Chapter 4

Experiments

Sommaire

In this chapter, we evaluate the efficiency of our proposed $\mathrm{M^2TML}$ algorithm on public datasets for classification problems of univariate time series. First, we describe the datasets. Secondly, we detail the experimental protocol. Finally, we present and discuss the obtained results.

4.1 Description

The efficiency of the learned multi-modal and multi-scale dissimilarities D and $D_{\mathscr{H}}$ is evaluated through a $1-$NN classification on 30 public datasets[1][Keo+11]. The $1-$NN classifier is used to make the results comparable with the results of the UCR time series data mining archive. Time series comes from several fields (simulated data, medical data, electrical data, etc.), are from variable lengths (from small ($q = 24$) to long lengths ($q = 1882$)) and the number of classes to discriminate evolves between 1 and 37 classes. Note that some of the datasets have a small number of time series in the training set ($n < 30$) and others have a large number of time series in the training set ($n > 100$). The results using standard metrics (Euclidean distance, Dynamic time warping) show both easy and challenging classifications problems, the latter being opened for improvements. The datasets encompass time series that involve global or local temporal differences, require or not time warping, with linearly or non linearly separable neighborhoods, discussed later. Table 4.1 gives a description of the datasets considered in

[1] PowerCons: `https://archive.ics.uci.edu/ml/datasets/Individual+household+electric+power+consumption`, BME and UMD: `http://ama.liglab.fr/~douzal/tools.html`.

the experiments and Fig. 4.1 gives the temporal representation for some of the datasets. Note that for some datasets (*e.g.*, SonyAIBO, ECG200, FaceFour, PowerConsumption), it is visually difficult to discriminate the classes using one modality (value, behavior, frequential).

Dataset	Nb. Class	Nb. Train	Nb. Test	TS length
1 CC	6	300	300	60
2 GunPoint	2	50	150	150
3 CBF	3	30	900	128
4 OSULeaf	6	200	242	427
5 SwedishLeaf	15	500	625	128
6 Trace	4	100	100	275
7 FaceFour	4	24	88	350
8 Lighting2	2	60	61	637
9 Lighting7	7	70	73	319
10 ECG200	2	100	100	96
11 Adiac	37	390	391	176
12 FISH	7	175	175	463
13 Beef	5	30	30	470
14 Coffee	2	28	28	286
15 OliveOil	4	30	30	570
16 CinCECGtorso	4	40	1380	1639
17 DiatomSizeReduc	4	16	306	345
18 ECG5Days	2	23	861	136
19 FacesUCR	14	200	2050	131
20 InlineSkate	7	100	550	1882
21 ItalyPowerD	2	67	1029	24
22 MedicalImages	10	381	760	99
23 MoteStrain	2	20	1252	84
24 SonyAIBOII	2	27	953	65
25 SonyAIBO	2	20	601	70
26 Symbols	6	25	995	398
27 TwoLeadECG	2	23	1139	82
28 PowerCons	2	73	292	144
29 BME	3	300	1500	128
30 UMD	3	360	1440	150

Table 4.1: Dataset description giving the number of classes (Nb. Class), the number of time series for the training (Nb. Train) and the testing (Nb. Test) sets, and the length of each time series (TS length).

The results of the learned metrics D and $D_{\mathscr{H}}$ are compared to those of three *a priori* combined metrics $D_{Lin}, D_{Geom}, D_{Sig}$ (Eqs. 2.16, 2.17, 2.18) and five alternative uni-modal metrics covering:

1. The standard Euclidean distance d_A (Eq. 2.1) and Dynamic time warping DTW
2. The behavior-based measures d_B (Eq. 2.6) and $d_{B-\text{DTW}}$ its counterpart for asynchronous time series, that is d_B is evaluated once time series synchronized by dynamic programing
3. The frequential-based metric d_F (Eq. 2.4).

Table 4.2 recalls briefly the considered metrics in our experiments.

