Machine Learning on temporal data

Classification Trees for Time Series

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

Plan

- Time Series classification approaches
- Classification trees for time series: main issues
- Distance-based time series split tests
- Adaptive split tests and discriminant sub-sequences extraction
- Applications: handwritten characters

Time Series Classification Approaches

Project time series into a given functional basis space.

- Mapping time series to a new description space where conventional classifiers can be applied (Fourier or wavelet transform, a polynomial or an ARIMA approximation,...)
- Signal processing or statistical tools are commonly used to project time series into a given functional basis space.
- The projected space performed by: Fourier or wavelet transform, a polynomial or an ARIMA approximation.
- Standard classifiers are subsequently applied on the fitted basis coefficients.

Garcia-Escudero et al. (2005), Serban et al. (2004), Caiado et al. (2006), Kakizawa et al. (1998), etc.

Time Series Classification Categories

Prototypes extraction heuristics

- The time series segmentation to extract prototypes (subsequences or regions of values) that best characterize the time series classes,
- Prototypes are described by a set of numerical features,
- Standard classifiers can be applied

Kudo et al. (1999), Rodriguez et al. (2001), Geurts (2001), etc.

Distance-based approaches

- Define finely new time series proximity measures,
- Extend conventional classification approaches based on the new time series dissimilarities

Classification Trees for Time Series: Distance-based approaches

Design Issues of Classification Trees Induction

1. How should the training sets be split?
   - Recursively the tree-growing process must select a variable split test to divide the samples (internal nodes) into smaller subsets.
   - The algorithm must provide a method for specifying the split test for different variable types (categorical, continous, ordered, binary, ...)
   - Provide an objective function for evaluating the goodness of each split test.

2. How should the splitting procedure stop?
   - Expanding a node until either all the samples belong to the same class
   - The same value for all the samples.
   - Other conditions may be specified to allow the tree-growing procedure to terminate earlier.

Time Series Split-test: distance-based strategies

- Build binary classification trees in which internal nodes are labeled by one or two time series,
- Proposed classifiers are mainly based on new split tests to bisect the set of time series within internal nodes most effectively

Yamada et al. (2003), Balakrishnan et al. (2006),...

Standard-example split test

- Uses an exhaustive search to select one existing time series (called the standard time series), leading to division with a maximum purity gain ratio (i.e., $-\sum_i p_i log(p_i)$).
- The first child node is composed of time series with a distance (DTW) to the standard time series that is less than a given threshold.
- 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.

Standard-example split test Algorithm (Yamada et al. (2003)

Procedure: *standardExSplit*
Input: Set of examples $e_1, e_2, \cdots, e_n$
Return value: Best split test ω

1 ω.gr $= 0$
2 **Foreach**(example e)
3 **Foreach**(time-series attribute a)
4 Sort examples $e_1, e_2, \cdots, e_n$ in the current node using $G(e(a), e_i(a))$ as a key to $\epsilon_1, \epsilon_2, \cdots, \epsilon_n$
5 **Foreach**(θ-guillotine cut ω' of $\epsilon_1, \epsilon_2, \cdots, \epsilon_n$)
6 **If** ω'.gr $>$ ω.gr
7 $\omega = \omega'$
8 **Else If** ω'.gr $==$ ω.gr **And** ω'.gap $>$ ω.gap
9 $\omega = \omega'$
10 **Return** ω

Cluster-example split test

- Performs an exhaustive search for two standard time series.
- The bisection is constructed by assigning each time series to the nearest standard time series (DTW).
- The purity gain ratio and the class isolation criterion are used to select the best split test.

Procedure: *clusterExSplit*
Input: Set of examples $e_1, e_2, \cdots, e_n$
Return value: Best split test ω

```
1   ω.gr = 0
2   Foreach(pair of examples e', e'')
3     Foreach(time-series attribute a)
4       ω' = σ'(e', e'', a)
6       If ω'.gr > ω.gr
7         ω = ω'
8       Else If ω'.gr == ω.gr And ω'.gap > ω.gap
9         ω = ω'
10  Return ω
```

Clustering-based split test (Balakrishnan et al. (2006))

- Look for a pair of reference time series that best bisects the set of time series according to a clustering-goodness criterion.
- For this, a k-means algorithm is performed.
- This algorithm ensures a partitioning that optimizes clustering criteria, namely, the compactness and isolation of the clusters but not their purity.
- Iteratively the k-means clustering is performed several times to select the partition that gives the highest Gini index.
- The centers of the clusters define the pair of reference time series for 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.

