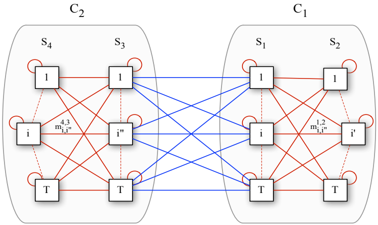
- Complex time series classification: different dynamics within classes, slight differences between classes

For this,

- Enlarge time series alignments to a less constrained temporal matching
- The learning process involves the whole dynamics within and between classes
- Match time series on their shared features within classes and distinctive ones between classes
- Derive a metric based on the highlighted discriminative features to be used for the time series classification.

Proposal's key

Given a set of linked time series (alignment, temporal matching,...)



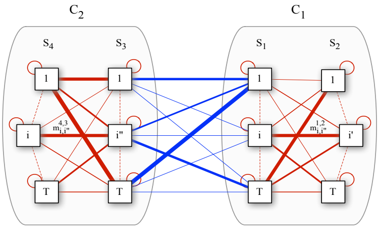
Idea

- Each link induces a variability corresponding to the divergence between the connected values
- To reveal shared features within a class, we minimize the within variance by removing links between non shared features
- To reveal differential features between classes, we maximize the between variance by removing links between shared features

Proposal's key

How?

- A new formalization of the classical variance/covariance for a set of time series, as well as for a partition of time series
- Strengthen or weaken links according to their contribution to the variances within and between classes



Variance/Covariance formalization for time series data

- $S_1, ..., S_n$ multivariate time series, of length T describing p variables
- X: description of $S_1, ..., S_n$ by p variables
- Assume time series linked through DTW alignment, temporal matching, ...

- We define $M_{(n,n)}(M_{ll'})$ as an adjacency block matrix
- A block $M_{ll'}$ specifies the linkage between S_l and $S_{l'}$
- A term of $M_{ll'}$ $m_{ii'}^{ll'} = 1$ if the instants i and i' of S_l and S_l' are aligned, 0 otherwise.

Variance induced by a set of time series

- **Variance/covariance induced by a set of time series**

$$V_M(X) \quad = \quad X^t(I - M')^t P(I - M')X$$

M': row normalized matrix of M

$(I - M')$: Laplacian matrix of the graph defined by the connected observations

Each observation is centered relative to the average of its neighborhood.

Remark: V_M leads to the total Variance/Covariance

- For a complete linkage defined by a unit matrix $M = 1$
- If each time series shrinks to one point

Variance induced by a partition of time series

- **Variance/covariance within et between classes of time series**

$$
\begin{aligned}
V_{M_W}(X) &= X^t(I - M'_W)^t P(I - M'_W)X \\
V_{M_B}(X) &= X^t(I - M'_B)^t P(I - M'_B)X
\end{aligned}
$$

intra-class matching M_W:
$m_{ii'}^{ll'} = 1$ if the linked time series belong to the same class, 0 otherwise.

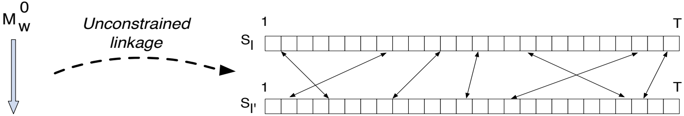
inter-class matching M_B:
$m_{ii'}^{ll'} = 1$ if the linked time series belong to different classes, 0 otherwise.

Remark: V_{M_W}, V_{M_B} lead to the within, between Variance/Covariance

- For a complete linkage defined by a unit matrix $M_W = 1$, $M_B = 1$
- If each time series shrinks to one point

