Dauphine | PSL

Masters 2 IASD

Project: SAX Paper Replication

Ms. REZKELLAH Fatma-Zohra Cohort: 2024/2025

Contents

1 Introduction

1.1 Motivation

In their paper "A Symbolic Representation of Time Series, with Implications for Streaming Algorithms," *Jessica Lin et al.* address the growing need for efficient methods to analyze time series data, which is prevalent in various fields such as finance, healthcare, and environmental monitoring. Time series data poses unique challenges due to its high dimensionality, temporal dependencies, and the necessity for real-time analysis. Traditional methods, such as **dynamic time warping (DTW)**, while effective, often suffer from high computational complexity, making them impractical for large datasets.

To tackle these challenges, the authors introduce the **Symbolic Aggregate approXimation (SAX)** method, a novel approach that transforms continuous time series data into a symbolic representation. SAX simplifies the analysis of time series data by discretizing it into a sequence of symbols that retain essential characteristics of the original series while reducing its dimensionality. This transformation not only facilitates more efficient processing but also enables the application of various data mining techniques, such as clustering, classification, and indexing.

1.2 Innovation & Proposed Solution

The SAX method consists of two main steps: Piecewise Aggregate Approximation (PAA) and quantization:

1. **Piecewise Aggregate Approximation:** In this step, the time series is divided into equal-length segments. The average value of each segment is calculated, resulting in a reduced representation of the original time series. This step effectively reduces the dimensionality of the data while preserving its overall trend and patterns.
2. **Discretization:**The PAA values are then quantized into discrete symbols based on a predefined set of breakpoints. These breakpoints can be determined using predefined thresholds, or probability distributions (typically determined based on the standard Gaussian distribution allowing for flexibility in the representation). The resulting symbolic representation makes it easier to analyze, compare, and query time series data without losing critical information.

1.3 Experimentations & Results

The authors conduct extensive experiments to evaluate the performance of SAX compared to traditional methods (Euclidean Distance, and other symbolic approaches (SDA, IMPACTS)), particularly focusing on five data mining tasks: Clustering, Classification, Query Indexing, Anomaly Detection, and Motif Discovery. The experimental results demonstrate that SAX significantly reduces the computational time required for these tasks while maintaining a high level of accuracy (The superiority of SAX and its ability to rival more sophisticated techniques).

2 Comments

2.1 Strengths

1. **Efficient Data Representation through Dimensionality and Numerosity Reduction:** SAX effectively reduces the complexity of time series data by first applying Piecewise Aggregate Approximation (PAA) to achieve dimensionality reduction, summarizing local trends while preserving essential patterns. It then applies discretization, converting numerical values into a compact symbolic form. This process also introduces numerosity reduction by eliminating redundant information, making SAX computationally efficient for large-scale datasets while retaining key structural properties.

2. **Robustness to Noise:** SAX's ability to generalize time series data points through symbolic representation enables it to perform resiliently in the presence of noise. This property is crucial in real-world applications where data can be erratic and unpredictable. The lower bounding property of SAX (MINDIST) ensures that the distance measures are a reasonable approximation of the original data, enabling effective analysis despite potential disturbances in the dataset.

3. **Facilitating Advanced Analyses:** The symbolic representation enables advanced analyses, such as motif discovery algorithms, which rely on discrete data forms. For instance, algorithms like PROJECTION leverage the discretization provided by SAX to perform motif discovery efficiently. This capability suggests that SAX not only simplifies the data but also enhances the potential for discovering significant patterns within time series data.

2.2 Weaknesses

1. **Dependence on Parameter Selection:** SAX's performance is heavily reliant on the selection of parameters, including the word size and the alphabet size. Poorly chosen parameters can lead to suboptimal representations, resulting in significant information loss or inaccuracies in analysis. This sensitivity can be a limitation for users who may not have the expertise to tune these parameters effectively.

   Improvement: Developing adaptive methods that dynamically adjust these parameters based on dataset characteristics could improve SAX's usability and robustness. Additionally, providing guidelines or tools for parameter selection would empower users to make informed decisions and enhance the overall effectiveness of SAX in practice.

2. **Underexplored Topics:** The paper touches on several important topics in its abstract and introduction, such as streaming algorithms, yet fails to adequately address them in the body of the text. While SAX holds great promise for applications in streaming data analysis, the authors could have done more to demonstrate this potential. The lack of emphasis on streaming algorithms leaves a significant gap in understanding the full scope of SAX's applicability.

   Improvement: Future iterations of the paper should delve deeper into these topics, providing examples and case studies that showcase SAX's effectiveness in streaming environments. This could strengthen the paper's contributions and underscore

SAX's relevance in contemporary data analysis contexts.

3. **Overall Organization:** One notable criticism of the paper is the significant redundancy present in the abstract and introduction sections. The repetitive nature of the content can give readers the impression of copy-pasting rather than a cohesive narrative, which detracts from the overall readability and flow of the paper. Furthermore, the paper, while informative, was at times more difficult to read than necessary. During my review, I found myself needing to read certain sections multiple times to fully comprehend the material.

   Improvement: Simplifying the language and structure could enhance accessibility for a broader audience, including those less familiar with the topic. In particular, the paper would benefit from better examples to clarify complex concepts. For instance, much of the discussion in Section 3 focuses on constructing a symbolic representation of a time series using an alpha size of three (yielding three possible values: a, b, or c). However, when the authors introduce the distance table (Table 4) to facilitate distance measures between symbolic pairs, they utilize an alphabet of size four. This inconsistency can confuse readers, as they may struggle to understand how the examples align with the theoretical discussions. Moreover, Figure 5, which also employs an alphabet of size three, could have been more effectively compared to Table 4 to illustrate the concepts more clearly.