Time Series split tests startegies

Remarks

- Looks for a set of time series bisections with the highest purity clusters (i.e., the highest Gini index), subsequently picks the one optimizing some clustering criteria (i.e., maximizing the separability of the clusters),
- Looks for a set of splits that optimize clustering criteria (i.e., k-means criteria) and accordingly selects the one exhibiting the highest purity clusters (i.e., maximizing the Gini index).
- Giving priority to a clustering criterion instead of the purity of the clusters, the split test may fail to select bisections of lower clustering criteria but of higher purity.

Time Series Classification trees

Challenges

- Extending classification trees to time series input variables (split test and objective function),
- Time series peculiarities may change from one node to another,
- Time series discrimination relies on some sub-sequences (i.e., segments of time series).

Objectives

- Split test criteria should involve adaptive time series metrics,
- Perform automatic extraction to extract the most discriminating sub-sequences (i.e., segments of time series).
- Outperforms temporal trees using standard time series distances,
- Leads to good performances compared to other competitive time series classifiers.

Time Series Classification Trees

Notations and specifications

- $\{s_1, ..., s_N\}$ a set of N multivariate time series partitioned into C classes,
- $I_1, ..., I_N$ $(I_i = [1, T_i])$ be their respective observation intervals,
- $s = (u_1, ..., u_p)$ a time series,
- $(u_i, ..., u_{i+q-1})$ $(1 \leq i \leq p - q + 1)$ a subsequence of s is a sampling of length $q < p$ of contiguous position from s.

Classification Trees algorithm

Time series length normalization

- Case 1 (Allowing time delays): time series are simply resampled (e.g., a linear interpolation) to make them of equal length I = [1, T].
- Case 2 (not allowing time delays): the smallest observation period $I = min(I_1, ..., I_N)$ is considered; data are resampled within I.

Classification Trees algorithm

Time series split (*TSSplit*) test algorithms

- $TSSplit(S, I, \alpha)$: to bisect a given node $S = \{s_1, ..., s_N\}$ composed of a set of time series,
- observed on the interval $I = [1, T]$,
- with a rate α the discriminant subsequences search parameter.
- $AdaptSplit(S, I)$ is performed to determine the best split of S involving the adaptive metric D_k evaluated on I.

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

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

Adaptive Split test

Adaptive Split algorithm

- Given a value of the parameter $k \in [0, 6]$ and two time series (l, r) from $S \times S$,
- $\sigma(l, r, k, I)$, a bisection of S, is obtained by:
  - Assigning each time series $ts \in S$ to the left node if it is closer to the time series l than to r ($D_k(ts, l) \leq D_k(ts, r)$), and to the right node otherwise.
- Iteratively, several values of the triplet (l, r, k) are explored to find the bisection that exhibits the minimum impurity Gini index.
- Returns the best split $\sigma(l^I_*, r^I_*, k^I_*, I)$ and its impurity Gini index $GI(\sigma(l^I_*, r^I_*, k^I_*, I))$.

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$(\sigma(l^I_*, r^I_*, k^I_*, I), e_*)$

Adaptive Split algorithm

Limitation of the adaptive split

- The best split $\sigma(I_*^I, r_*^I, k_*^I, I)$ is obtained by comparing the time series proximities according to their observations within I,
- The split $\sigma(I_*^I, r_*^I, k_*^I, I)$ fails to reach higher purity classes when time series differentiation is induced by subsequences instead of implicating all observations of I

Dichotomic Split algorithm

Dichotomic search for discriminant subsequences extraction

- *DichoSplit* allows us to determine subsequences of I entailing a bisection of lower impurity than $\sigma(I_*^I, r_*^I, k_*^I, I)$,
- A dichotomy search is performed within the left and right subsequences of I.

