Préparée à Université PSL

Research Article Replication: “A Symbolic Representation of Time Series, with Implications for Streaming Algorithms”

Student

Caio Rocha

9 mars 2025

Professor

Paul Boniol

PSL

1 Introduction

Methods The paper [Lin, 2003] introduced SAX, a new symbolic representation of time series that allows dimensionality and numerosity reduction. The symbolic approach also allows distance measures that lower bound corresponding distance measures defined in the original series. The authors introduced an algorithm that allows a time series of arbitrary length n to be represented as a string of arbitrary length w containing a characters, where $w < n$ and $a > 2$. This is done by a discretization procedure that first transforms the data into the Piecewise Aggregate Approximation (PAA) and then symbolize the PAA into a discrete string.

Results The introduction of a symbolic representation that lower bounds the Euclidean distance between the original time series, allowing for efficient similarity search and indexing. In addition, given a single time series, their approach leads to dimensionality and numerosity reduction (reduction of the volume of the original data). Several data mining techniques were compared with SAX, highlighting its similar or superior performance across multiple applications, such as clustering, classification and anomaly detection.

2 Strong Points

Simple and Effective Approach : The algorithm behind SAX is simple but powerful, lying in a three-step process : normalization, Piecewise Aggregate Approximation (PAA), and discretization. It is computationally efficient because it reduces the time series to a lower-dimensional representation using PAA, which simply computes the mean of equal-sized segments of the series and then maps the series into a symbolic representation using a lookup table of breakpoints. This newly created compressed representation speeds up similarity searches and reduces the required storage size.

Inclusion of Graphic Examples : The instructive description of the algorithm and its advantages is accompanied by illustrative examples, which helps the reader associate the mathematical formulations with the transformations incurred in the time series. For instance, Figure 3 is a clear graphic representation of the PAA step applied to reduce the time series from n dimensions to w dimensions by dividing the data into w equal sized “frames”. Figure 5 also illustrates well how a final SAX representation looks like. Figure 6 and 7 represent how SAX and its parameters relate to the distance measures and the tightness of lower bounds.

Extensive Tests on Data Mining Techniques : The authors conducted comprehensive experiments on classification, clustering, anomaly detection, motif discovery, and indexing tasks. These tests were crucial in proving the article’s main point that SAX is a powerful and efficient symbolic representation for time series data. The results showed that SAX is competitive with, or superior to, other representations on a wide variety of data mining problems, proving the quality of the approximation and the utility of this framework. SAX also has the benefit of being applicable to the generic time series data mining approach introduced in Table 1.

3 Weak Points

Limited Discussion of the Methods used for the Experiments : Although SAX was tested on several data mining problems, the article lacked a more detailed discussion of the methods employed for each experiment, possibly because of space constraints in the paper. The authors proposed several adaptations of SAX to address these problems, but a comprehensible and step-by-step description of these modifications would have been highly valuable to the scientific community, as well as the parameters used. Including the sources of the datasets used is also crucial. This would allow for an easier reproduction of the results.

Limited Exploration of Results : The results presented in Figures 11 and 12 are left to open interpretation. For instance, it is not clear at first if / how Figure 11 supports the authors' claim in Section 2.2. that none of the other techniques allows a distance measure that lower bounds a distance measure defined on the original time series. Even though we can understand what is being shown after a careful analysis, it would be beneficial to have short explanations after each figure. This would help consolidate the authors' arguments and would allow for an easier comprehension of the experiments' conclusions.

Limited Discussion of SAX's drawbacks : The paper falls short in completely addressing SAX's limitations, focusing primarily on the issue of normalizing near-constant subsequences in Section 3.4. While it proposes a solution for this specific problem, it neglects to discuss other potential drawbacks such as information loss during discretization, sensitivity to noise, or limitations in capturing complex temporal patterns. A more comprehensive examination of SAX's weaknesses would have been beneficial to the scientific community, allowing researchers to better understand what can be improved.

4 Questions

a. Is the paper solving an ongoing research problem ? Yes, the SAX paper addressed an ongoing research problem in time series data mining. It introduced the first lower bounding symbolic representation, which is also dimensionality reducing and can be obtained in streaming fashion. Therefore, it allows one to run classical data mining algorithms on the obtained symbolic representation, while producing comparable results to the algorithms that operate on the original data.

b. Is the paper improving existing state-of-the-art performances of a task ? Yes, the paper improved existing state-of-the-art methods for mining time series data, which prevously explicitly assumes that the time series data is real valued, limiting its applications to streaming. SAX circumvented that by allowing an accurate representation the data in a symbolic space. It demonstrated competitive or superior results compared to established methods like Discrete Wavelet Transform (DWT) and Discrete Fourier Transform (DFT) for classification, clustering, and other tasks, while requiring less storage space.

c. Is the paper opening new research directions and problems ? Yes, SAX also opens up new research directions because it enables the use of algorithms from bioinformatics and text mining in time series analysis. For example, we can now use hashing, Markov models, suffix trees, decision trees, and the Jaccard coefficient for comparing and analyzing time series. The authors highlight the possibility of creating a lower bounding approximation of Dynamic Time Warping by slightly modifying the classic string edit distance, and the utility to representing multidimensional time series.