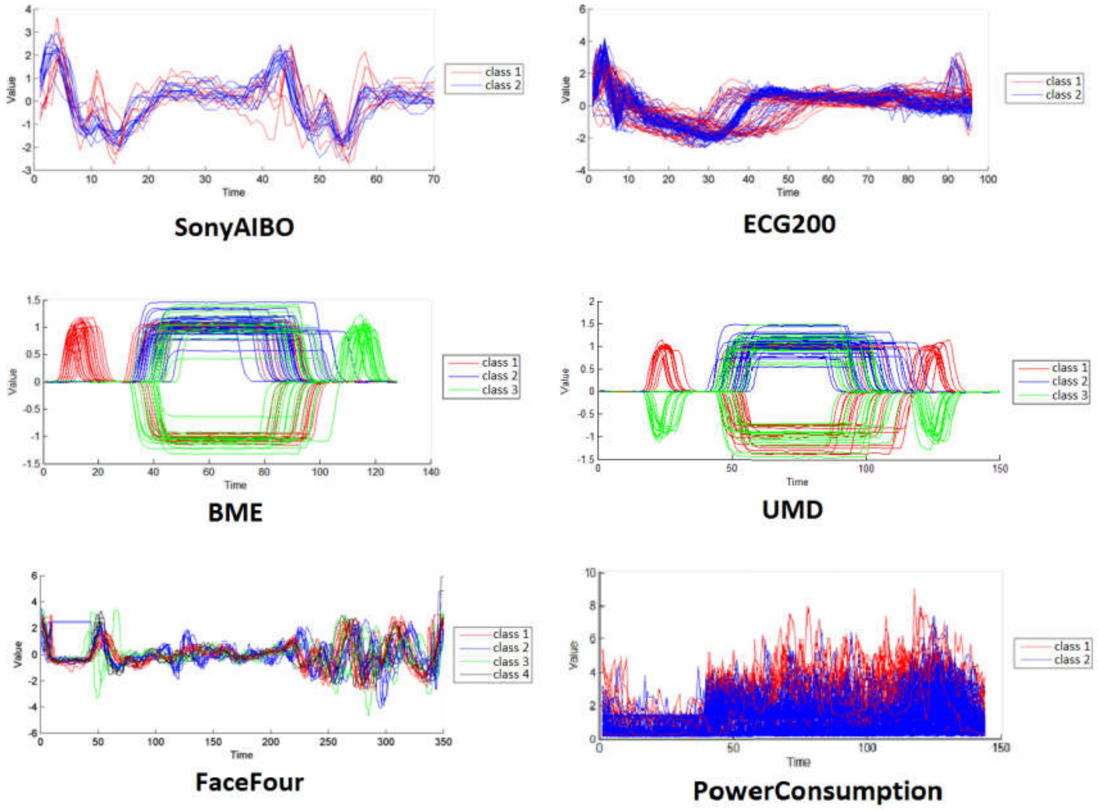


Figure 4.1: Temporal representation of some datasets (SonyAIBO, ECG200, BME, UMD, FaceFour, PowerConsumption) considered in the experiments.

Symbol	Name	Equation reference	Description
d_A	Value-based dissimilarity	Eq. 2.1	Euclidean distance
d_B	Behavior-based dissimilarity	Eq. 2.6	Behavior metric based on cort
DTW	Dynamic time warping	Eqs. 2.13 & 2.1	Euclidean distance after alignment
$d_{B-\text{DTW}}$	Behavior-based aligned dissimilarity	Eqs. 2.13 & 2.6	Behavior metric based on cort after alignment
d_F	Frequential-based dissimilarity	Eq. 2.4	Frequential based on Fourier transform
D_{Lin}	Linear combined metric	Eq. 2.16	Combines d_A and d_B (resp. DTW and $d_{B-\text{DTW}}$)
D_{Geom}	Geometric combined metric	Eq. 2.17	Combines d_A and d_B (resp. DTW and $d_{B-\text{DTW}}$)
D_{Sig}	Sigmoid combined metric	Eq. 2.18	Combines d_A and d_B (resp. DTW and $d_{B-\text{DTW}}$)
D	Linear learned metric	Eq. 3.36	M^{2}TML linear combined metric
$D_{\mathscr{H}}$	Non-linear learned metric	Eq. 3.37	M^{2}TML non-linear combined metric with a Gaussian kernel

Table 4.2: Considered metric in the experiments

The *a priori* combined metrics ($D_{Lin}, D_{Geom}, D_{Sig}$) relies, for synchronous (resp. asynchronous), on 2 log-normalized dissimilarities d_A, d_B (resp. DTW, $d_{B-\text{DTW}}$). The alternative metrics and the *a priori* combined metrics are evaluated as usual by involving the all time series elements (*i.e.* at the global scale). For D and $D_{\mathscr{H}}$, we consider a 21-dimensional embedding space $\mathscr{E}$ that relies, for synchronous (resp. asynchronous) data, on 3 log-normalized dissimilarities d_A^s, d_B^s (resp. DTW^s, $d_{B-\text{DTW}}^s$), and d_F^s, at 7 temporal granularities $s \in \{0, ..., 6\}$ obtained by binary segmentation, described in Section 3.2.

4.2 Experimental protocol

The different metrics can be split into two categories. For those without parameters to tune (d_A, DTW), the 1−NN classifier is applied directly on the test set. For those that require to tune parameters (d_B, $d_{B-\text{DTW}}$, D_{Lin}, D_{Geom}, D_{Sig}, D, $D_{\mathscr{H}}$), we recall briefly the grid search and cross-validation procedure (Section 1.1.2). When a learning algorithm requires to tune some parameters, to avoid overfitting, the training set can be divided into two sets: a learning and a validation set. The model is learnt for each combination of parameters (grid search) on the learning set and evaluated on the validation set. The model with the lowest error on the validation set is retained. An other alternative is cross-validation, which partitions the training set into v folds, performs the learning on one subset, and validates on the $v-1$ other subsets. To take into account of variability within the data, multiple rounds of cross-validation are performed using different partitions, and the validation results are averaged over the rounds. Note that for unbalanced datasets in classification problems, it is recommended to use stratified sampling . Table 4.3 resumes the parameter ranges for each metric. We recall that the parameters retained are those that:

- **First**, minimize the average classification error on the validation set.
- **Secondly**, in the case of multiple solutions leading to equal performances, the most discriminant one is retained (*i.e.*, making closer pull pairs and far away push pairs). Precisely, it minimizes the ratio $\frac{d_{intra}}{d_{inter}}$ where d_{intra} and d_{inter} stands respectively to the mean of all intraclass and interclass distances.