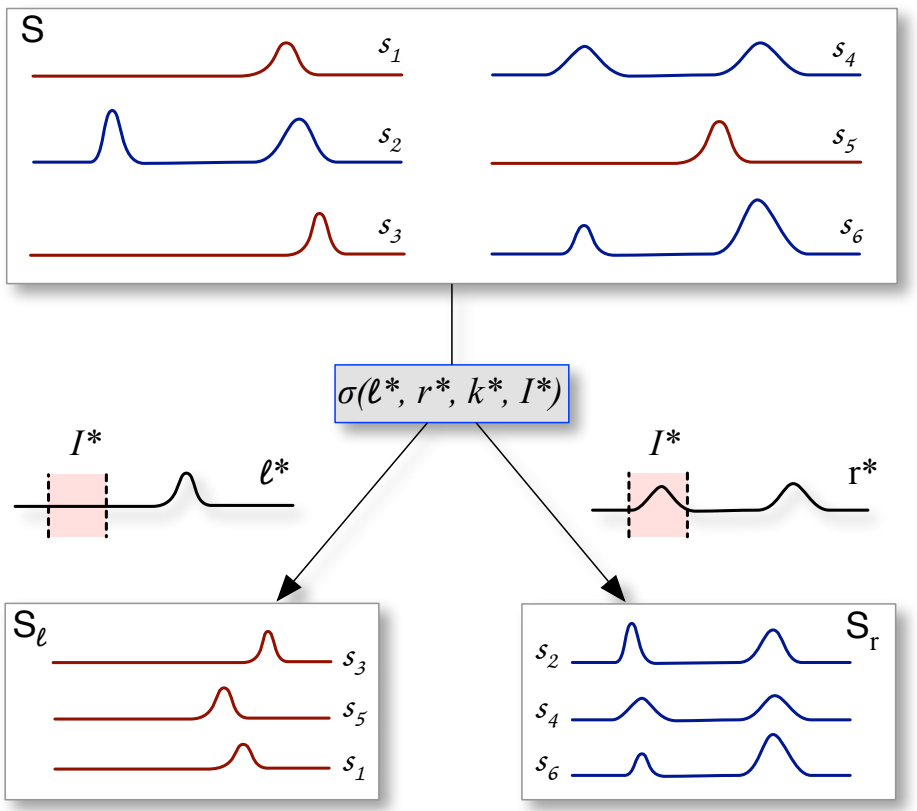
Algorithm 3 $DichoSplit(S, \sigma(I_*^I, r_*^I, k_*^I, I), e_I, \alpha)$

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

Time Series Split algorithm

The induced time series classification tree ($TSTree$) is characterized by:

- A split test $\sigma(l_*, r_*, k_*, I_*)$.
- Two representative time series (l_*, r_*).
- The optimal value k_* of the learned metric D_{k_*}.
- The most discriminating subsequence I_*.



Time Series Split algorithm

Time series classification rule

- A new time series ts is assigned to the left sub-node if it is closer to the left time series I_* than to r_* ($D_{k_*}(ts, I_*) <= D_{k_*}(ts, r_*)$); otherwise, it is 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 it falls in.

Experimental study

Time series datasets

- The proposed time series classification tree *TSTree* is applied to four public datasets CBF, CBF-TR (Geurts 2002) , CC (Asuncion et al. 2007), and TWO-PAT (Geurts 2002).

