Temporal Decision Trees

V. Shalaeva, G. Bisson, C. Amblard

Université Grenoble Alpes, Laboratoire d'Informatique de Grenoble

Introduction

Nowadays an enormous amount of temporal data is produced and need to be processed.

There are a lot of applications where it's necessary to mine temporal data such that genomic analysis, information retrieval, finance, energy data analytics, airplane tracking and so forth[1]. We are focused on the **classification temporal mining task** is, where aim is to assign labels to each time series of an input set.

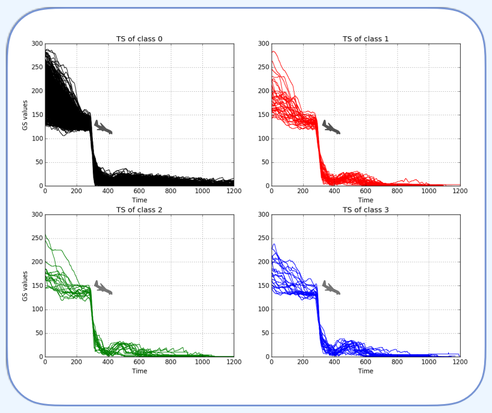

Fig.1: Tracking of airplane's speed landing

Objective

The goal of this PhD research is develop **a rapid and accurate classification algorithm which outputs are easily interpretable by the end users.**

Method

Decision Tree for Time Series (TS) [2] - modification of classical decision tree algorithm to handle temporal data. Input data is a collection of **labelled time-series.** Split at each node is based on distance between chosen splitting pair of TS and input set. Selection of the best pair of TS (i.e. the most discriminant one) is based on Gini impurity index. The algorithm allows also to explore different distance measures at each node and to find the most discriminative time interval by dichotomy search. The split process terminates when a leaf of the TS tree contains a set of TS owning to a single class.

Input: (set of time series, labels), distance measures
Function Split&CreateNode:
  for each (*pair of TS)* **do**:
    while Gini improve **do**:
      for each time sub-interval and distance measure **do**:
        - evaluate Gini of each *(pair of TS)*
        - select pair with the *best Gini value*
      split TS time interval on two sub-intervals
Return the most discriminative *(pair of TS)*

Fig.2: Algorithm of node split

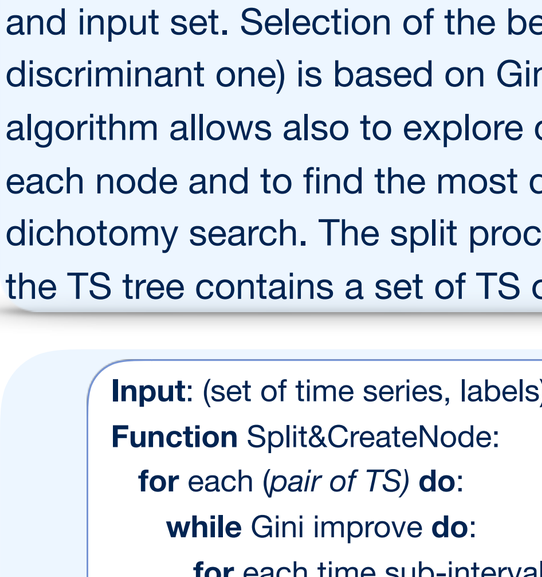

Fig.3: Principle of algorithm

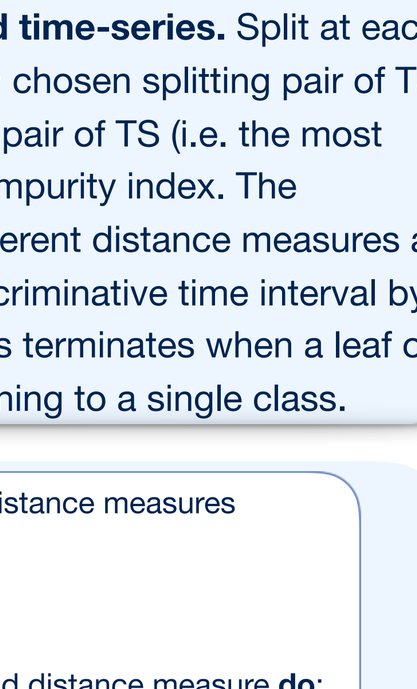

Fig.4: Output tree - 'Arrow Head' dataset[3]

Ongoing research

With strong characteristics of **interpretability** and **performance** the current drawback of the algorithm is its **high algorithmic complexity O(N^3*K*log(T))**, where **N** is number of time-series, **K** is number of explored similarity metrics, **T** is the average length of TS.