As D and $D_{\mathscr{H}}$ involves several parameters to be tuned, we detail hereafter the procedure. The combined metrics D and $D_{\mathscr{H}}$ (κ as the Gaussian kernel) are learned respectively under L_1 and L_2 regularization, using LIBLINEAR and LIBSVM libraries. The parameters are estimated on a validation set by line/grid search. A cross-validation and stratified sampling for unbalanced datasets are used. Particularly, for each couple $(r,\ \lambda)$ $r \in \{1, 4, 10\}$ and $\lambda \in \{0, 10, 30\}$, the pairwise SVM parameters (C, α, γ) are learned by grid search as indicated in Table 4.3. The temporal order r for the behavior-based metrics d_B is noise-dependent, typically 1 is retained for noise-free data. The parameter λ corresponds to the strength of the '*push*' term; precisely, if no, moderate or strong '*push*' is required during the training process, a λ value of 0, 10 and 30 is learned, respectively.

Dissimilarity	Parameter	Ranges	Description
d_B	r	$\{1, 2, 3, , \ldots, q-1\}$	Order of behavior-based metric
$D_{Lin}, D_{Geom}, D_{Sig}$	α	$\{0, 0.1, \ldots, 1\}$	Trade-off between value and behavior components
D, $D_{\mathscr{H}}$	λ	$\{0, 10, 30\}$	Strength of the '*push*' term
D, $D_{\mathscr{H}}$	C	$\{10^{-3}, 0.5, 1, 5, 10, 20, 30, ..., 150\}$	Parameter of SVM
D, $D_{\mathscr{H}}$	α	$\{1, 2, 3\}$	Size of the $m = \alpha.k$ neighborhood
$D_{\mathscr{H}}$	γ	$\{10^{-3}, 10^{-2}, \ldots, 10^3\}$	Parameter of the Gaussian kernel

Table 4.3: Parameter ranges

4.3 Results

Table 4.4 reports the 1-NN classification test errors when based on uni-modal metrics (first 5 columns), on three *a priori* combined metrics ($D_{Lin}, D_{Geom}, D_{Sig}$) and on D and $D_{\mathscr{H}}$. The results for each dataset that are statistically and significantly better than the rest are indicated in bold (z-test at 5% risk detailed in Section 1.1.3.a). The last column 'WARP' indicates the synchronous (✓) or asynchronous (×) data type.

Michèle trouve que ce serait bien de ranger les datasets par warp/non-warp et de classer des moins challenge au plus challenge pour aider la lecture.

Dataset	Alternative uni-modal metrics					A priori combinations			M²TML		WARP
	d_A	d_B	d_F	DTW	$d_{B-\mathrm{DTW}}$	D_{Lin}	D_{Geom}	D_{Sig}	$D(\lambda^*)$	$D_{\mathscr{H}}(\lambda^*)$	WARP
1 CC	0.120	0.113	0.383	**0.007**	0.027	**0.007**	**0.007**	**0.007**	**0.003** (0)	**0.007** (0)	✓
2 GunPoint	0.087	0.113	**0.027**	0.093	**0.027**	**0.027**	**0.027**	**0.040**	**0.020** (10)	**0.040** (10)	✓
3 CBF	0.148	0.140	0.382	**0.003**	**0.000**	**0.000**	**0.000**	**0.000**	0.031 (30)	**0.003** (0)	✓
4 OSULeaf	0.484	0.475	0.426	0.409	**0.265**	**0.264**	**0.264**	**0.322**	0.380 (0)	0.376 (0)	✓
5 SwedishLeaf	0.211	0.186	0.146	0.208	**0.109**	**0.115**	**0.110**	**0.125**	**0.110** (0)	**0.114** (0)	✓
6 Trace	0.240	0.240	0.140	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000** (0)	**0.010** (0)	✓
7 FaceFour	0.216	0.216	0.239	0.170	0.136	0.170	0.170	0.170	**0.000** (0)	0.034 (0)	✓
8 Lighting2	**0.246**	**0.246**	**0.148**	**0.131**	**0.213**	**0.131**	**0.131**	**0.131**	**0.148** (0)	**0.131** (0)	✓
9 Lighting7	**0.425**	**0.411**	**0.316**	**0.274**	**0.288**	**0.342**	**0.356**	**0.342**	**0.397** (0)	**0.233** (0)	✓
10 ECG200	**0.120**	**0.070**	0.160	0.230	0.190	**0.070**	**0.070**	**0.070**	**0.080** (0)	**0.080** (0)	×
11 Adiac	0.389	**0.297**	**0.261**	0.396	0.338	0.373	0.363	0.402	0.358 (0)	0.361 (0)	×
12 FISH	0.217	**0.149**	0.229	**0.166**	**0.137**	**0.109**	**0.137**	**0.126**	**0.149** (0)	0.240 (0)	✓
13 Beef	0.467	0.300	0.500	0.500	0.500	0.367	0.267	0.467	**0.033** (0)	0.267 (0)	×
14 Coffee	0.250	**0.000**	0.357	0.179	0.143	**0.000**	**0.000**	**0.071**	**0.000** (0)	**0.000** (10)	×
15 OliveOil	**0.133**	**0.133**	**0.167**	**0.200**	**0.100**	**0.133**	**0.133**	**0.133**	**0.167** (0)	**0.100** (10)	✓
16 CinCECGtorso	0.103	0.367	0.167	0.349	0.367	**0.094**	**0.094**	**0.093**	**0.092** (0)	**0.079** (0)	×
17 DiatomSizeR	0.065	0.076	0.069	**0.033**	**0.029**	**0.033**	**0.033**	**0.042**	**0.026** (0)	**0.029** (0)	✓
18 ECG5Days	0.203	0.153	**0.006**	0.232	0.236	0.203	0.203	0.203	**0.007** (10)	0.024 (0)	×
19 FacesUCR	0.231	0.227	0.175	0.095	0.102	0.098	0.098	0.099	**0.068** (10)	**0.059** (0)	✓
20 InlineSkate	**0.658**	**0.658**	0.675	**0.616**	**0.623**	**0.605**	**0.605**	**0.602**	0.733 (10)	**0.625** (0)	✓
21 ItalyPowerD	0.045	**0.028**	0.078	0.050	0.055	**0.028**	**0.028**	**0.030**	**0.028** (30)	**0.037** (10)	×
22 MedicalImages	0.316	0.313	0.345	**0.263**	0.290	**0.263**	**0.263**	**0.263**	**0.237** (0)	**0.236** (10)	✓
23 MoteStrain	**0.121**	0.263	0.278	0.165	0.171	0.260	0.248	0.188	0.185 (0)	0.153 (10)	✓
24 SonyAIBOII	**0.141**	**0.142**	**0.128**	0.169	0.194	**0.142**	**0.142**	**0.144**	**0.155** (0)	**0.131** (0)	×
25 SonyAIBO	0.305	0.308	0.258	0.275	0.343	0.308	0.308	0.293	**0.188** (0)	**0.165** (30)	×
26 Symbols	0.101	0.111	0.080	**0.050**	**0.043**	**0.051**	**0.050**	**0.052**	**0.034** (30)	**0.046** (30)	✓
27 TwoLeadECG	0.253	0.153	0.103	0.096	**0.008**	**0.005**	**0.005**	0.018	**0.006** (0)	0.016 (10)	✓
28 PowerCons	**0.366**	0.445	**0.315**	0.397	0.401	0.401	0.401	0.401	**0.318** (0)	**0.308** (0)	✓
29 BME	0.173	0.180	0.373	0.107	0.120	0.107	0.107	0.107	**0.040** (30)	**0.000** (10)	✓
30 UMD	0.194	0.222	0.299	0.118	**0.090**	0.111	0.111	0.118	0.104 (0)	**0.042** (0)	✓