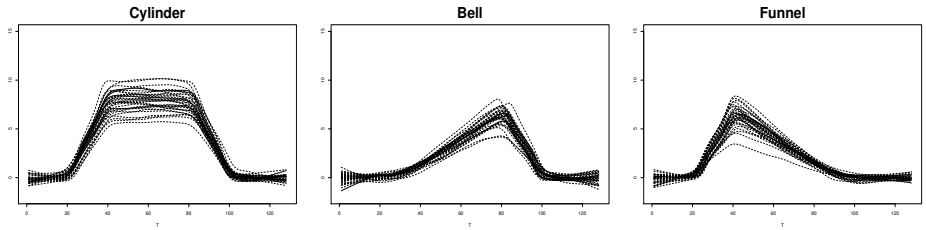
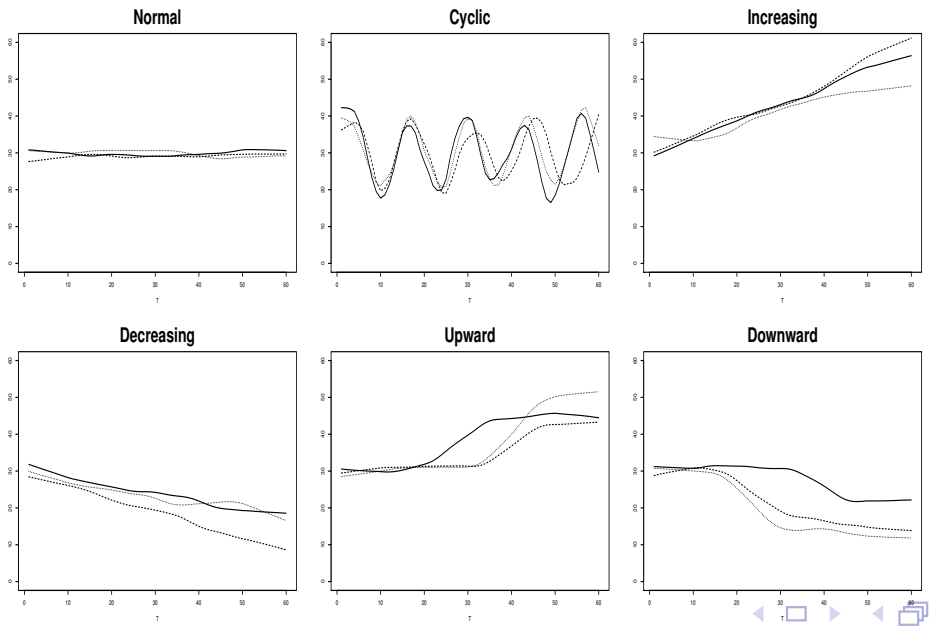


Figure: CBF (Saito 1994)



Experimental study

Time series datasets characteristics

- Each class identifies a distinctive global behavior,
- Classes are well discriminated by their global behaviors,
- time series progress in relatively close domains....

Time series with complex peculiarities

- Time series progressing in different ranges of values (CBF-RANGVAR),
- Time series discrimination based on local events (LOCAL-DISC dataset), .
- Character trajectories (CHAR-TRAJ) (Asuncion et al. 2007)
  - Pen tip trajectories recorded while writing individual characters, captured using a WACOM tablet
  - All samples are from the same writer,
  - Each handwritten character trajectory is a 3-dimensional time series: x, y for the pen positions and z for the pen tip force.
- Handwritten digits (DIGITS)(Asuncion et al. 2007)
  - Samples are collected from 11 writers,
  - x and y coordinate information was recorded along the strokes,
  - A class (character) may be composed of time series of different global behaviors.

