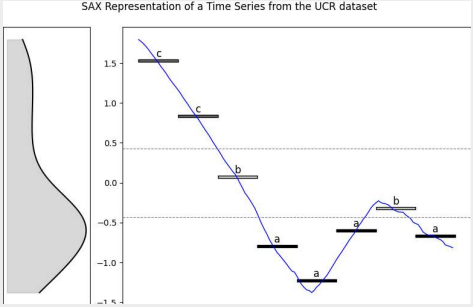
A Symbolic Representation of Time Series, with Implications for Streaming Algorithms

Caio Rocha
Prof. Paul Boniol
NoSQL 2025 class
Université PSL

Contents

Introduction

The paper by Lin et al. (2003) introduced SAX, a new symbolic representation of time series that allows dimensionality and numerosity reduction. SAX enables efficient similarity searches and indexing by transforming time series into a lower-dimensional symbolic form.



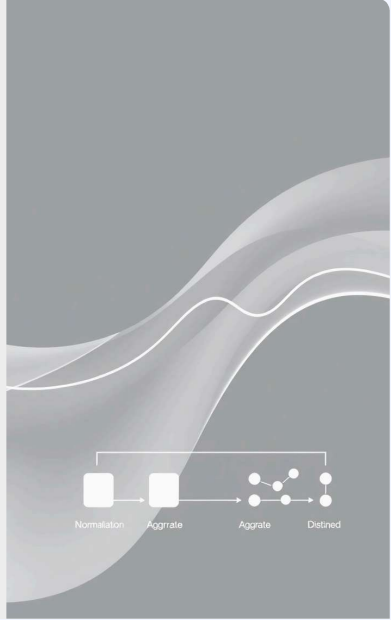
SAX Algorithm Overview

Three-step process

Normalization
Transform the data into a normalized form.

Piecewise Aggregate Approximation (PAA)
Aggregate the normalized data into equal-sized segments.

Discretization
Convert the aggregated segments into a discrete symbolic string using breakpoints.



Strong Points

Key Positive Points of the SAX paper

Simple and Effective Approach
SAX reduces time series to a lower-dimensional representation, enabling faster similarity searches and reducing storage requirements in applications.

Inclusion of Graphic Examples
Illustrative examples follow the algorithm's description, helping readers understand SAX and the mathematical formulations behind it, such as PAA.

Extensive Tests on Data Mining Techniques
Comprehensive experiments on classification, clustering, anomaly detection, motif discovery, and indexing tasks, proving SAX's competitive and efficient symbolic representation for time series data.

Weak Points

Areas for Improvement

Limited Discussion of Methods
Details on experiment methods and parameters are lacking, as well as the sources of the datasets. They are required for an accurate reproduction of the results.

Limited Exploration of Results
Figures 11 and 12 are not well-explained, making interpretation difficult. Example: it is not clear that Figure 11 supports the following claim "No other technique provides a lower bound for the distance measure on the original time series."

Limited Discussion of Drawbacks
Potential drawbacks, such as sensitivity to noise, limitations in capturing complex temporal patterns, and information loss, are not discussed.

Research Questions Addressed

Key Research Questions

Solving an Ongoing Problem
SAX introduces the first lower bounding symbolic representation, which is dimensionality reducing and can be obtained in streaming fashion.

Improving State-of-the-Art
SAX shows competitive or superior performance compared to established methods like DWT and DFT for classification, clustering, and other tasks, while requiring less storage space.

Opening New Research Directions
SAX enables the use of bioinformatics and text mining algorithms in time series analysis. We can now use hashing, Markov models, suffix trees, decision trees, and the Jaccard coefficient with SAX.

Research Questions Addressed

Key Research Questions

Comprehensive Experimental Results
Extensive tests demonstrate SAX's versatility for classic data mining techniques, though more detailed explanations of the the results and experimental setup are needed.

Missing Related Work
The authors explained PAA without citing Keogh (2001), the original reference, and overlooked Keogh's (2000) similar dimensionality reduction technique for time series indexing, which SAX extended for symbolic representation.

Algorithm

Algorithm 1 SAX

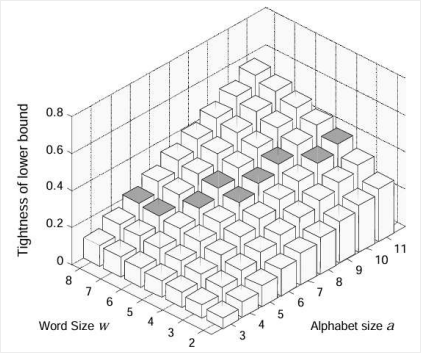
Require: Time series ts, word length w, alphabet size a
Ensure: SAX representation as a string

1: Normalize ts : $ts_{norm} \leftarrow (ts - \mu)/\sigma$
2: Compute PAA representation : $PAA \leftarrow \text{PAA}(ts_{norm}, w)$
3: Compute breakpoints : $breakpoints \leftarrow \Phi^{-1}(\{i/a\}_{i=1}^{a-1})$
4: Initialize $sax \leftarrow \emptyset$
5: **for** each p in PAA **do**
6: Determine symbol $s \leftarrow \text{digitize}(p, breakpoints)$
7: Append character $sax \leftarrow sax \cup \text{char}(s)$
8: **end for**
9: **return** sax

Figure: SAX pseudocode.