Table 4.4: 1-NN test error rates for standard, *a priori* combined and M²TML measures.

4.4 Discussion

4.4.1 Comparison of the classification performances on the test set

From Table 4.4, we can see first that the 1-NN classification reaches the best results in:

1. Less than one-third of the data when based on d_A, d_B or d_F
2. Slightly more than one-third for DTW and $d_{B-\text{DTW}}$
3. Two-thirds (19 - 20 times on 30) when based on D_{Lin}, D_{Geom} and D_{Sig}
4. More than two-thirds (23 times on 30) when based on D or $D_{\mathscr{H}}$.

Particularly, note that for nearly all datasets for which an uni-modal metric succeeds, the M^{2}TML metrics succeed similarly or lead to equivalent results. However, for several challenging datasets (*e.g.* FaceFour, Beef, FaceUCR, SonyAIBO, BME), M^{2}TML realizes drastic improvements, to the best of our knowledge never achieved before for these challenging public data. For instance, the impressive scores of 3% obtained for Beef against an error rate varying from 30% to 50% for alternative metrics, and of 0% obtained for FaceFour v.s. 13% to 23% for alternative metrics. Finally, D and $D_{\mathscr{H}}$ are all the more outperforming if only compared to the standard metrics d_A (the Euclidean distance) and DTW.

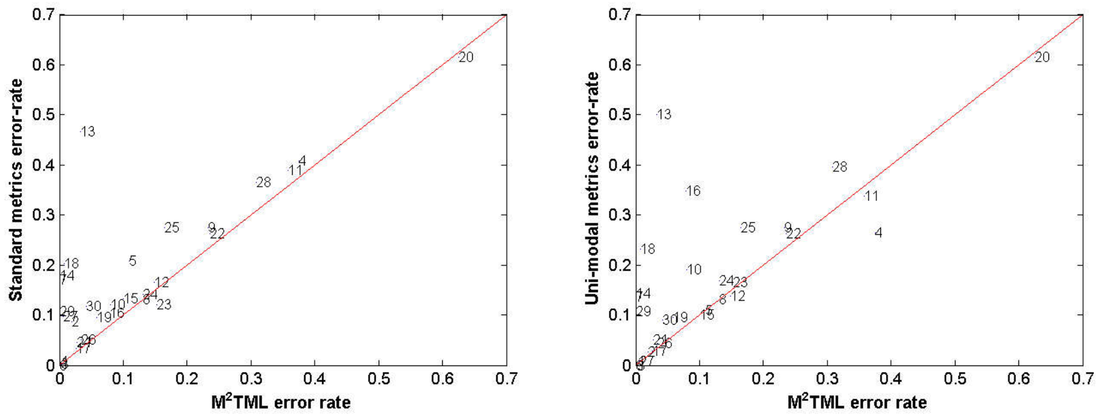
If we compare the *a priori* combined metrics (D_{Lin}, D_{Geom}, D_{Sig}) based on only the unimodal metrics involved in the combination (either d_A and d_B or DTW and $d_{B-\text{DTW}}$), we observe that *a priori* combined metrics achieved on two-third of the data with an equivalent or better score. Compared to the learnt metrics (D, $D_{\mathscr{H}}$), the results are globally similar except for 8 datasets where the learned metrics perform better (FaceFour, Beef, ECG5Days, FaceUCR, SonyAIBO, PowerCons, BME, UMD) and one where the *a priori* combined metrics perform better (OSULeaf). Note that the combined metric D_{Sig} is limited to two components and can't be easily extend to other metrics in its combination. D_{Lin} and D_{Geom} could be easily extended and a proposition could be:

$$D_{Lin}(\mathbf{x}_i, \mathbf{x}_j) = \sum_{h=1}^{p} \alpha_h d_h(\mathbf{x}_i, \mathbf{x}_j) \tag{4.1}$$

$$D_{Geom}(\mathbf{x}_i, \mathbf{x}_j) = \prod_{h=1}^{p} \alpha_h d_h(\mathbf{x}_i, \mathbf{x}_j) \tag{4.2}$$

However, by considering p metrics d_h the resulting models requires to optimize p parameters. The grid search to find the best parameters α_h can become time consuming.

In the second part, we perform a graphical analysis for a global comparison on the whole datasets. In Fig. 4.2-a, each dataset is projected according to, on the x-axis its best error rate obtained for D and $D_{\mathscr{H}}$, and on y-axis its best performance w.r.t the standard metrics d_A and DTW. In Fig. 4.2-b, the y-axis is related to the best error rate w.r.t DTW and $d_{B-\text{DTW}}$, the two most performant uni-modal metrics. For both plots we can note that the datasets are principally projected above the first bisector, indicating higher error rates mostly obtained for alternative metrics than for M^{2}TML. For the less challenging datasets (near the origin of each graph), although almost projected near the bisector denoting equal performances for the compared metrics, M^{2}TML still bring improvements with projections clearly positioned above the bisector. Finally, from Fig. 4.2 (b) we can see that M^{2}TML metrics perform significantly lower than $d_{B-\text{DTW}}$ on OSUleaf, while InlineSkate dataset remains challenging for all studied metrics.



(a) (b)

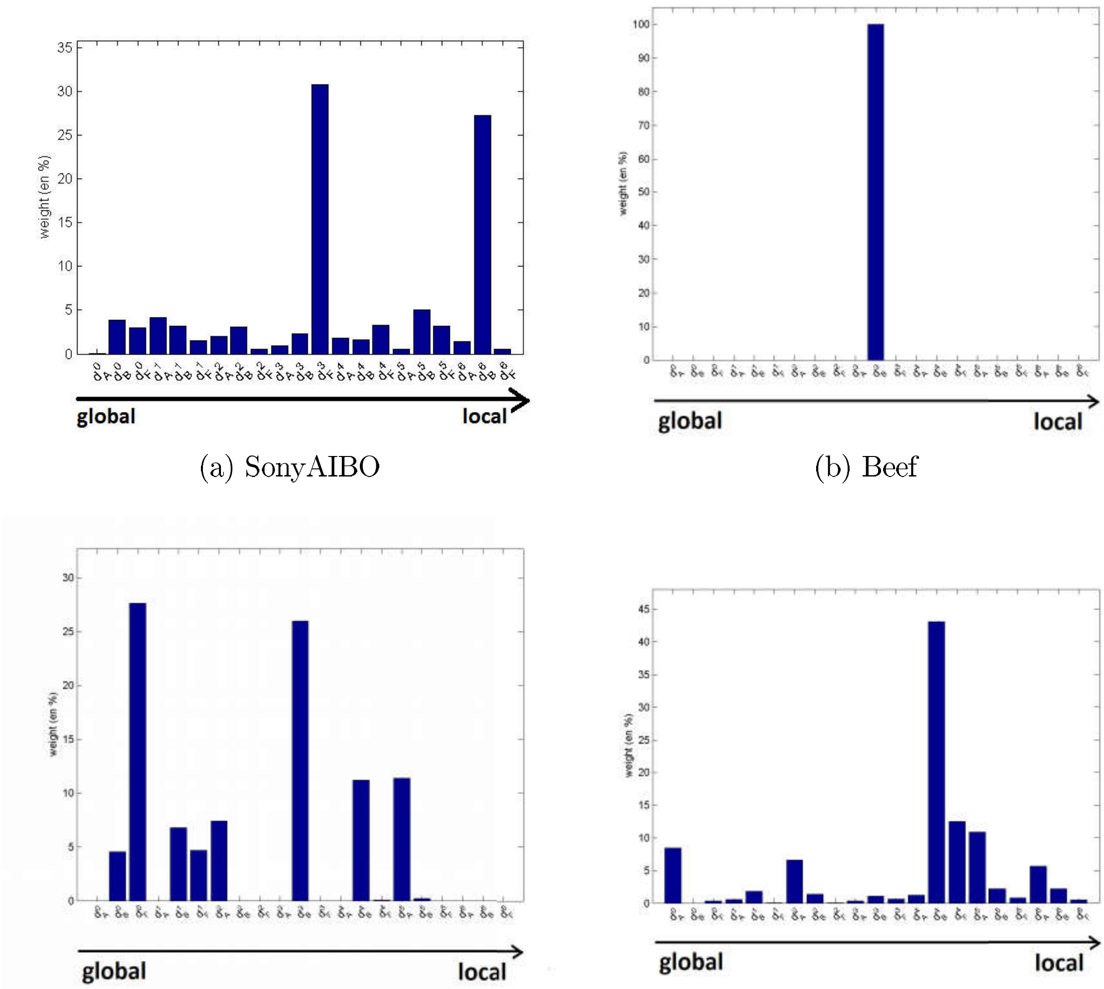
Figure 4.2: (a) Standard (Euclidean distance d_A and DTW) *vs.* M^{2}TML (D and $D_{\mathscr{H}}$) metrics. (b) Best Uni-modal (DTW and $d_{B-\text{DTW}}$) *vs.* M^{2}TML (D and $D_{\mathscr{H}}$) metrics.