d. Are the experimental results covering the claim of the paper ? Yes, the experimental results presented in the paper covered its claims. The authors conducted comprehensive tests on classic data mining techniques, e.g. classification, clustering, anomaly detection, motif discovery, and indexing tasks. These experiments demonstrated SAX's versatility across different domains and datasets. Nevertheless, the paper could benefit from a more detailed explanation of the the results. Even though the experiments were extensive, the paper lacked an in-depth discussion of the experimental setup and use cases.

e. Are there missing related work ? Yes, even though the paper provided a thorough overview of SAX and other approaches, some related papers were left out. The authors explained in the article the PAA (Piecewise Aggregate Approximation), but they did not reference the original paper [Keogh, 2001]. In addition, [Keogh, 2000] proposed a dimensionality reduction technique that allows fast indexing of time series. The step of calculating the mean of segments as an intermediary dimensionality reduction step is very similar to SAX's, though SAX extended this technique for a symbolic representation support.

5 Implementation

Below is a pseudocode for the SAX algorithm followed by an explanation.

Algorithm 1 SAX

```
Require: Time series $ts$, word length $w$, alphabet size $a$
Ensure: SAX representation as a string
Normalize $ts$ : $ts_{norm} \leftarrow (ts - \mu)/\sigma$
Compute PAA representation : $PAA \leftarrow \mathrm{PAA}(ts_{norm}, w)$
Compute breakpoints : $breakpoints \leftarrow \Phi^{-1}(\{i/a\}_{i=1}^{a-1})$
Initialize $sax \leftarrow \emptyset$
for each $p$ in $PAA$ do
    Determine symbol $s \leftarrow \mathrm{digitize}(p, breakpoints)$
    Append character $sax \leftarrow sax \cup \mathrm{char}(s)$
end for
return $sax$
```

where Φ^{-1} is the inverse cumulative distribution function of a standard normal distribution.

6 Reproduction of the Results

The replicated results were the experiment represented in Figure 7 of the paper and the one represented in Figure 12, since they cover the authors' main claims.

In Figure 7, the authors show that SAX provides a tight lower bound of corresponding distance measures. They chose a dataset of 50 time series from the UCR Time Series Classification Archive and tested combinations of 7 word sizes and 11 alphabet sizes. They computed the tightness of lower bounds given every time series combination and averaged the results for each (w, a) pair. The total number of combinations was of order of magnitude 10^5. Because I did not have access to the exact time series that were used in the experiment, I chose 51 time series from the UCR database and did a loop calculating the average tightness of lower bounds for every (w, a) pair, obtaining Figure 1. The reason why the results are a bit different could be the dataset selection or the normalization procedure, even though I tried to follow exactly the paper's descriptions.

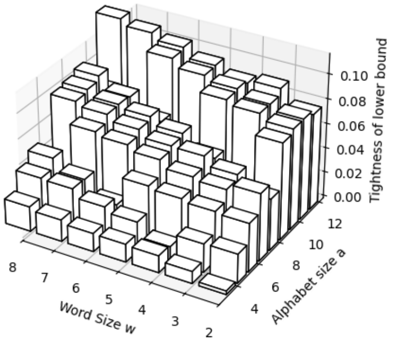

Figure 1 – Reproduction of Figure 7. Even though it has the same pattern, the z scale is different (my highest tightness of lower bound is 0.12 vs 0.55 in the paper) and there are some (w, a) pairs with low values in the middle of the plot.

The following experiment refers to the effectiveness of using SAX for K-Means clustering. The authors conducted a comparison between SAX versus raw data, employing a dataset of 1,000 subsequences of length 512 from Space Shuttle telemetry. They even found that clustering performed on the SAX approximation yielded better results than when applied to the original raw data. I couldn't find the exact dataset that the authors used either, so I used instead a 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. As done in the paper, I split it into subsequences of 512 observations (rows). One important parameter that is not specified in the paper is the number of clusters, which in this case I used $K = 2$, obtaining Figure 2.

As in the paper, K-Means performed better with SAX than with the raw data, which is reasonable since SAX provides a representative simplification of the time series, which is useful for a faster convergence of the K-Means algorithm.

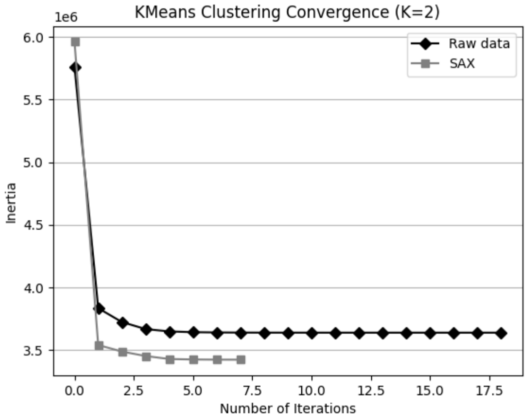

FIGURE 2 – K-Means convergence with K=2.

7 Conclusion

The SAX paper is well-structured and introduces a novel symbolic representation of time series that allows effective application of data mining techniques unlike previous representation algorithms. The reproduction of Figure 7 confirmed the accuracy of SAX's lower bound but showed differences in scale, possibly due to dataset selection or normalization. The reproduction of the K-Means clustering experiment also supported the claim that SAX improves clustering performance, though the exact dataset used by the authors was unavailable. The lack of dataset sources in the original paper made full replication difficult, so including these references in the final version would be useful.

Bibliographie

[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 : Springer 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).