Adaptive split test for multivariate time series classification trees

Ahlame Douzal-Chouakria, Cécile Amblard

Adaptive split test for multivariate time series classification trees

Ahlame Douzal Chouakria[1], Cécile Amblard[1]

LIG (Lab. d'Informatique de Grenoble)
BP 53 - 38041 Grenoble cedex 9, France
ahlame.douzal,cecile.amblard@imag.fr

Abstract : This paper proposes an extension of the classification trees to time series input variables. A new split criterion based on time series proximities is introduced. First, it relies on an adaptive (i.e., parametrized) time series metric to cover both behavior and values proximities. The metric's parameters may change from one internal node to another to best bisect the set of time series. Second, it involves the automatic extraction of the most discriminating sub-sequences. The proposed time series classification tree is applied to a wide range of datasets: public and new, real and synthetic, univariate and multivariate data. We show, through the carried out experiments, that the proposed tree outperforms temporal trees using standard time series distances, and leads to good performances compared to other competitive time series classifiers.

Keywords: time series proximity measures, classification trees, learning metric

1. Introduction

Time series classification has prompted extensive research in the last few years. A first category of proposals consists of mapping time series to a new description space where conventional classifiers can be applied. Signal processing or statistical tools are commonly used to project time series into a given functional basis space. For instance, this projection can be performed by a Fourier or Wavelet transform, a polynomial or an ARIMA approximation. Standard classifiers are then applied on the fitted basis coefficients (e.g., Garcia-Escudero & Gordaliza (2005), Serban & Wasserman (2005)). A second class of works proposes new heuristics starting generally with the time

series segmentation to extract prototypes that best characterize the time series classes. The prototypes, defined by a set of sub-sequences, regions of values, etc., are then described by a set of features where standard classifiers can be applied (e.g., Kudo *et al.* (1999), Geurts & Wehenkel (2005)). This paper focuses on a distance-based approach to extend classification trees to temporal data. We propose a new time series split criterion characterized by, on the one hand, the use of an adaptive metric to cover both behavior and values proximities. This metric may change from one node to another according to the set of time series to divide. On the other hand, the proposed split involves an automatic extraction of the most discriminating sub-sequences. We show, through the carried out experiments, that the proposed tree outperforms temporal trees using standard time series distances, and leads to good performances compared to some competitive time series classifiers.
The rest of the paper is organized as follows. In the next section, we discuss two distance-based temporal trees proposed in Yamada *et al.* (2003) and Balakrishnan & Madigan (2006). In Section 3., the major metrics for time series are presented in a novel unified formalism. Section 4. presents the new time series classification tree, gives the main algorithms and discusses their complexity. In Section 5., the proposed classification tree is performed on six public and three new simulated datasets. The induced trees are compared to temporal trees using standard distances.

2. Related works

In this section, we describe two temporal classification trees proposed in Yamada *et al.* (2003) and Balakrishnan & Madigan (2006) . Both works build binary classification trees where internal nodes are labeled by one or two time series. Proposed classifiers are mainly based on new split tests to best bisect the set of time series within internal nodes. Yamada et al. Yamada *et al.* (2003) propose two split tests. The first one, called the *Standard-example* split test, selects through an exhaustive search one existing time series (called the standard time series), leading to division with a maximum purity gain ratio. The first child node is composed of time series whose distance to the standard time series is less than a given threshold, while the second child node contains the remaining time series. If more than one standard time series provides the largest value of the purity gain ratio, a class isolation criterion is used to select the split that exhibits the most dissimilar child nodes. The second proposed split test, called the *Cluster-example* split test, looks through an exhaustive

search for a couple of standard time series. The bisection is constructed by assigning each time series to the nearest standard time series. Similarly, the purity gain ratio and the class isolation criterion are used to select the best split test. For both split tests, the dynamic time warping is considered as the time series proximity measure. Balakrishnan and Madigan Balakrishnan & Madigan (2006) look for a couple of reference time series that best bisects the set of time series according to a clustering goodness criterion. For this, a kmeans algorithm is used, it ensures a partitioning that optimizes clustering criteria, namely the compactness and isolation of the clusters, but not their purity. To alleviate this problem, authors perform several times the kmeans clustering and select the partition that gives the highest Gini index. The centers of the clusters define the pair of reference time series of the split test. For the time series proximities, both the Euclidean distance and the dynamic time warping are used to compare the efficiency of the obtained classification trees.