Experimental study

Time series with complex peculiarities

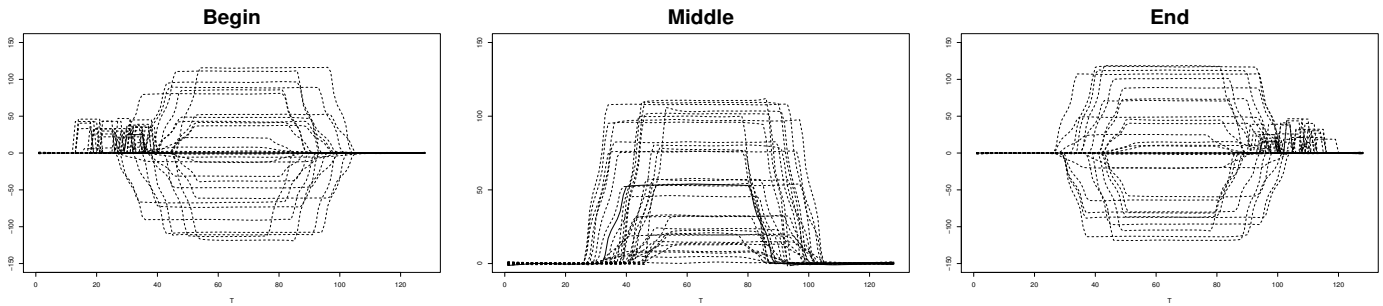


Figure: LOCAL-DISC time series classes

Experimental study

Classification trees performances

Datasets	Metric	Adap.	Dicho.	Error rate	Nb. leaves
LOCAL-DISC	DTW_k^{Cort}	Yes	Yes	**0.020**	3
	DTW_k^{Cor}	Yes	Yes	**0.020**	5
	DTW_k^{Cort}	Yes	No	0.073	13
	DTW_k^{Cor}	Yes	No	0.096	22
	d_{Dtw}	No	No	0.096	30
CBF-RANGVAR	DTW_k^{Cort}	Yes	Yes	**0.006**	3
	DTW_k^{Cor}	Yes	Yes	0.053	10
	DTW_k^{Cort}	Yes	No	**0.006**	3
	DTW_k^{Cor}	Yes	No	0.070	15
	d_{Dtw}	No	No	0.060	21
GENES	DE_k^{Cort}	Yes	Yes	**0.004**	5
	DE_k^{Cor}	Yes	Yes	**0.004**	5
	DE_k^{Cort}	Yes	No	**0.004**	5
	DE_k^{Cor}	Yes	No	**0.004**	5
	d_E	No	No	0.036	8
CHAR-TRAJ	DTW_k^{Cort}	Yes	Yes	**0.075**	20
	DTW_k^{Cor}	Yes	Yes	0.082	20
	DTW_k^{Cort}	Yes	No	**0.075**	24
	DTW_k^{Cor}	Yes	No	0.095	24
	d_{Dtw}	No	No	0.080	24
DIGITS	DTW_k^{Cort}	Yes	Yes	**0.065**	12
	DTW_k^{Cor}	Yes	Yes	0.141	11
	DTW_k^{Cort}	Yes	No	0.141	13
	DTW_k^{Cor}	Yes	No	0.161	12
	d_{Dtw}	No	No	0.247	16