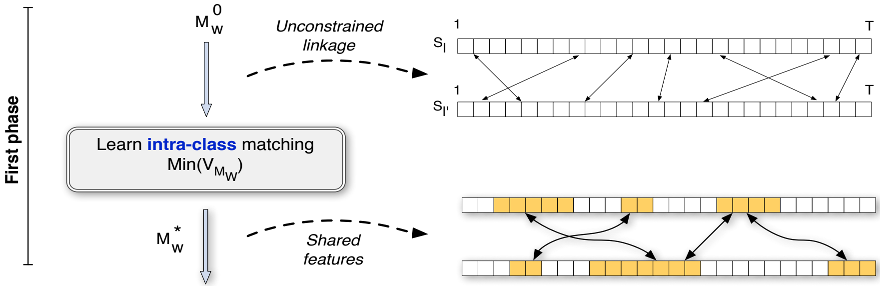
Learning a discriminative temporal matching

Two consecutive phases algorithm



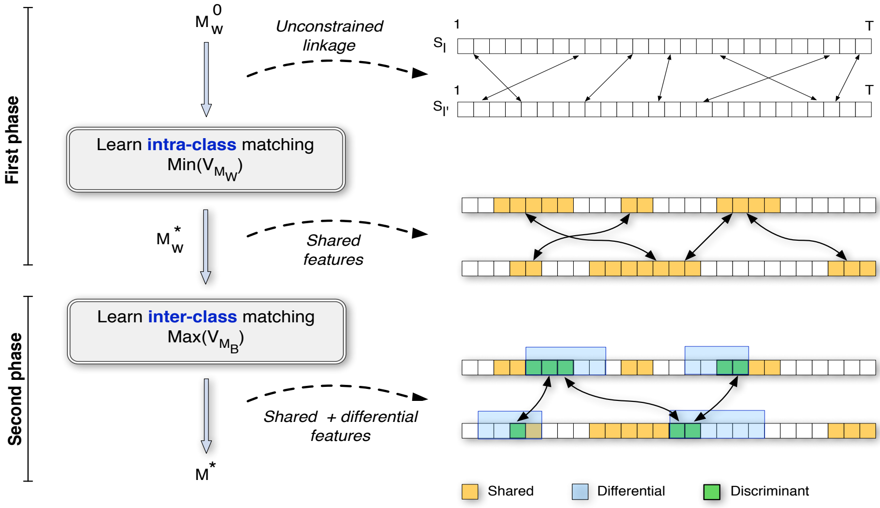
Learning a discriminative temporal matching

Two consecutive phases algorithm



Learning a discriminative temporal matching

Two consecutive phases algorithm



Learning the intra-class temporal matching

$S_l = (x_1^l, ..., x_T^l)$, $S_{l'} = (x_1^{l'}, ..., x_T^{l'})$ belonging to C_k ($|C_k| = n_k$)

$M \backslash (i, i', l, l')$: M after the removal of the link (i, i') between S_l and $S_{l'}$ ($m_{ii'}^{ll'} = 0$)

Outlines of the algorithm

1. Initialise M_W as a complete linkage

$$\forall\ i, i' \in \{1, ... T\} \text{ and } S_l,\ S_{l'} \text{ of the same class } m_{ii'}^{ll'} = 1$$

2. Calculate the contribution $C_{ii'}^{ll'}$ to the variance V_{M_W} of each link i, i' between S_l et $S_{l'}$

$$C_{ii'}^{ll'} = V_{M_W} - V_{M_W \backslash (i,i',l,l')}$$

3. Delete links (i, i') ($m_{ii'}^{ll'} = 0$) of positive contributions $C_{ii'}^{ll'} > 0$

4. Iterate steps 2 and 3 until V_{M_W} stabilization

Learning the intra-class temporal matching

$S_l = (x_1^l, ..., x_T^l)$, $S_{l'} = (x_1^{l'}, ..., x_T^{l'})$ belonging to C_k ($|C_k| = n_k$)

$M \backslash (i, i', l, l')$: M after the removal of the link (i, i') between S_l and $S_{l'}$ ($m_{ii'}^{ll'} = 0$)

Outlines of the algorithm

1. Initialise M_W as a complete linkage

$$\forall\ i, i' \in \{1, ... T\} \text{ and } S_l,\ S_{l'} \text{ of the same class } m_{ii'}^{ll'} = 1$$