In summary, the *Cluster-example* split test of Yamada et al. Yamada *et al.* (2003) and the one proposed by Balakrishnan and Madigan Balakrishnan & Madigan (2006) are highly similar. The former, first looks for a set of time series bisections with the highest purity clusters (i.e., here the highest Gini index) and then picks the one optimizing some clustering criteria (i.e., maximizing the separability of the clusters), whereas the latter, first looks for a set of splits optimizing clustering criteria (i.e., kmeans criteria) and then selects the one exhibiting the highest purity clusters (i.e., maximizing the Gini index). When giving priority to a clustering criterion instead to the purity of the clusters, the split test may fail to select bisections of lower clustering criteria but of higher purity.

Let us make some remarks about the above proposed split tests. First, as for many distance-based approaches, the Euclidean distance and the dynamic time warping are considered for time series proximities. These standards measures are value-based metrics and ignore the time series behaviors, as discussed in Section 3.. Second, the proposed splits involve the same metric to divide all the nodes, whereas the time series peculiarities may change from one node to another. Finally, the time series distances are calculated by using the whole time series values, even though the discrimination is determined by some sub-sequences.

3. Time series metrics

We present briefly, in a unified formalism, three categories of time series metrics (deeply detailed in Douzal-Chouakria & Amblard (2012)). The first category relies on two standard values-based metrics: the dynamic time warping and the Euclidean distance. In the second category, we recall the definition of the correlation coefficient and the temporal correlation coefficient, two behavior-based metrics. In the third category, we present a model to cover both behavior and values time series components. In particular, an extension of the Euclidean distance and of the dynamic time warping are provided to cover both behavior and values proximities. Let $S_1 = (u_1, ..., u_p)$ and $S_2 = (v_1, ..., v_q)$ be two time series of p and q values observed at the time instants $(t_1, ..., t_p)$ and $(t'_1, ..., t'_q)$, respectively. A mapping r between S_1 and S_2 is defined as a sequence of m pairs of observations $((u_{a_1}, v_{b_1}), (u_{a_2}, v_{b_2}), ..., (u_{a_m}, v_{b_m}))$, with $a_i \in \{1, .., p\}$, $b_i \in \{1, ..q\}$, and $i \in \{1, .., m-1\}$ obeying to the order constraints:

$$\begin{aligned} a_1 &= 1,\ a_m = p,\ a_{i+1} = a_i \text{ or } a_i + 1 \text{ and,} \\ b_1 &= 1,\ b_m = q,\ b_{i+1} = b_i \text{ or } b_i + 1. \end{aligned}$$

with $m \in [max(p, q), p + q - 1]$. Let R be a subset of such mappings, satisfying possibly some additional constraints, and $c(r)$ ($r \in R$) be the mapping cost function measuring the distance between the coupled values in r. A unified formalism of the time series proximity measures, denoted $dUnif$, may be presented as an optimization problem minimizing the cost function $c(r)$ on the search space R:

$$dUnif_{(c,R)}(S_1, S_2) = \min_{r \in R} c(r). \tag{1}$$

3.1. Values-based metrics

For the cost function definition $c(r) = \sum_{i=1}^{m} |u_{a_i} - v_{b_i}|$, $dUnif_{(c,R)}$ (Eq. 1) leads to the standard dynamic time warping Kruskall & Liberman (1983):

$$d_{Dtw}(S_1, S_2) = \min_{r \in R}(\sum_{i=1}^{m} |u_{a_i} - v_{b_i}|). \tag{2}$$

In the case of times series of the same length ($m = p = q$), and for the cost function definition $c(r) = (\sum_{i=1}^{m}(u_i - v_i)^2)^{\frac{1}{2}}$ minimized on $R = \{r_0\}$, $dUnif_{(c,R)}$ gives the Euclidean distance, with:

$$r_0 = ((u_1, v_1), (u_2, v_2), ..., (u_m, v_m)) \tag{3}$$

$$d_E(S_1, S_2) = c(r_0) = (\sum_{i=1}^{m}(u_i - v_i)^2)^{\frac{1}{2}}. \quad (4)$$

The above cost functions $c(r)$ involve the differences between the aligned values, without allowance for the values neighborhoods.