Experimental study

Induced Classification Trees

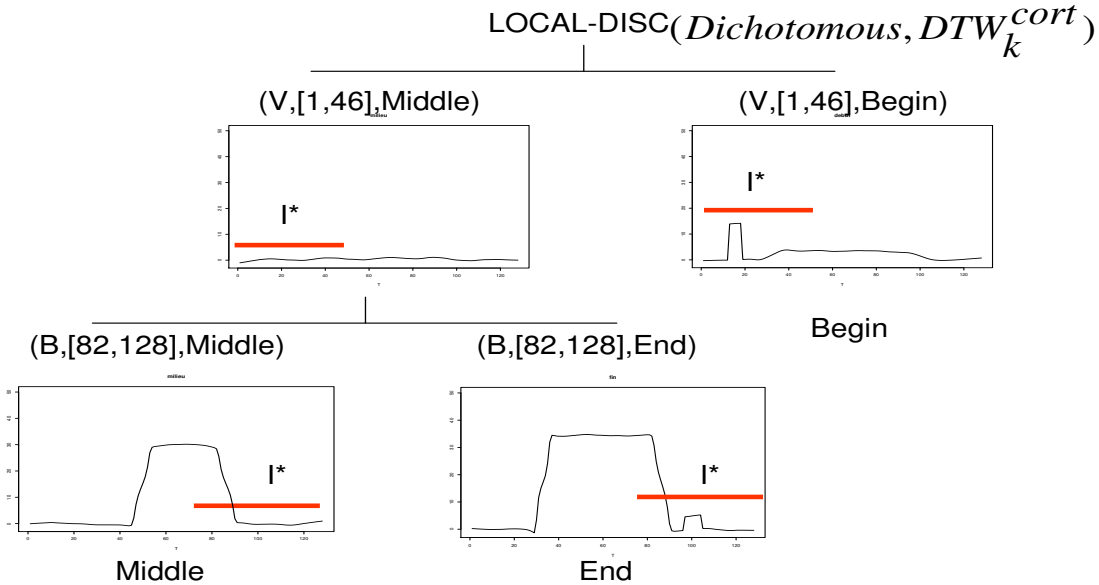


Figure: Classification tree of LOCAL-DISC data

Experimental study

Classification Tree performances

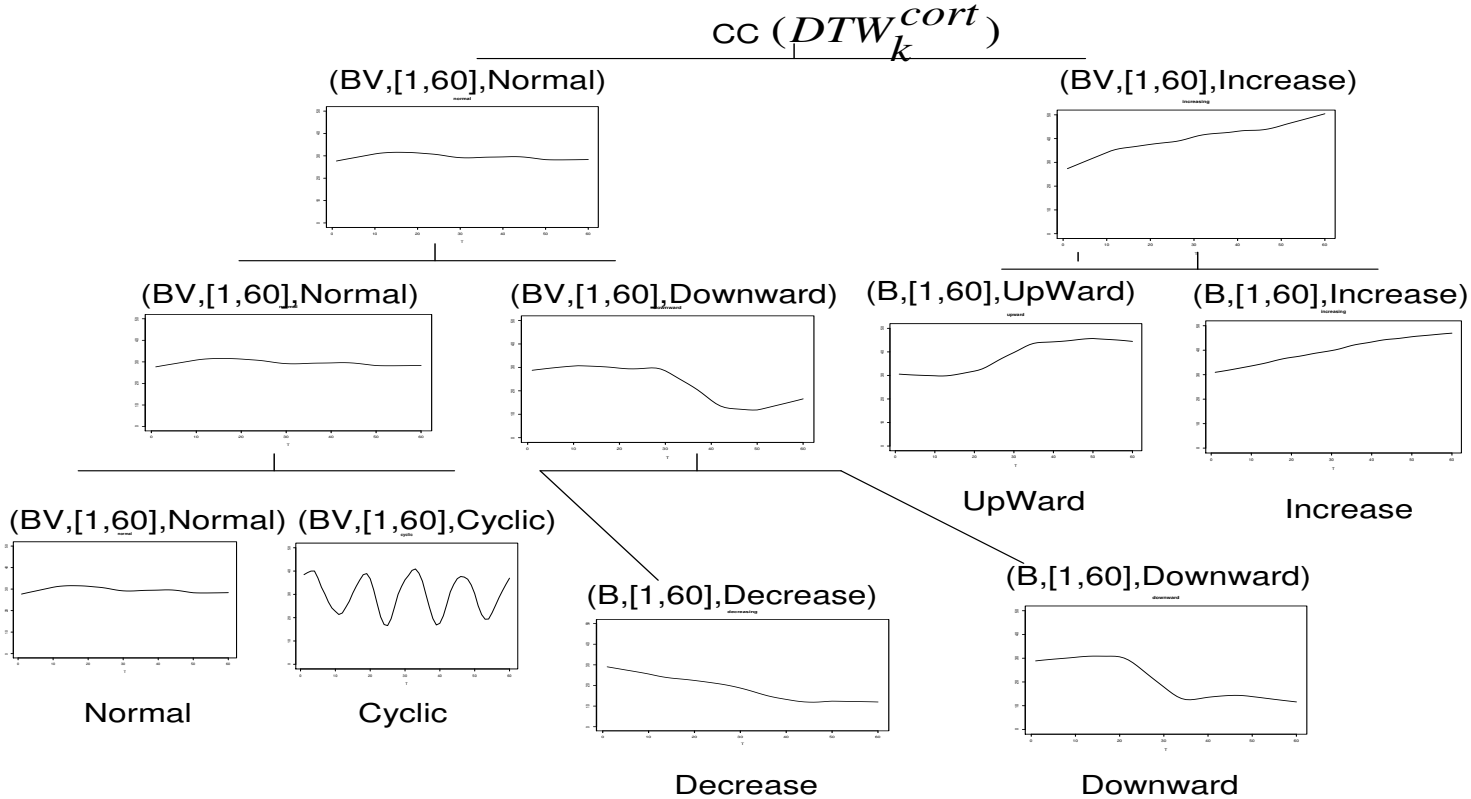
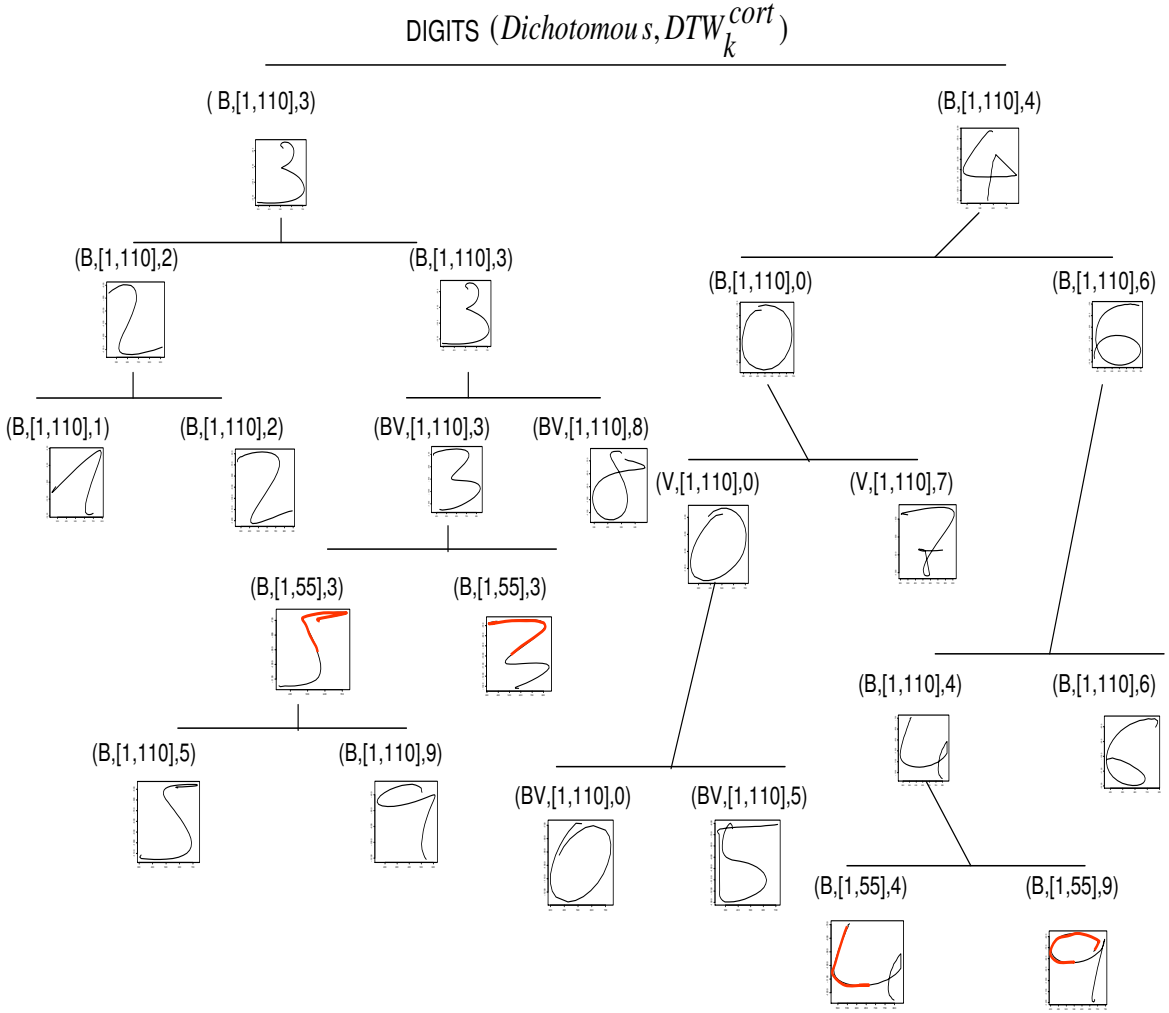