2. Calculate the contribution $C_{ii'}^{ll'}$ to the variance V_{M_W} of each link i, i' between S_l et $S_{l'}$

$$C_{ii'}^{ll'} = V_{M_W} - V_{M_W \backslash (i,i',l,l')}$$

3. Delete links (i, i') ($m_{ii'}^{ll'} = 0$) of positive contributions $C_{ii'}^{ll'} > 0$

4. Iterate steps 2 and 3 until V_{M_W} stabilization

Learning the intra-class temporal matching

$S_l = (x_1^l, ..., x_T^l)$, $S_{l'} = (x_1^{l'}, ..., x_T^{l'})$ belonging to C_k ($|C_k| = n_k$)

$M \backslash (i, i', l, l')$: M after the removal of the link (i, i') between S_l and $S_{l'}$ ($m_{ii'}^{ll'} = 0$)

Outlines of the algorithm

1. Initialise M_W as a complete linkage

$$\forall\ i, i' \in \{1, ...T\} \text{ and } S_l,\ S_{l'} \text{ of the same class } m_{ii'}^{ll'} = 1$$

2. Calculate the contribution $C_{ii'}^{ll'}$ to the variance V_{M_W} of each link i, i' between S_l et $S_{l'}$

$$C_{ii'}^{ll'} \quad = \quad V_{M_W} - V_{M_W \backslash (i,i',l,l')}$$

3. Delete links (i, i') ($m_{ii'}^{ll'} = 0$) of positive contributions $C_{ii'}^{ll'} > 0$

4. Iterate steps 2 and 3 until V_{M_W} stabilization

Learning the intra-class temporal matching

$S_l = (x_1^l, ..., x_T^l)$, $S_{l'} = (x_1^{l'}, ..., x_T^{l'})$ belonging to C_k ($|C_k| = n_k$)

$M \backslash (i, i', l, l')$: M after the removal of the link (i, i') between S_l and $S_{l'}$ ($m_{ii'}^{ll'} = 0$)

Outlines of the algorithm

1. Initialise M_W as a complete linkage

$$\forall\ i, i' \in \{1, ... T\} \text{ and } S_l,\ S_{l'} \text{ of the same class } m_{ii'}^{ll'} = 1$$

2. Calculate the contribution $C_{ii'}^{ll'}$ to the variance V_{M_W} of each link i, i' between S_l et $S_{l'}$

$$C_{ii'}^{ll'} = V_{M_W} - V_{M_W \backslash (i,i',l,l')}$$

3. Delete links (i, i') ($m_{ii'}^{ll'} = 0$) of positive contributions $C_{ii'}^{ll'} > 0$
4. Iterate steps 2 and 3 until V_{M_W} stabilization

Non degenerate and convergence conditions

$\forall k \in \{1, ..., K\}, \ \forall (l, l') \in C_k, \ \forall (i, i') \in [1, T]^2$

Variance definition

1- $m_{ii}^{ll} > 0$

2- M_W row-normalized : $\sum_{l'=1}^{n_k} \sum_{i'=1}^{T} m_{ii'}^{ll'} = 1$

Non-degenerate variance

3- Each obs. of S_l should be linked to at least one obs. of $S_{l'}$: $\sum_{i'=1}^{T} m_{ii'}^{ll'} > 0$

Convergence of the variance minimization process

4- The delete of (i, i') impacts the i et i' neighborhoods (rows i and i'): at each iteration, delete the link of maximal positive contribution per row

Non degenerate and convergence conditions

$\forall k \in \{1, ..., K\},\ \forall (l, l') \in C_k,\ \forall (i, i') \in [1, T]^2$

Variance definition

1- $m_{ii}^{ll} > 0$

2- M_W row-normalized : $\sum_{l'=1}^{n_k} \sum_{i'=1}^{T} m_{ii'}^{ll'} = 1$

Non-degenerate variance

3- Each obs. of S_l should be linked to at least one obs. of $S_{l'}$: $\sum_{i'=1}^{T} m_{ii'}^{ll'} > 0$

Convergence of the variance minimization process

4- The delete of (i, i') impacts the i et i' neighborhoods (rows i and i'): at each iteration, delete the link of maximal positive contribution per row