3.2. Behavior-based metrics

One may define two time series S_1 and S_2 as similar on behavior if at any observed period $[t_i, t_{i+1}]$, they increase or decrease simultaneously with the same growth rate. In contrast, they are considered as opposite on behavior if at any observed period $[t_i, t_{i+1}]$ where S_1 increases, S_2 decreases and vice-versa with the same growth rate (in absolute value). Until nowadays, many applications in different domains use the Pearson correlation coefficient as a behavior proximity measure between signals. The correlation coefficient assumes the data independent as based on the differences between all the pairs of values observed at $[t_i, t_{i'}]$; whereas the behavior proximity needs only to capture how time series behave at $[t_i, t_{i+1}]$. Thus, the correlation coefficient is biased by all the remaining pairs of values observed at $[t_i, t_{i'}]$ with $|i - i'| > 1$. To overcome the limitations of the Pearson correlation coefficient, the temporal correlation coefficient is used, which reduces the Pearson correlation coefficient to the first order differences:

$$Cort(S_1, S_2) = \frac{\sum_i (u_{i+1} - u_i)(v_{i+1} - v_i)}{\sqrt{\sum_i (u_{i+1} - u_i)^2}\sqrt{\sum_i (v_{i+1} - v_i)^2}} \quad (5)$$

with $Cort(S_1, S_2)$ belonging to $[-1, 1]$. The value $Cort(S_1, S_2) = 1$ indicates that S_1 and S_2 exhibit similar behavior. The value $Cort(S_1, S_2) = -1$ indicates that S_1 and S_2 exhibit opposite behavior. Finally, $Cort(S_1, S_2) = 0$ expresses that the growth rates S_1 and S_2 are stochastically, linearly independent, thereby identifying time series of different behaviors, namely that they are neither similar nor opposite.

3.3. Values and behavior based metrics

To define a proximity measure covering both the behavior and values components, we consider the cost function $c_k(r)$ Douzal-Chouakria *et al.* (2009) modulating the values-based proximity according to the behavior-based proximity:

$$c_k(r) = \frac{2}{1 + exp(k\ Co(r))}.c(r), \quad k \geq 0 \quad (6)$$

where $c(r)$ and $Co(r)$ define, respectively, values-based (e.g., Eqs. (2), (4)) and behavior-based (e.g. Eqs. (5)) cost functions. The parameter k defines the relative contributions of the behavior and values components to $c_k(r)$. For a mapping cost function $c_k(r)$ covering only the values proximity component (i.e., ignoring the behavior component), k is fixed to 0 and $c_{k=0}(r) = c(r)$. On the other hand, for $k >= 6$, $c_{k=6}(r)$ completely includes the behavior proximity component. Hence, if $Co(r) = 1$, then $c_{k=6}(r) \approx 0$, which means that if two time series are of similar behavior, the cost function is reduced to zero regardless of the value of $c(r)$. If $Co(r) = -1$, then $c_{k=6}(r) \approx 2c(r)$; this corresponds, in the case of time series of opposite behaviors, to penalty of a factor of 2 to $c(r)$. Finally, if $Co(r) = 0$, then $c_{k=6}(r) \approx c(r)$, which means that in the case of time series of different behaviors, the mapping cost $c_{k=6}(r)$ is determined by the only available information $c(r)$.

Based on the cost function $c_k(r)$, the definition of the adaptive dissimilarity covering both values and behavior proximities Douzal-Chouakria *et al.* (2009) is:

$$D_k(S_1, S_2) = \min_{r \in R}\left(\frac{2}{1 + exp(k\ Co(r))} c(r)\right).$$

4. Time series classification trees

In this section, we present a new split test for multivariate time series classification trees, characterized by two additive values and deeply detailed in Douzal-Chouakria & Amblard (2012). First, the use of an adaptive metric which may change from one internal node to another to best bisect the set of time series. Second, the involvement of the automatic extraction of the most discriminating sub-sequences. Let $\{s_1, ..., s_N\}$ be a set of N multivariate time series partitioned into C classes, and $I_1, ..., I_N$ ($I_i = [1, T_i]$) their respective observation intervals. Before building the classification tree, time series are preprocessed to make them of equal length $I = [1, T]$, and pairwise time series dissimilarities computed.

4.1. Time series length normalization

To make the time series of the same length, two cases have to be considered. For data allowing time delays, time series are simply resampled by a linear interpolation to make them of equal length $I = [1, T]$. In the case of data that do

not allow time delays, the smallest observation period $I = min(I_1, ..., I_N)$ is considered; and linear interpolations may be used to resample the data within I.