Figure: Classification tree of cc data

Experimental study

Classification Tree performances



References

- Asuncion, A. and Newman, D.J. (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. ICDM International Conference on Data Mining, 798-802.
- Caiado, J., Crato, N., Pena, D. (2006). A periodogram-based metric for time se- ries classification. Computational Statistics and Data Analysis. 50, 2668-2684.
- Garcia-Escudero, L. A., Gordaliza, A. (2005). A proposal for robust curve clustering. Journal of Classification. 22, 185-201.
- Geurts, P. (2001). Pattern extraction for time series classification. LNCS Principles of Data Mining and Knowledge Discovery, 115-127.
- Kakizawa, Y., Shumway, R.H., Taniguchi, N. (1998). Discrimination and clustering for multivariate time series. Journal of the American Statistical Association. 93, 328-340.
- Kudo, M., Toyama, J., and Shimbo, M. (1999). Multidimensional Curve Classification Using Passing-Through Regions. Pattern Recognition Letters, Vol. 20, No. 11-13, 1103-1111.
- Saito, N. (1994). Local feature extraction and its application using a library of bases. Phd thesis. Department of Mathematics, Yale University.
- Rodrguez, J.J., Alonso, C.J., Bostrom, H. (2001). Boosting interval-based literals. Intelligent Data Analysis, vol 5, issue 3, 245-262.
- Serban, N., Wasserman, L. (2004). CATS: Cluster After Transformation and Smoothing. Journal of the American Statistical Association. 100, 990-999.
- Yamada, Y., Suzuki, E., Yokoi, H., and Takabayashi, K. (2003). Decision-tree induction from time-series data based on standard-example split test. International Conference on Machine Learning.