4.4.2 Analysis of the discriminative features

For the learnt metric D, thanks to the L_1 regularization, the learned SVM reveals the features that most differentiate pull from push pairs. Table 4.5 shows the sparse, muti-modal and multi-scale potential of M^{2}TML approach. It gives for each dataset, the weights of the top five 'discriminative' features that contribute to the definition of D. For instance, for FaceFour D reaches the 0% by combining, in the order of importance, the behavior, frequential and amplitude modalities, at the global (I^0) and local (I^4, I^5, I^2) scales. For Beef, besides the impressive error rate of 3%, the learned model is very sparse as D involves only the behavior modality based on the segment I^3 (d_B^3). Similarly for Coffee, the obtained 0% involves only the behavior and frequential modalities at several scales. Note that if we look at only the most discriminative feature (1st column), the M^{2}TML method helps to localize discriminative modality and a specific temporal scale that could not be easily guessed *a priori* (*e.g.*, Lightning7: $d^6_{B-\text{DTW}}$, OliveOil: d^5_F, TwoLeadECG: $d^4_{B-\text{DTW}}$)

In Fig. 4.3, we plot the weights of all features for SonyAIBO, Beef, CincECGtorso and FaceFour cases as an example. It illustrates both the sparsity of the $\mathrm{M^2TML}$ approach (Beef, CincECGtorso and FaceFour) and the ability of the algorithm to combine all the features into the metric D (SonyAIBO). In particular, the approach is able to either select one single feature (Beef) or combine several selected features (CinCECGTorso, FaceFour). Fig. 4.4 illustrates the temporal locations of the most discriminative features for these datasets. Note that from looking at the temporal representation, it is not easy to determine *a priori* which modality (value, behavior, frequential) and at which temporal scale is the most discriminative feature to separate the classes.

In summary, we can emphasize that for almost all datasets, the definition of D involves no more than five features (the most contributive ones), that assesses not only the model's sparsity but also the representativeness of the revealed features.



(a) SonyAIBO (b) Beef

(c) CinC ECG torso (d) FaceFour

Figure 4.3: $\mathrm{M^2TML}$ feature weights for 4 datasets.