4.2. Time series split ($TSSplit$) test algorithms

To split a given node S composed of a set of time series, the procedure $TSSplit(S, I, \alpha)$ is called with as input parameters: the set $S = \{s_1, ..., s_N\}$ of time series to bisect, the observation interval $I = [1, T]$, and a rate α needed for the discriminant sub-sequences search. In $TSSplit(S, I, \alpha)$ (Algorithm 1), a first call to $AdaptSplit(S, I)$ is performed to determine the best split of S involving the adaptive metric D_k evaluated on I.
The main idea behind the procedure $AdaptSplit(S, I)$ (Algorithm 2) is that, given a value of the parameter $k \in [0, 6]$ and two time series (l, r) from $S \times S$, a bisection of S, denoted $\sigma(l, r, k, I)$, is obtained by assigning each time series $ts \in S$ to the left node if it is closest to the time series l than to r, namely if $D_k(ts, l) \leq D_k(ts, r)$, and to the right node otherwise (see Figure 1). To determine the best split, several values of the triplet (l, r, k) are explored to find the bisection exhibiting the minimum impurity Gini index. As output, $AdaptSplit(S, I)$ returns the best split $\sigma(l^I_*, r^I_*, k^I_*, I)$ and its impurity Gini index $GI(\sigma(l^I_*, r^I_*, k^I_*, I))$.
The best split $\sigma(l^I_*, r^I_*, k^I_*, I)$ is obtained by comparing the time series proximities according to their observations within I. In the case of time series differentiation induced by some sub-sequences instead of implicating all the observations of I, the split $\sigma(l^I_*, r^I_*, k^I_*, I)$ may fail to reach higher purity classes. To alleviate this limitation, $DichoSplit$ allows us to determine, through a dichotomy search between left and right sub-sequences of I, those entailing a bisection of lower impurity than $\sigma(l^I_*, r^I_*, k^I_*, I)$.
The $DichoSplit(S, \sigma(l^I_*, r^I_*, k_*, I), e_I, \alpha)$ (Algorithm 3) is called with as input parameters: the set of time series S, the best split $\sigma(l^I_*, r^I_*, k^I_*, I)$ of S obtained by comparing time series over I, its corresponding impurity Gini index error named e_I, and the rate $\alpha \geq 0$ needed to define the boundaries of the left I_L and right I_R sub-intervals of I. Two calls to $AdaptSplit$ are performed to split S based on the observations of I_L and I_R, respectively.
If the impurity Gini index is not improved ($e_I \leq \min(e_{I_L}, e_{I_R})$), all the observations within I are needed to best discriminate time series. $DichoSplit$ stops and returns the split $\sigma(l^I_*, r^I_*, k^I_*, I)$. However, if the impurity Gini index is improved by at least one of the splits based on I_L or I_R, a call to $DichoSplit$

is pursued with the most discriminative sub-interval.

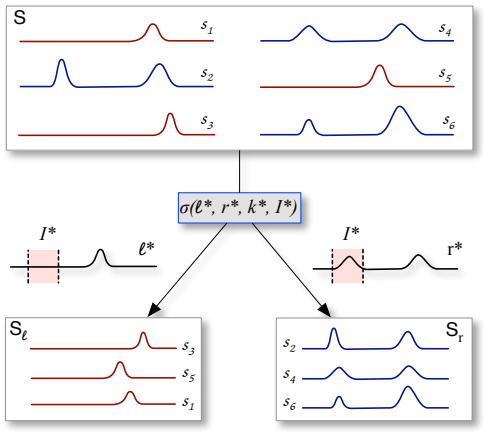

Figure 1: The adaptive time series split test

4.3. Algorithms specifications

In $AdaptSplit$, the explored splits $\sigma(l, r, k, I)$ (Algorithm 2, line 3) rely on the adaptive metric D_k, for seven values of $k \in [0, 6]$.
In $DichoSplit$, a value of $\alpha = 0.6$ is taken to divide I into the left and right covering intervals, this allows to cover discriminating sub-sequences in the central region of I. Of course, other values of α may be taken to divide I either into disjoint ($\alpha = 0.5$) or more covering ($\alpha > 0.6$) left and right sub-intervals. Once the two calls to $AdaptSplit$ based on I_L and I_R achieved (Algorithm 3 lines 4 and 5), some options are taken when faced with the equality cases. First, in the case of $e_I = e_{I_L} = e_{I_R}$, $DichoSplit$ (Algorithm 3, line 6) stops and returns the best split based on I. Whereas a more costly variant may continue by exploring the splits based on I_L and I_R and stops only if $e_I < min(e_{I_L}, e_{I_R})$. Second, if $e_{I_L} = e_{I_R} < e_I$ (Algorithm 3 line 8); the split continues with only the left sub-interval.
The three algorithms $TSSplit$, $AdaptSplit$, and $DichoSplit$ were implemented in C and integrated to the CART algorithm of the R package tree proposed by B. Ripley (http://cran.r-project.org/web/packages/tree/). As default parameters, the time series decision tree is induced without pruning and with a minimum size of leaves of 2 time series.

