Additional material: scalablity of the EAST representation

Xavier Renard[1,3], Maria Rifqi[2], Gabriel Fricout[3] and Marcin Detyniecki[1,4]

[1]Sorbonne Universités, UPMC Univ Paris 06, CNRS, LIP6, Paris, France.
[2]Université Panthéon Assas, Univ Paris 02, LEMMA, Paris, France.
[3]Arcelormittal Research, Maizières-lès-Metz, France.
[4]Polish Academy of Sciences, IBS PAN, Warsaw, Poland.

Demonstration

The objective is to establish a lower bound on the probability to draw a pattern $z_1 \in Z = \{z_1\}$ that appears exactly one time in all the time series of D of class $y_1 \in Y$ and only y_1. S is the set of all the subsequences that we can enumerate from D. The probability to draw a subsequence $s \in S$ that is z_1 is:

$$P(s = z_1) = \frac{|S_{z_1}|}{|S|} \tag{1}$$

Where $S_{z_1} \subseteq S$ is the set of all the subsequences $s \in S$ such as $s = z_1$. If a unique subsequence satisfies this condition by time series (most pessimistic assumption), then $|S_{z_1}| = N_{y_1}$ with $0 < N_{y_1} \leq N$ is the number of time series of class y_1 in D.

$$|S| = \frac{1}{2} \sum_{i=1}^{N} L_i(L_i + 1) \tag{2}$$

With $L_{min} \leq L_i \leq L_{max}$:

$$\frac{N * L_{min}(L_{min} + 1)}{2} \leq |S| \leq \frac{N * L_{max}(L_{max} + 1)}{2} \tag{3}$$

$$\frac{2 * N_{y_1}}{N * L_{min}(L_{min} + 1)} \geq P(s = z_1) \geq \frac{2 * N_{y_1}}{N * L_{max}(L_{max} + 1)} \tag{4}$$

$$\boxed{\frac{2 * F_{y_1}}{L_{min}(L_{min} + 1)} \geq P(s = z_1) \geq \frac{2 * F_{y_1}}{L_{max}(L_{max} + 1)}} \tag{5}$$

$F_{y_1} = \frac{N_{y_1}}{N}$ is the proportion of time series of class y_1 in D, that is specific of the use case and remains constant independently of N.

We now consider the case where several patterns $z_i \in Z$ are sought, each of them being discriminant or characteristic of a class or a set of classes. The probability to draw them all is:

$$P(Z) = \prod_{k=1}^{|Z|} P(s = z_k) \tag{6}$$

$$\boxed{\prod_{k=1}^{|Z|} \frac{2 * F_{z_k}}{L_{min}(L_{min}+1)} \geq P(Z) \geq \prod_{k=1}^{|Z|} \frac{2 * F_{z_k}}{L_{max}(L_{max}+1)}} \tag{7}$$

Where F_{z_k} is the proportion of time series for which z_k is discriminant.

Hence a lower bound on the probability to discover Z independent of the number of time series in D exists. The assumption that at most one subsequence is discriminant by time series is pessimistic. Relevant patterns may be encountered in smaller or longer enumerated subsequences from D, possibly affected by noise or time warping, while still being discriminant enough.