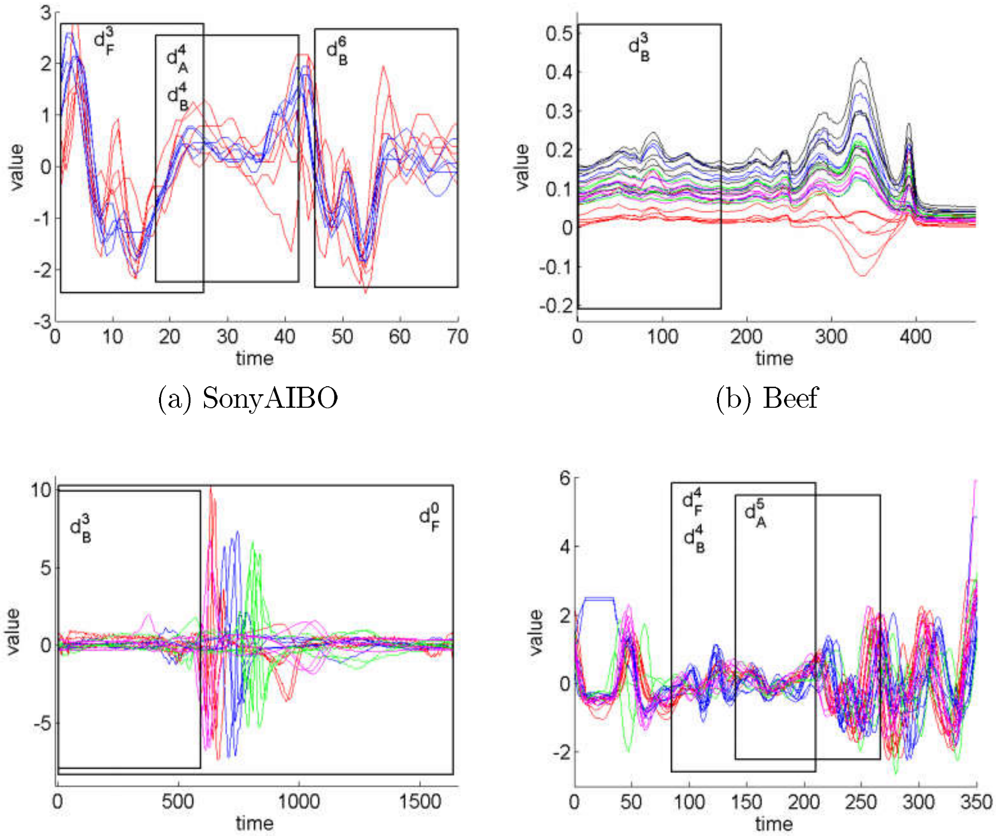
Dataset	Feature weights (%)				
CC	DTW^2 (56.3%)	d_F^0 (18.8%)	d_F^4 (5.9%)	d_F^1 (4%)	$d_{B-\text{DTW}}^0$ (3.2%)
GunPoint	$d_{B-\text{DTW}}^0$ (42.1%)	DTW^5 (10.9%)	DTW^1 (10.2%)	$d_{B-\text{DTW}}^6$ (8.4%)	d_F^4 (8.1%)
CBF	DTW^4 (56.5%)	d_F^3 (43.4%)	$d_{B-\text{DTW}}^1$ (0.2%)	DTW^0 (0%)	$d_{B-\text{DTW}}^0$ (0%)
OSULeaf	DTW^2 (23.7%)	d_F^0 (19.5%)	$d_{B-\text{DTW}}^0$ (14.6%)	$d_{B-\text{DTW}}^2$ (9.4%)	DTW^1 (9%)
SwedishLeaf	d_F^0 (21.5%)	DTW^0 (15.9%)	$d_{B-\text{DTW}}^0$ (15.2%)	DTW^6 (11.5%)	$d_{B-\text{DTW}}^1$ (6.1%)
Trace	DTW^0 (58.3%)	DTW^6 (6.9%)	$d_{B-\text{DTW}}^0$ (5.8%)	DTW^2 (5.6%)	DTW^5 (5.5%)
FaceFour	$d_{B-\text{DTW}}^4$ (44.5%)	d_F^4 (12.7%)	DTW^5 (11.1%)	DTW^0 (8.3%)	DTW^2 (6.4%)
Lighting2	$d_{B-\text{DTW}}^0$ (30.4%)	DTW^6 (18.7%)	d_F^1 (16.5%)	$d_{B-\text{DTW}}^6$ (13.4%)	DTW^0 (10.7%)
Lighting7	$d_{B-\text{DTW}}^6$ (87.4%)	d_F^6 (8.6%)	$d_{B-\text{DTW}}^5$ (4%)	-	-
ECG200	d_B^0 (89.6%)	d_B^6 (2.4%)	d_A^3 (2.3%)	d_B^1 (2.2%)	d_B^4 (2%)
Adiac	d_F^0 (79.2%)	d_B^4 (13.8%)	d_A^4 (3.5%)	d_F^5 (1.7%)	d_B^5 (1.2%)
FISH	$d_{B-\text{DTW}}^5$ (17.9%)	d_F^0 (10.5%)	$d_{B-\text{DTW}}^6$ (9.9%)	$d_{B-\text{DTW}}^4$ (8.3%)	$d_{B-\text{DTW}}^3$ (7.8%)
Beef	d_B^3 (100%)	-	-	-	-
Coffee	d_B^2 (22.4%)	d_F^4 (20.1%)	d_B^6 (14.6%)	d_B^0 (8.1%)	d_F^5 (7%)
OliveOil	d_F^5 (97%)	$d_{B-\text{DTW}}^2$ (3%)	-	-	-
CinCECGtorso	d_F^0 (38.4%)	d_A^5 (13.1%)	d_B^4 (11.5%)	d_F^1 (11.2%)	d_A^2 (9.8%)
DiatomSizeR	d_F^5 (39.1%)	d_F^0 (36%)	$d_{B-\text{DTW}}^4$ (24.9%)	-	-
ECG5Days	d_B^5 (59.5%)	d_B^6 (32.3%)	d_A^4 (3.9%)	d_B^2 (3.1%)	d_B^4 (1.2%)
FacesUCR	d_F^2 (21.5%)	$d_{B-\text{DTW}}^0$ (19.5%)	d_F^4 (16.7%)	DTW^0 (12.6%)	$d_{B-\text{DTW}}^2$ (8.6%)
InlineSkate	d_F^4 (42.5%)	DTW^5 (22.8%)	DTW^4 (17.6%)	DTW^2 (6.7%)	$d_{B-\text{DTW}}^6$ (5.9%)
ItalyPowerD	d_B^6 (68.7%)	d_B^0 (25.9%)	d_B^3 (5.2%)	d_B^4 (0.2%)	-
MedicalImages	$d_{B-\text{DTW}}^1$ (53.3%)	d_F^3 (12.9%)	$d_{B-\text{DTW}}^2$ (10.7%)	$d_{B-\text{DTW}}^3$ (10.1%)	$d_{B-\text{DTW}}^0$ (3.8%)
MoteStrain	$d_{B-\text{DTW}}^5$ (93.2%)	$d_{B-\text{DTW}}^6$ (6.8%)	-	-	-
SonyAIBOII	d_B^3 (100%)	-	-	-	-
SonyAIBO	d_F^3 (30.8%)	d_B^6 (27.3%)	d_B^5 (5%)	d_A^1 (4.1%)	d_B^0 (3.9%)
Symbols	$d_{B-\text{DTW}}^0$ (45.6%)	$d_{B-\text{DTW}}^6$ (35.3%)	$d_{B-\text{DTW}}^5$ (19%)	DTW^0 (0.1%)	-
TwoLeadECG	$d_{B-\text{DTW}}^4$ (60%)	d_F^1 (12%)	DTW^4 (11.4%)	$d_{B-\text{DTW}}^6$ (7.6%)	$d_{B-\text{DTW}}^1$ (4.2%)
PowerCons	d_F^0 (26.1%)	DTW^0 (20.3%)	d_F^1 (19.3%)	$d_{B-\text{DTW}}^0$ (6.1%)	d_F^2 (5.1%)
BME	$d_{B-\text{DTW}}^0$ (75.2%)	d_F^4 (15.5%)	$d_{B-\text{DTW}}^2$ (5.8%)	$d_{B-\text{DTW}}^1$ (1.9%)	d_F^1 (0.7%)
UMD	$d_{B-\text{DTW}}^0$ (99.8%)	$d_{B-\text{DTW}}^5$ (0.2%)	-	-	-