4.4. Time series classification

Each node of the induced time series classification tree ($TSTree$) is characterized by a split test $\sigma(l_*, r_*, k_*, I_*)$ described by the two representative time series (l_*, r_*), the optimal value k_* of the learned metric D_{k_*} and the most discriminating sub-sequence I_*. A new time series ts is assigned to the left sub-node if it is closest to the left time series l_* than to r_* with $D_{k_*}(ts, l_*) <= D_{k_*}(ts, r_*)$, otherwise assigned to the right sub-node. The time series proximities D_{k_*} are evaluated over the discriminant period I_*. As in conventional classification trees, ts is assigned to the class of the leaf in which it falls.

Algorithm 1 $TSSplit(S, I, \alpha)$

$(\sigma(l_*^I, r_*^I, k_*^I, I), e_I) = AdaptSplit(S, I)$
$(\sigma(l_*, r_*, k_*, I_*), e_{I_*})$=$DichoSplit(S, \sigma(l_*^I, r_*^I, k_*^I, I), e_I, \alpha)$
return$(\sigma(l_*, r_*, k_*, I_*), e_{I_*})$

Algorithm 2 $AdaptSplit(S, I)$

$e_* = \infty$
for k in $[0; 6]$ **do**
  $(l_k, r_k) = \arg\min_{(l,r)}(GI(\sigma(l, r, k, I)))$
  if $GI(\sigma(l_k, r_k, k, I)) < e_*$ **then**
    $e_* = GI(\sigma(l_k, r_k, k, I))$
    $l_*^I = l_k$, $r_*^I = r_k$, $k_*^I = k$
  end if
end for
return$(\sigma(l_*^I, r_*^I, k_*^I, I), e_*)$

4.5. $TSSplit$ complexity

Let N be the number of time series, K the number of explored values of the parameter k, T the initial time series length, and α the cover rate for the dichotomous search. On the one hand, the core of the complexity of $AdaptSplit$ is determined by the exhaustive search of the triple (l, r, k) which is $O(KN^2)$. For each explored triple, the Gini index (complexity of $O(N)$) is evaluated, leading to a total complexity for $AdaptSplit$ of $O(KN^3)$. The $AdaptSplit$ complexity can be reduced by limiting the exhaustive search to the pairs of

Algorithm 3 $DichoSplit(S, \sigma(l_*^I, r_*^I, k_*^I, I), e_I, \alpha)$

$[a, b] = I$
$I_L = [a, a + \alpha(b - a)]$
$I_R = [b - \alpha(b - a), b]$
$(\sigma(l_*^{I_L}, r_*^{I_L}, k_*^{I_L}, I_L), e_{I_L}) = AdaptSplit(S, I_L)$
$(\sigma(l_*^{I_R}, r_*^{I_R}, k_*^{I_R}, I_R), e_{I_R}) = AdaptSplit(S, I_R)$
if $e_I \leq \min(e_{I_L}, e_{I_R})$ **then**
    return$(\sigma(l_*^I, r_*^I, k_*^I, I), e_I)$
else if $e_{I_L} \leq e_{I_R}$ **then**
    $DichoSplit(S, \sigma(l_*^{I_L}, r_*^{I_L}, k_*^{I_L}, I_L), e_{I_L}, \alpha)$
else
    $DichoSplit(S, \sigma(l_*^{I_R}, r_*^{I_R}, k_*^{I_R}, I_R), e_{I_R}, \alpha)$
end if

time series of different classes. On the other hand, $DichoSplit$ performs two calls to AdapSplit and a recursive call to DichoSplit if either I_L or I_R provides a better purity Gini index than the interval I. The maximum number of recursive calls is $Log_{\frac{1}{\alpha}}(T)$, corresponding to the number of dichotomous splits of $I = [0, T]$ until having a sub-interval of length one. Thus, in the worst case, the complexity of $DichoSplit$ is $O(Log_{\frac{1}{\alpha}}(T)2KN^3)$. Finally, based on the complexities of $AdaptSplit$ and $DichoSplit$, the complexity of $TSSplit$ is dominated by $O(Log_{\frac{1}{\alpha}}(T)2KN^3 + KN^3)$ that is globally about $O(Log_{\frac{1}{\alpha}}(T)KN^3)$.

