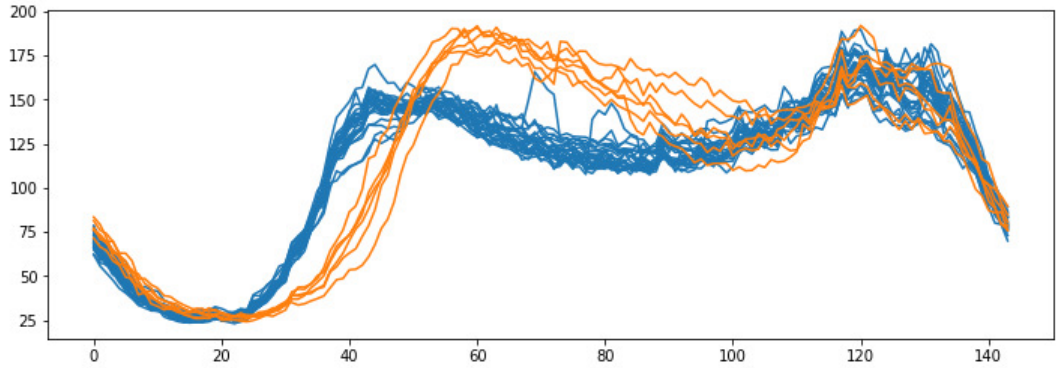
Advanced Data Mining

Piotr Lipiński

Time Series Classification

- How to classify time series?
  - see the jupyter python notebook with some examples



  - **APPROACH 1:** (for regular time series)
    - define patterns of daily (or weekly, monthly, annual) profiles for each class
    - each day (or week, month, year) time series can be matched to these patterns
    - the time series represents the class of the best matched pattern

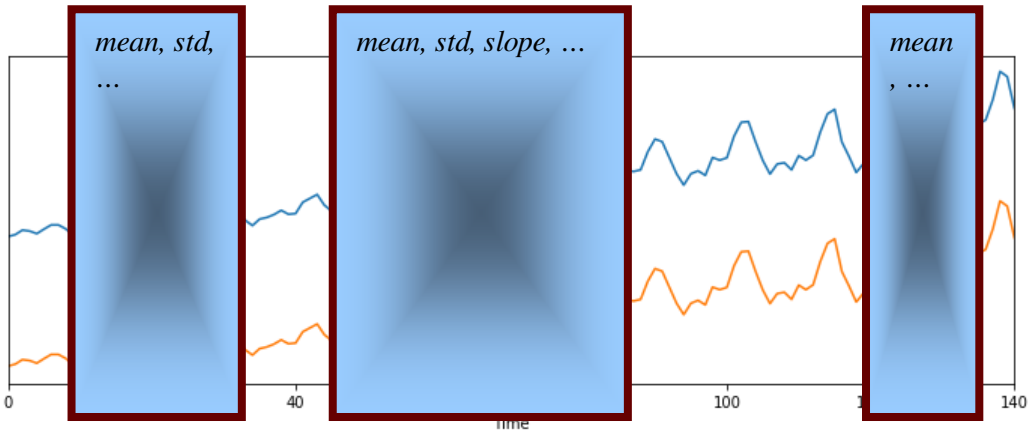
Time Series Classification

- How to classify time series?
  - **APPROACH 2:**
    - consider the time series as a vector of numbers, apply one of classic classification algorithms, e.g. kNN, SVM, etc.
    - **Difficulties:**
      - time series may be of various length
      - similar, but shifted or rescaled time series will lead to dissimilar vectors
  - **APPROACH 3:**
    - as above, but use the DTW distance measure

Time Series Classification

- How to classify time series?
  - **APPROACH 4:**
    - consider the time series in a feature-based representation
    - **Time Series Forest**

