Symbolic Aggregate approXimation (SAX): A symbolic representation for time series

Sylvain W. Combettes, Laurent Oudre, and Charles Truong

Centre Borelli, École Normale Supérieure Paris-Saclay, Université Paris-Saclay

Objective

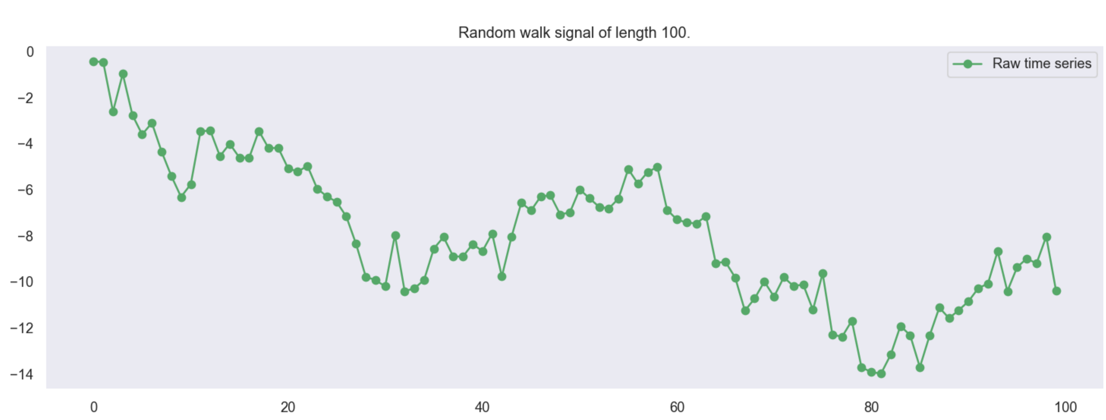
Better represent complex signals with interpretable symbols, to make them easier to analyze, and improve data mining efficiency.

Introduction

SAX was introduced in [1].

SAX allows a time series of arbitrary length n to be reduced to a string of arbitrary length w.

- $w < n$, typically $w \ll n$
- The alphabet size is also an arbitrary integer α, where $\alpha > 2$.
- w and α are chosen.



SAX

`ddcbccbaab`

Figure 1:An example of SAX representation of a signal

Why can SAX be useful?

- Symbolic representations allow researchers to benefit from data structures and algorithms from the **text processing and bioinformatics** communities.
- SAX is competitive with, or superior to, other representations on a wide variety of classic data mining problems (e.g. classification or clustering), mainly due to the **smoothing effect of dimensionality reduction**.
- SAX can deal with the **multi-modality** aspect of physiological signals.

How SAX works

SAX steps:

1. **Normalize** each time series to have a mean of zero and a standard deviation of one.
2. Transform the normalized data into the PAA (**Piecewise Aggregate Approximation**) representation: taking the mean for each of the w uniform segments.
3. Encode the PAA representation into a discrete string (**quantization**).
   - We want to produce symbols with **equiprobability**.
   - Given that the normalized time series have highly **Gaussian distribution** (assumption), we can simply determine the quantization steps β_i that will produce equal-sized areas under Gaussian curve.