5. Experimental study

The proposed time series classification tree $TSTree$ is first applied to four public datasets frequently used in the literature for the validation of the major competitive approaches: CBF Saito (1994), CBF-TR Geurts (2002), CC Asuncion & Newman (2007), and TWO-PAT Geurts (2002). Note that the four datasets share some similar characteristics: each class identifies a distinctive global behavior, classes are well discriminated by their global behaviors, and time series progress in relatively close domains. It is indisputable that in real applications time series specifications may be more complex. For instance, time series may involve time delays, have tendency or amplitude variations, may share a global common profile or be characterized by some local com-

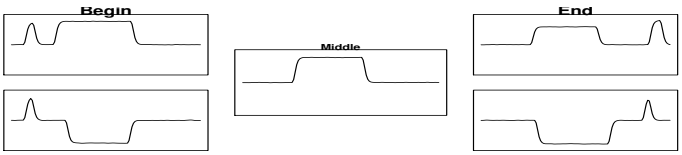

Figure 2: LOCAL-DISC classes

mon events. To complete and broaden the validation process to properties frequently encountered in temporal applications, we propose three additional datasets. On the one hand, a synthetic time series datasets with time series discrimination based on local events rather than on the global behaviors. On the other hand, two real and multivariate time series, describing character trajectories and handwritten digits Asuncion & Newman (2007). Let us detail in the following section the specifications of these additional datasets.

5.1. Additional time series datasets

5.1.1. LOCAL-DISC

The aim of the LOCAL-DISC dataset is to study the efficiency of time series classification trees when time series of a same class may have distinctive global behaviors while sharing common local features. The LOCAL-DISC dataset is composed of 3 time series classes, Begin, Middle and End. In the Begin class, the time series share a common local event characterized by a little bell arising at the begin period; the Middle class consists of time series sharing similar global behavior characterized by a centered large bell, and the time series of the End class share a common local event corresponding to a bell arising at the end period. The Begin, Middle and End time series classes are illustrated in Figure 2. First, remark that the global behavior is not a discriminative criterion as time series of different classes may share similar global behaviors (e.g., a cylinder shape). Second, time series of the same class may eventually have different global behaviors and progress in different ranges of values.

5.1.2. Character trajectories

The character trajectories dataset Asuncion & Newman (2007) consists of a set of pen tip trajectories recorded while writing individual characters. All samples are from the same writer, for the purposes of primitive extraction.

Only characters with a single pen-down segment were considered. The data were captured using a WACOM tablet. Each handwritten character trajectory is a 3-dimensional time series: x, y for the pen positions and z for the pen tip force.

5.1.3. Handwritten digits

The handwritten digits data are extracted from the UJI Pen Characters database Asuncion & Newman (2007). Samples are collected from 11 writers, with two samples for each pair writer/digit. Only x and y coordinate information was recorded along the strokes by the acquisition program, without, for instance, pressure level values or timing information. As several handwritten prototypes may be used by the 11 writers to generate a same digit, a class may be composed of time series of different global behaviors. For usual and additional datasets, Table 1 gives the main characteristics of the above datasets.

Name	Sample size	Num. classes	Num. TS/class	TS lengths	Multi. TS.	Real data
CBF	300	3	100	128	No	No
CBF-TR	300	3	100	128	No	No
CC	600	6	100	60	No	No
TWO-PAT	400	4	100	128	No	No
LOCAL-DISC	300	3	100	128	No	No
CHAR-TRAJ	400	20	20	[100-200]	Yes	Yes
DIGITS	220	10	22	110	Yes	Yes

Table 1: Usual and additional datasets description

5.2. Validation protocol

To highlight the additive value of the new temporal classification tree, several configurations of the split procedure are considered: an adaptive metric (i.e., a behavior and values based metric) vs. a non-adaptive metric (i.e., a classical values based metric), a dichotomous vs. a non-dichotomous search, and a temporal correlation vs. a Pearson correlation for the behavior cost-function. In addition, according to classes including or not time delays, these configurations are modulated for several variants of the dynamic time warping and of the Euclidean distance. A misclassification error rate, based on a 10-fold stratified cross-validation, is estimated.