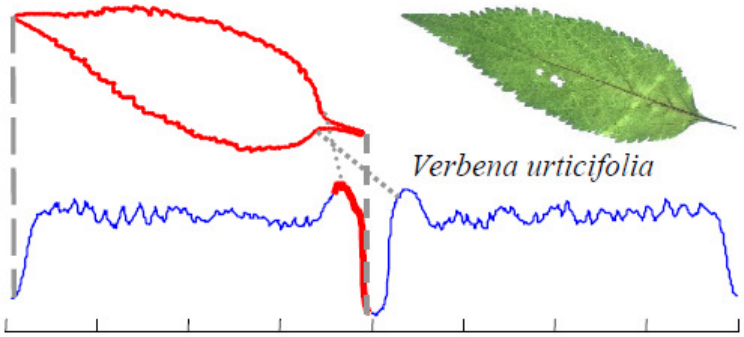
      H. Deng, G. Runger, E. Tuv, M. Vladimir, "A time series forest for classification and feature extraction". Information Science 239, 2013, pp.142-153.
      - TSF extends the **Decision Tree** and **Random Forest** classifiers
      - the time domain of the time series is splitted into a number of intervals
      - for each interval, a given set of features is evaluated (e.g. mean, std, slope)
      - this gives a vector of features describing the time series
    - similar approaches based on XGBoost or DNN



Time Series Classification

- How to classify time series?
  - **APPROACH 5:**
    - **shapelet-based representation**

      L. Ye, E. Keogh, "Time Series Shapelets: A New Primitive for Data Mining". International Conference on Knowledge Discovery and Data Mining, 15, 2009, pp.947-956.



    - Generally: *a shapelet is a characteristic pattern in time series*
    - Formally:
      - a shapelet S is a sequence $s_1, s_2, \ldots, s_L$
      - for each time series X, we can evaluate the distance between S and X, e.g. DTW(X, S), to verify whether the shapelet S is characteristic to the time series X (and how much)
    - Therefore, for a given set of shapelets, time series may be encoded in the shapelets-based representation.

Time Series Classification

- How to classify time series?
  - **APPROACH 5:**
    - **brute-force algorithm for discovering valuable shapelets**

      L. Ye, E. Keogh, "Time Series Shapelets: A New Primitive for Data Mining". International Conference on Knowledge Discovery and Data Mining, 15, 2009, pp.947-956.

      - consider a train dataset D composed of N time series $X_1, X_2, \ldots, X_N$ labelled with target values $y_1, y_2, \ldots, y_N \in \{1, 2, \ldots, C\}$, where C is the number of classes
      - for the sake of simplicity, consider the binary classification problem (C = 2)
      - similarly to the **Decision Tree** classifier, for a shapelet S and a distance threshold d, the entire train dataset may be splitted into two parts with the condition

        $$DTW(X, S) < d$$

        and the information gain (e.g. based on entropy) may be evaluated
      - for a shapelet S, the optimal split point is the distance threshold maximizing the information gain among all distance thresholds
      - the information gain of a shapelet S is the information gain for the shapelet S and its optimal split point
      - the brute-force algorithm considers the set of all possible shapelets of a given length and searches for the shapelet with the maximum information gain
      - it may be easily extended to find more than one shapelet and to general classification problems (C > 2)

Time Series Classification

- How to classify time series?
  - **APPROACH 5:**
    - **brute-force algorithm for discovering valuable shapelets**

      L. Ye, E. Keogh, "Time Series Shapelets: A New Primitive for Data Mining". International Conference on Knowledge Discovery and Data Mining, 15, 2009, pp.947-956.

```
candidates := Candidate-Shapelets(maxlen, minlen)
best_gain := 0
for each S in candidates
        gain := Information-Gain(S)
        if gain > best_gain then
        best_gain := gain
        best_shapelet := S
return best_shapelet
```

    - possible improvements by subsequence distance early abandon and admissible entropy pruning

Time Series Classification

- How to classify time series?
  - **APPROACH 6:**
    - try to define the shapelets in a learning process

      J. Grabocka, N. Schilling, M. Wistuba and L. Schmidt-Thieme, "Learning Time-Series Shapelets". International Conference on Data Mining, 14, 2014, pp.392-401.