Table 4.5: Top 5 multi-modal and multi-scale features involved in D

4.4.3 Effect on the neighborhood before and after learning

In the last part, we compare the global effect of the alternative and M^2TML metrics on the 1-NN neighborhood distribution and class discrimination. For that, a multidimensional scaling (MDS)[2] is used to visualize the distribution of samples according to their pairwise dissimilarities.

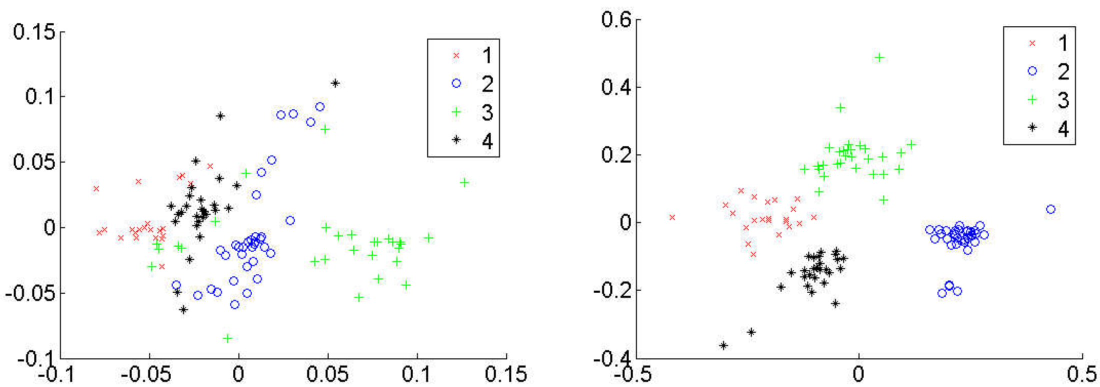
[2]Matlab function: mdscale for metrics and non metrics



(a) SonyAIBO (b) Beef

(c) CinC ECG torso (d) FaceFour

Figure 4.4: Temporal representation of the top M^{2}TML feature weights for 4 datasets.



(a) FaceFour ($d_{B-\mathrm{DTW}}$, stress: 20,1%) (b) FaceFour (D, stress: 17,2%)

Figure 4.5: MDS visualization of the $d_{B-\mathrm{DTW}}$ (Fig. a) and D (Fig. b) dissimilarities for FaceFour

We briefly recall the principle of MDS. MDS is a method of visualizing the proximity between samples in a dataset. A MDS algorithm aims to place each sample in P-dimensional space (in general, $P = 2$ or 3) such that the between-object distances are preserved as well as possible. Classical MDS takes an input matrix giving dissimilarities between pairs of samples and outputs a coordinate for each sample whose configuration minimizes a loss function called stress. Note that the MDS representation has no link with the dissimilarity space representation whose dimensions are basic temporal metrics.

For FaceFour, Fig. 4.5 shows the first obtained plans and their corresponding stresses, the classes being indicated in different symbols and colors. We can see distinctly the effect of the learned D that leads to more compact and more isolated classes with robust neighborhoods for 1-NN classification (*i.e.* closer pull pairs and far away push pairs) than the best alternative metric $d_{B-\mathrm{DTW}}$ that shows more overlapping classes and heterogeneous neighborhoods.

4.5 Conclusion of the chapter

The large conducted experiments and the impressive performances obtained attest the efficiency of the learned M^{2}TML metrics for time series nearest neighbors classification. Finally, let us underline the merit of the M^{2}TML solution, that not only leads to equivalent or better performances from the standard metrics (Euclidean distance, Dynamic time warping), but also provides a comprehensive and fine-grained information about which modalities are mostly discriminant, how they should be combined and precisely at which temporal granularity.