Time delay	Adap. metric	Dich. search	Behav. cost	Metric
Yes	Yes	Yes	$Cort$	DTW_k^{cort}
	Yes	Yes	Cor	DTW_k^{cor}
	Yes	No	$Cort$	DTW_k^{cort}
	Yes	No	Cor	DTW_k^{cor}
	No	No	-	d_{Dtw}
No	Yes	Yes	$Cort$	DE_k^{cort}
	Yes	Yes	Cor	DE_k^{cor}
	Yes	No	$Cort$	DE_k^{cort}
	Yes	No	Cor	DE_k^{cor}
	No	No	-	d_E

Table 2: The studied configurations for $TSSplit$

5.3. Performances results

Table 3 and 4 give for each usual and additional dataset the misclassification error rates and the number of leaves of the induced trees. These results allow us to study the effect of each $TSSplit$'s configuration (Table 2) on the performances of the induced tree. In particular, it allows us to compare the decision trees performances when the split criterion uses an adaptive metric versus a standard one, involves a dichotomous search versus not. For instance, Figure 3 visualizes the DIGITS trees of minimum error rate over the studied $TSSplit$ configurations. Let us first bring some interpretation elements of the built classification trees. Each node is characterized by the triplet $(Type, I_*, Class)$ indicating respectively: the metric's type "B", "V", or "BV" indicating, respectively, if the learned D_{k_*} is behavior-based (k_* greater than 3), values-based (k_* lower than 3), or equally behavior and values-based (for k=3), the most discriminating interval or sub-interval I_* on which D_{k_*} will be evaluated, and the class label of the representative time series.

5.4. Discussion

From Table 3 we can see for all datasets (except for the noisy TOW-PAT) that a $TSTree$ based on an adaptive metric (the 4th first configurations) outperforms a tree based on the standard metrics (d_E, d_{DTW}). The performances of $TSTree$ remain the same when it involves or not the dichotomous search revealing that each class is characterized by one distinctive global behavior and a discrimination mainly based on the global behaviors of time series. Finally, for all the datasets the performances of $TSTree$ are improved when involving the temporal correlation instead of the Pearson correlation. From Table 4, we can see for all additional datasets, introducing several temporal peculiarities, that the $TSTree$ based on an adaptive metric outperforms a tree based on the standard metrics. These performances are always improved

Datasets	Metric	Dicho.	Error rate	Nb. leaves
CBF	DE_k^{Cort}	Yes	**0.000000**	3
	DE_k^{Cor}	Yes	**0.000000**	3
	DE_k^{Cort}	No	**0.000000**	3
	DE_k^{Cor}	No	**0.000000**	3
	d_E	No	0.006667	3
CBF-TR	DTW_k^{Cort}	Yes	**0.023333**	3
	DTW_k^{Cor}	Yes	0.170000	22
	DTW_k^{Cort}	No	**0.023333**	3
	DTW_k^{Cor}	No	0.183333	23
	d_{Dtw}	No	0.136667	30
CC	DTW_k^{Cort}	Yes	**0.005000**	6
	DTW_k^{Cor}	Yes	0.028333	7
	DTW_k^{Cort}	No	**0.005000**	6
	DTW_k^{Cor}	No	0.025000	10
	d_{Dtw}	No	0.021667	13
TWO-PAT	DTW_k^{Cort}	Yes	0.002632	6
	DTW_k^{Cor}	Yes	0.002632	4
	DTW_k^{Cort}	No	0.002632	6
	DTW_k^{Cor}	No	0.002632	4
	d_{Dtw}	No	**0.000000**	4

Table 3: Times series classification trees on the usual datasets: adaptive vs. standard metrics

when involving the temporal correlation instead of the Pearson correlation. The dichotomous search improves significantly the results for LOCAL-DISC and DIGITS, as classes may be composed of time series of different global behaviors. In fact, for DIGITS the 11 writers may follow different trajectories to write a same digit. From the tree given in Figure 3, we can see that the dichotomous search plays a part at two nodes: to separate the digits 3 and 5, then 4 and 9. In fact, time series of the digits 3 and 5 (resp. 4 and 9) provided by different writers may be very close on the second half of the trajectories. Thus, the dichotomous search selects the first half of the trajectories (underlined in red in Figure 3) as best discriminating these digits. In other words, to best separate the digits 3 and 5 (resp. 4 and 9) the dissimilarities between those digits are evaluated based on their first half trajectories. Finally, one may know from the induced trees the typical profiles of the time series belonging to each child node, whether these profiles cover all the observation period, or reference sub-sequences, and whether the assignation rules are values or behavior-based. Although the best specification for $TSSplit$ involves an adaptive metric, the temporal correlation and a dichotomous search, studying the other configurations may be informative: it may reveals whether the