Replication of Results

Experiment 1



Paper Data

50 time series from the UCR Time Series Classification Archive.

Goal

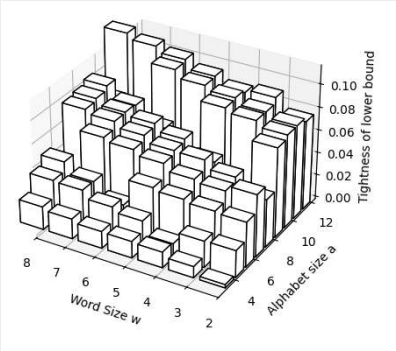
To test combinations of 7 word sizes and 11 alphabet sizes, calculating the tightness of SAX lower bounds for each (w, a) pair, with 100.000 combinations.

Expected Results

The tightness of lower bound increases linearly with a and w.

Replication of Results

Experiment 1



My Data

51 random time series from the UCR Time Series Classification Archive.

Results

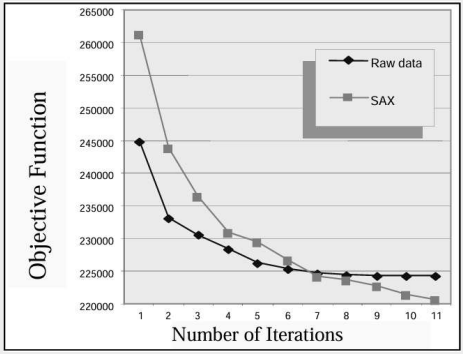
Some differences compared to the original plot, such as a lower z scale (0.12 vs. 0.55) and some (w, a) pairs with low values in the middle of the plot.

Possible causes

Variations in dataset selection or normalization.

Replication of Results

Experiment 2

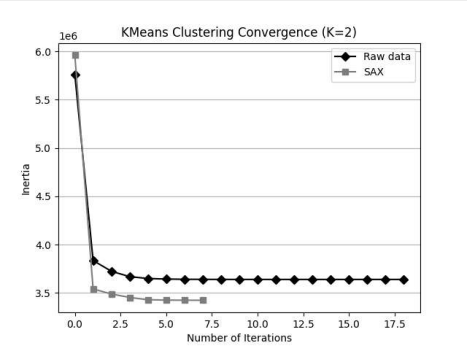


Paper Data
Space Shuttle telemetry, 1,000 subsequences of length 512.

Goal
Perform k-means on both the original raw data and SAX.

Expected Results
SAX gives better results than working with the original data.

Replication of Results



My Data

Household power consumption dataset [Learning, 2016], which shows measurements of electric power consumption in one household with a one-minute sampling rate over a period of approximately 4 years. Subsequences of 512 observations.

Results

K-Means performed better with SAX as concluded in the paper. Why? The simplified symbolic representation is useful for a faster convergence of the K-Means algorithm.

Conclusion

The SAX paper provides a novel symbolic representation for time series, allowing effective data mining applications. While the reproduced experiments confirmed SAX's advantages, the absence of dataset sources complicated full replication. Providing more detailed discussions on methods, results, and drawbacks would benefit the scientific community.

Sources

[Keogh, 2001] Eamonn Keogh, Kaushik Chakrabarti, Michael Pazzani et Sharad Mehrotra. « Dimensionality Reduction for Fast Similarity Search in Large Time Series Databases ». 2001 (cf. p. 3).

[Keogh, 2000] Eamonn J. Keogh et Michael J. Pazzani. « A Simple Dimensionality Reduction Technique for Fast Similarity Search in Large Time Series Databases ». Knowledge Discovery and Data Mining. Current Issues and New Applications. Sous la dir. de Takao Terano, Huan Liu et Arbee L. P. Chen. Berlin, Heidelberg, 2000, p. 122-133 (cf. p. 3).

[Learning, 2016] UCI Machine Learning. Electric Power Consumption Data Set. Kaggle. 2016 (cf. p. 4).

[Lin, 2003] Jessica Lin, Eamonn Keogh, Stefano Lonardi et Bill Chiu. « A symbolic representation of time series, with implications for streaming algorithms ». Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery. DMKD '03. San Diego, California : Association for Computing Machinery, 2003, p. 2-11 (cf. p. 1).