Non degenerate and convergence conditions

$\forall k \in \{1, ..., K\},\ \forall (l, l') \in C_k,\ \forall (i, i') \in [1, T]^2$

Variance definition

1- $m_{ii}^{ll} > 0$

2- M_W row-normalized : $\sum_{l'=1}^{n_k} \sum_{i'=1}^{T} m_{ii'}^{ll'} = 1$

Non-degenerate variance

3- Each obs. of S_l should be linked to at least one obs. of $S_{l'}$: $\sum_{i'=1}^{T} m_{ii'}^{ll'} > 0$

Convergence of the variance minimization process

4- The delete of (i, i') impacts the i et i' neighborhoods (rows i and i'): at each iteration, delete the link of maximal positive contribution per row

Derive discriminative metric

M_*: the learned discriminative matching

- Let $M_*^{l\cdot}$ be the average matching to S_l:

$$M_*^{l\cdot} = \frac{1}{(n - n_k)\,T} \sum_{l'} M_*^{ll'}$$

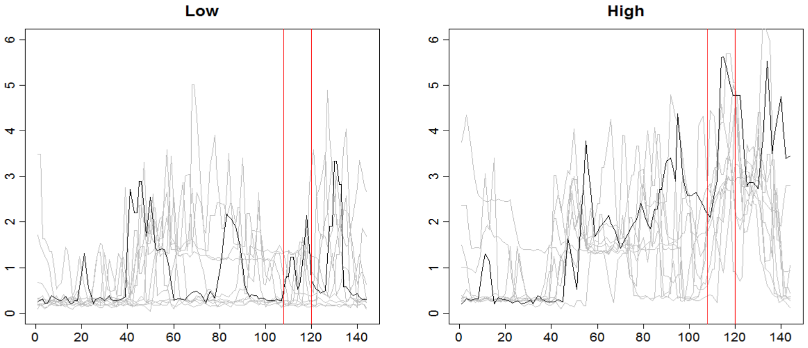
with $y_{l'} \neq y_l = k$

- The discriminative dissimilarity between S_{New} and S_l

$$D_l(S_l, S_{New}) = \min_{r \in \{0,..,T-1\}} \Big(\sum_{|i-i'| \leq r;\,(i,i') \in [1,T]^2} \frac{m^l_{ii'}}{\sum_{|i-i'| \leq r} m^l_{ii'}} (x^l_i - x^{New}_{i'})^2 \Big)$$

where r corresponds to the Sakoe-Chiba band width.

Classification of the household electric power consumption

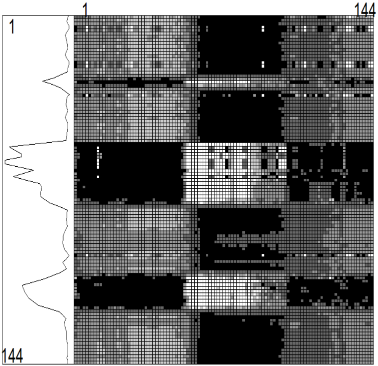


Objective: Early classification of consumption profiles for consumer demand prediction on 6-8pm

Classification of the household electric power consumption

Learned discriminant matching (CONSLEVEL)

M_W^* (Low)

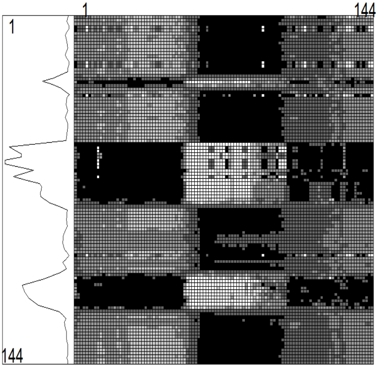


M_B^* (Low vs. High)

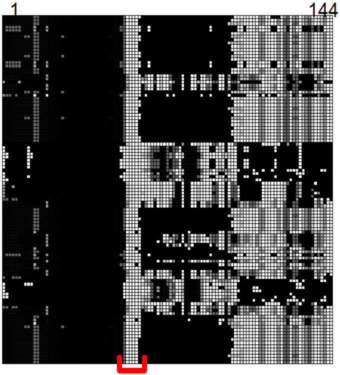
Classification of the household electric power consumption