Datasets	Metric	Dicho.	Error rate	Nb. leaves
LOCAL-DISC	DTW_k^{Cort}	Yes	**0.020000**	3
	DTW_k^{Cor}	Yes	**0.020000**	5
	DTW_k^{Cort}	No	0.073333	13
	DTW_k^{Cor}	No	0.096667	22
	d_{Dtw}	No	0.096667	30
CHAR-TRAJ	DTW_k^{Cort}	Yes	**0.075000**	20
	DTW_k^{Cor}	Yes	0.082500	20
	DTW_k^{Cort}	No	**0.075000**	24
	DTW_k^{Cor}	No	0.095000	24
	d_{Dtw}	No	0.080000	24
DIGITS	DTW_k^{Cort}	Yes	**0.065657**	12
	DTW_k^{Cor}	Yes	0.141414	11
	DTW_k^{Cort}	No	0.141414	13
	DTW_k^{Cor}	No	0.161616	12
	d_{Dtw}	No	0.247475	16

Table 4: Times series classification trees on the additional datasets: adaptive vs. standard metrics

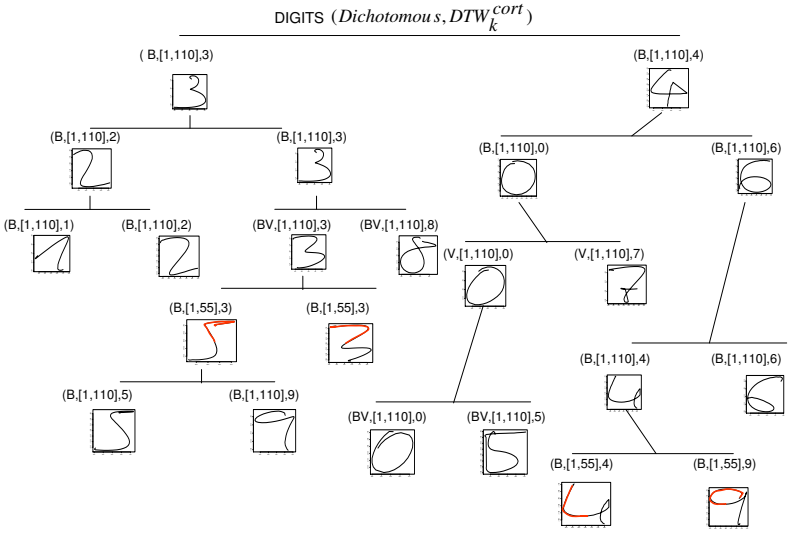

Figure 3: Classification tree of DIGITS data

time series involve amplitude variations and drifts, whether the discrimination is due to some local events or to the global behavior, and whether it is values

or behavior-based.

References

ASUNCION A. & NEWMAN D. (2007). UCI, machine learning repository. [http://www.ics.uci.edu/ mlearn/MLRepository.html]: Irvine, CA: University of California, School of Information and Computer Science.

BALAKRISHNAN S. & MADIGAN D. (2006). Decision trees for functional variables. In *International Conference on Data Mining*, p. 798 –802.

DOUZAL-CHOUAKRIA A. & AMBLARD C. (2012). Classification trees for time series. *Pattern Recognition*, **45**(3), 1076–1091.

DOUZAL-CHOUAKRIA A., DIALLO A. & GIROUD F. (2009). Adaptive clustering for time series: application for identifying cell cycle expressed genes. *Computational Statistics and Data Analysis*, **53**(4), 1414–1426.

GARCIA-ESCUDERO L. A. & GORDALIZA A. (2005). A proposal for robust curve clustering. *Journal of Classification*, **22**, 185–201.

GEURTS P. (2002). *Contributions to decision tree induction: bias/variance tradeoff and time series classification.* PhD thesis, Department of Electrical Engineering, University of Liege, Belgium.

GEURTS P. & WEHENKEL L. (2005). Segment and combine approach for non-parametric time-series classification. In *PKDD*, p. 478–485.

KRUSKALL J. & LIBERMAN M. (1983). *The symmetric time warping algorithm: From continuous to discrete. In Time Warps, String Edits and Macromolecules.* Addison-Wesley.

KUDO M., TOYAMA J. & SHIMBO M. (1999). Multidimensional curve classification using passing-through regions. *Pattern Recognition Letters*, **20**(11), 1103–1111.

SAITO N. (1994). *Local feature extraction and its application using a library of bases*. PhD thesis, Department of Mathematics, Yale University.

SERBAN N. & WASSERMAN L. (2005). CATS: Cluster after transformation and smoothing. *Journal of the American Statistical Association*, **100**(471), 990–999.

YAMADA Y., SUZUKI E., YOKOI H. & TAKABAYASHI K. (2003). Decision-tree induction from time-series data based on standard-example split test. In *Proceedings of the 20th International Conference on Machine Learning*, p. 840–847: Morgan Kaufmann.