Complexity reduction approaches:

1. **Surrogate evaluation criterion.** The pairwise distance matrix of size **NxN** is computed to evaluate and select the best pair of TS. The complexity of one distance computation is **O(T^2)**. An approximation of evaluation criterion is suggested to use to reduce number of distance computations, in that way reduce the complexity.

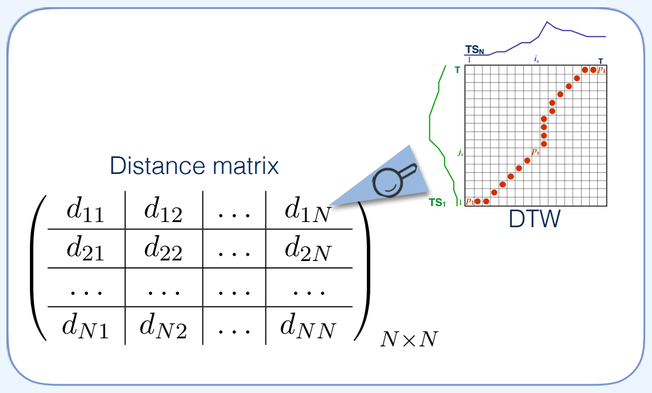

Fig.5: Distance matrix

2. **Reduce number of time-series pairs under split ability examining at each node. N** input time-series effects ~ **N^2** split candidates. Some of splitters have the similar discriminative ability and it's useless to evaluate all of them. We are investigating the methods to find groups of splitter with the similar split ability using small subsample of input time series.

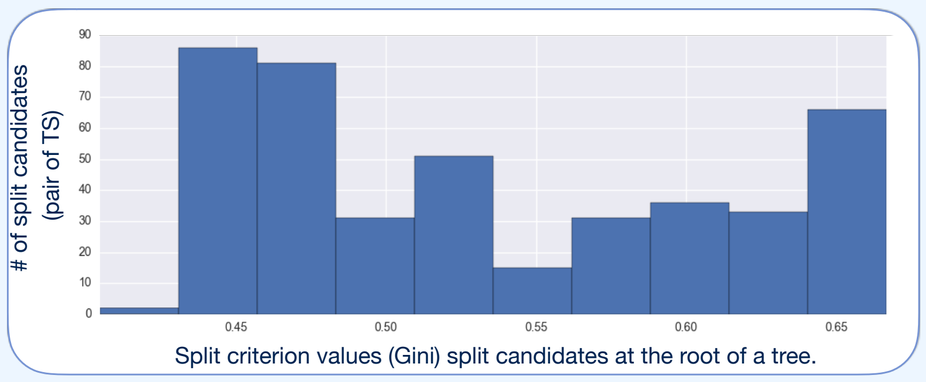

Fig.6: Split criterion values (Gini) distribution of all split candidates at the root of a tree, 'ArrowHead' dataset[3].

Acknowledgements

This PhD research is funded by **IKATS** project **- Innovative Toolkit for Analysing Time Series.**

References

1. Philippe Esling, Carlos Agon. Time-series Data Mining (2012)
2. Ahlame Douzal-Chouakria, Cécile Amblard. Classification trees for time series (2012)
3. Y. Chen, E. Keogh, B. Hu, N. Begum, A. Bagnall, A. Mueen and G. Batista (2015). The UCR Time Series Classification Archive. URL www.cs.ucr.edu/~eamonn/time_series_data/

AMA - Grande Echelle - Traitement de Données et de Connaissances à Grande Echelle