Learned discriminant matching (CONSLEVEL)

M^*_W (Low)



M^*_B (Low vs. High)



Preliminary results

Table 2: k-Nearest Neighbor classification error rates

	k	D	DE	DTW
BME	1	**0.032**	0.165	0.130
BME	3	0.034	0.208	0.132
BME	5	0.062	0.234	0.136
BME	7	0.079	0.297	0.191
UMD	1	**0.055**	0.173	0.121
UMD	3	0.111	0.333	0.177
UMD	5	0.173	0.343	0.225
UMD	7	0.222	0.378	0.274
CONSLEVEL	1	**0.056**	0.306	**0.289**
CONSLEVEL	3	**0.044**	0.267	**0.261**
CONSLEVEL	5	**0.028**	0.233	**0.239**
CONSLEVEL	7	**0.017**	0.233	**0.233**
CONSSEASON	1	**0.094**	0.239	0.283
CONSSEASON	3	0.128	0.228	0.311
CONSSEASON	5	0.205	0.200	0.300
CONSSEASON	7	0.111	0.222	0.306
TRAJ	1	0.014	**0.012**	0.019
TRAJ	3	0.018	0.017	0.022
TRAJ	5	0.022	0.021	0.028
TRAJ	7	0.019	0.021	0.026

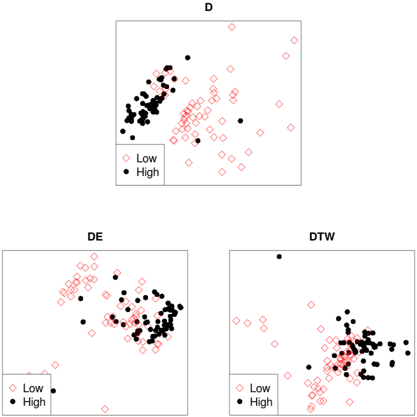
- Classes compactness/separability by MDS

Preliminary results

Table 2: *k*-Nearest Neighbor classification error rates

	k	D	DE	DTW
BME	1	**0.032**	0.165	0.130
	3	0.034	0.208	0.132
	5	0.062	0.234	0.136
	7	0.079	0.297	0.191
UMD	1	**0.055**	0.173	0.121
	3	0.111	0.333	0.177
	5	0.173	0.343	0.225
	7	0.222	0.378	0.274
CONSLEVEL	1	**0.056**	0.306	**0.289**
	3	**0.044**	0.267	**0.261**
	5	**0.028**	0.233	**0.239**
	7	**0.017**	0.233	**0.233**
CONSSEASON	1	**0.094**	0.239	0.283
	3	0.128	0.228	0.311
	5	0.205	0.200	0.300
	7	0.111	0.222	0.306
TRAJ	1	0.014	**0.012**	0.019
	3	0.018	0.017	0.022
	5	0.022	0.021	0.028
	7	0.019	0.021	0.026

- Classes compactness/separability by MDS



Preliminary results

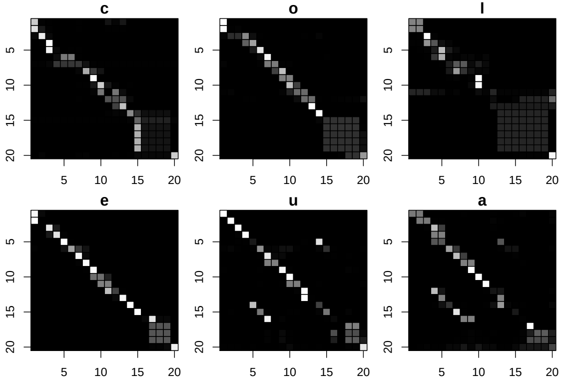


Figure: The learned discriminative matching for the characters "c'', ''o'', ''l'', "e", "u" and "a" of TRAJ data (UCI ML)