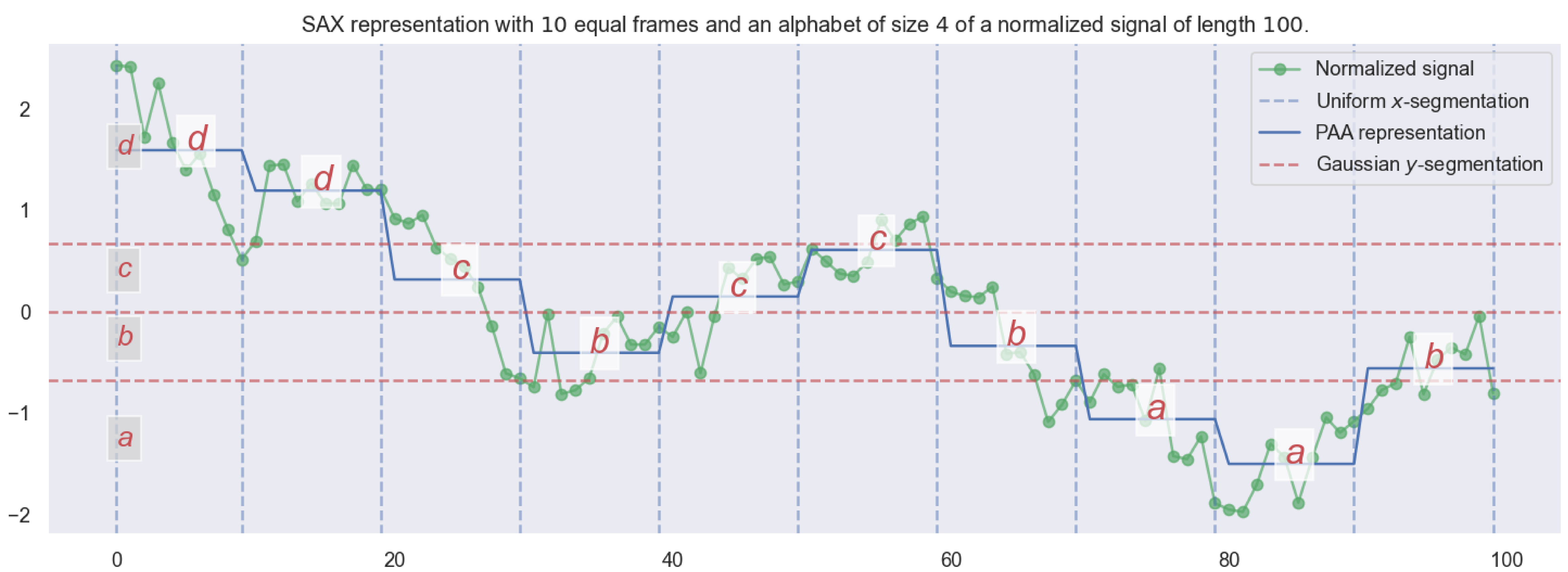


Figure 2:An example of how SAX works. SAX representation obtained here: `ddcbccbaab`.

Distance measures

Notation:

- time series $S = (s_1, \dots, s_n)$
- symbolic representation of a time series: $\bar{S} = \bar{s}_1, \dots, \bar{s}_w$

Euclidean distance on the signals:

$$EUCLIDEAN(S,T) = \sqrt{\sum_{i=1}^{n} (s_i - t_i)^2} \quad (1)$$

SAX uses the following distance:

$$MINDIST(\bar{S},\bar{T}) = \sqrt{\frac{n}{w}}\sqrt{\sum_{i=1}^{w} \left(\delta\left(\bar{t}_i, \bar{s}_i\right)\right)^2} \quad (2)$$

Note: MINDIST is more a discrepancy measure than a true distance.

The δ function is evaluated using a precomputed lookup table, for example for $\alpha = 4$:

	a	b	c	d
a	0	0	0.67	1.34
b	0	0	0	0.67
c	0.67	0	0	0
d	1.34	0.67	0	0

Note: $\delta(a,b) = 0$.

For a given value of the alphabet size α, the table needs only be calculated once, then stored for fast lookup. The value in cell (r,c) for any lookup table is given by:

$$cell_{r,c} = \begin{cases} 0 & \text{if } |r-c| \leq 1 \\ \beta_{\max(r,c)-1} - \beta_{\min(r,c)} & \text{otherwise} \end{cases} \quad (3)$$

The lower bounding property

The MINDIST distance on SAX insures the **lower bounding property**:

$$MINDIST(\bar{S},\bar{T}) \leq EUCLIDEAN(S,T) \quad (4)$$

Hence, we can assume that the similarity matching in the reduced space **maintains its meaning** (e.g. indexing of data with no false negatives).

Research objectives

Overcoming the shortcomings of SAX:

1. Taking the average value of a subsequence causes a loss of information.
2. Fixed sized frames, not adaptive segmentation.
3. SAX requires that the PAA data approximates a Gaussian distribution.
4. SAX requires to first select the number of segments w and the alphabet size of symbols α.
5. No multivariate nor multi-modal modelling.
6. The distance between a symbol and its adjacent symbol is 0.

References

[1] Jessica Lin, Eamonn J. Keogh, S. Lonardi, and B. Chiu. A symbolic representation of time series, with implications for streaming algorithms. In *DMKD '03*, 2003.

[2] C. Cassisi, P. Montalto, M. Aliotta, A. Cannata, and A. Pulvirenti. Similarity measures and dimensionality reduction techniques for time series data mining. 2012.

Contact Information

- Web: https://sylvaincom.github.io/
- Email: sylvain.combettes@ens-paris-saclay.fr