3 Answers to the Specific Questions

3.1 Is the paper solving an ongoing research problem?

Yes. The paper addresses the challenge of efficiently representing and analyzing large time series datasets. While traditional methods for time series analysis are computationally expensive and memory-intensive, SAX provides a compact, symbolic representation that reduces dimensionality while preserving essential features.

3.2 Is the paper improving existing state-of-the-art performances of a task?

Yes. SAX improves upon traditional methods like DFT (Discrete Fourier Transform) and DWT (Discrete Wavelet Transform) in terms of efficiency and interpretability. Before SAX, DFT and DWT were commonly used for dimensionality reduction but lacked interpretability and were computationally expensive (SAX combines dimensionality reduction with symbolic representation, enabling efficient storage (Less memory footprint, indexing, and analysis)).

3.3 Is the paper opening new research directions and problems?

Yes. SAX has inspired research in:

- Adaptive SAX variants (e.g., adaptive alphabet size, dynamic segmentation).
- Integration with machine learning models and NLP techniques for improved time series analysis.

3.4 Are the Experimental Results Valid?

Yes. The experimental results presented in the paper confirm the claims regarding SAX's efficiency and effectiveness in various tasks, showing considerable improvements in processing time while maintaining classification accuracy. However, as stated in the weaknesses points, One would like to have more details about testing this framework with streaming data as stated in the abstract and introduction.

3.5 Are there missing related work?

No, the paper does not appear to overlook significant previous works in the field. It effectively cites several foundational studies relevant to the development and context of the SAX method.

4 Final Decision

Almost all the questions asked in the previous section were answered with a "Yes", therefore, I **accept** this paper.

5 Reproducibility of Experimental Evaluations

In this section, I detail the experiments I reproduced based on the SAX paper. I focused on the most significant experiments that substantiate the paper's key claims.

5.1 Implementation of SAX Representation

I first implemented the SAX representation and tested it on some basic generated time series. The implementation seems to work correctly. Figure 1 shows the different generated time series and their SAX representations.

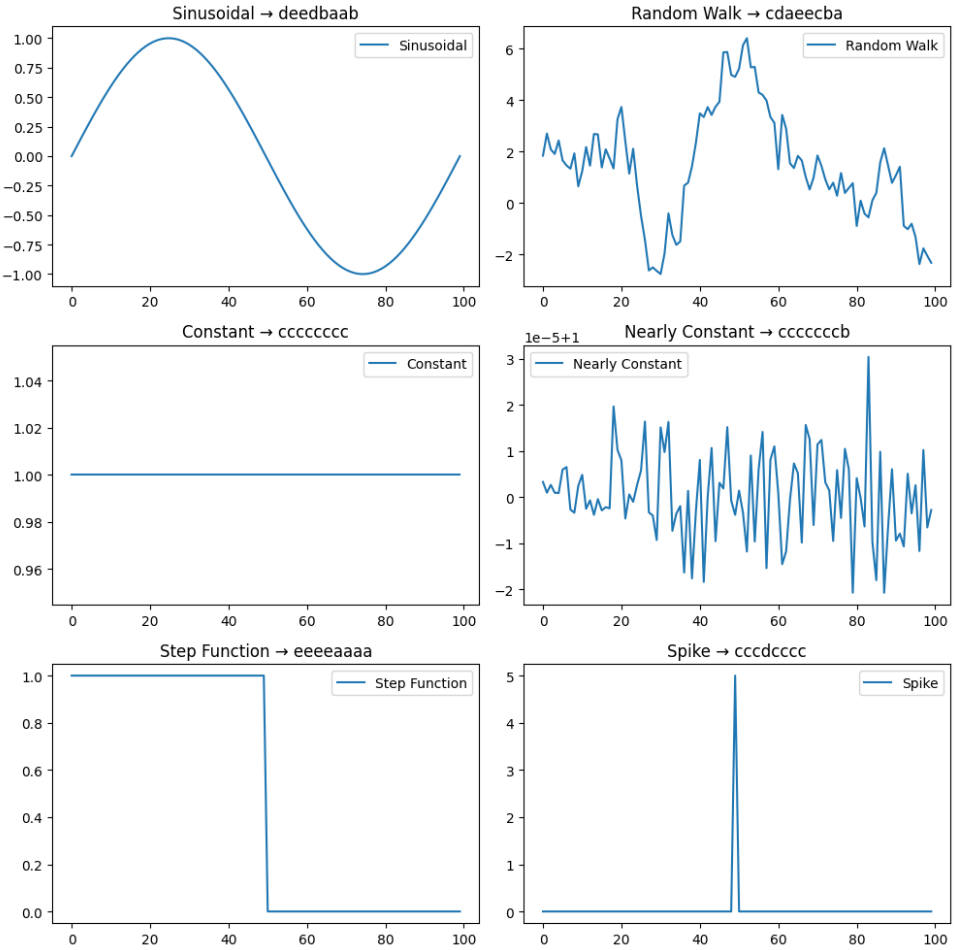

Figure 1: Different generated time series SAX

5.2 Clustering Experiments

I replicated the clustering experiments presented in the paper on the control chart dataset. I ran the hierarchical clustering algorithm on raw data using Euclidean distance, as well as the two symbolic representations SDA and Impacts, alongside SAX. The dendrograms obtained (Figures 2) affirm the claims of the paper that SAX is superior, as it correctly assigns each class to its own subtree. This is a side effect due to the smoothing effect of dimensionality reduction. More generally, I observed that SAX closely mimics Euclidean distance, while the two other methods were a bit far from that.

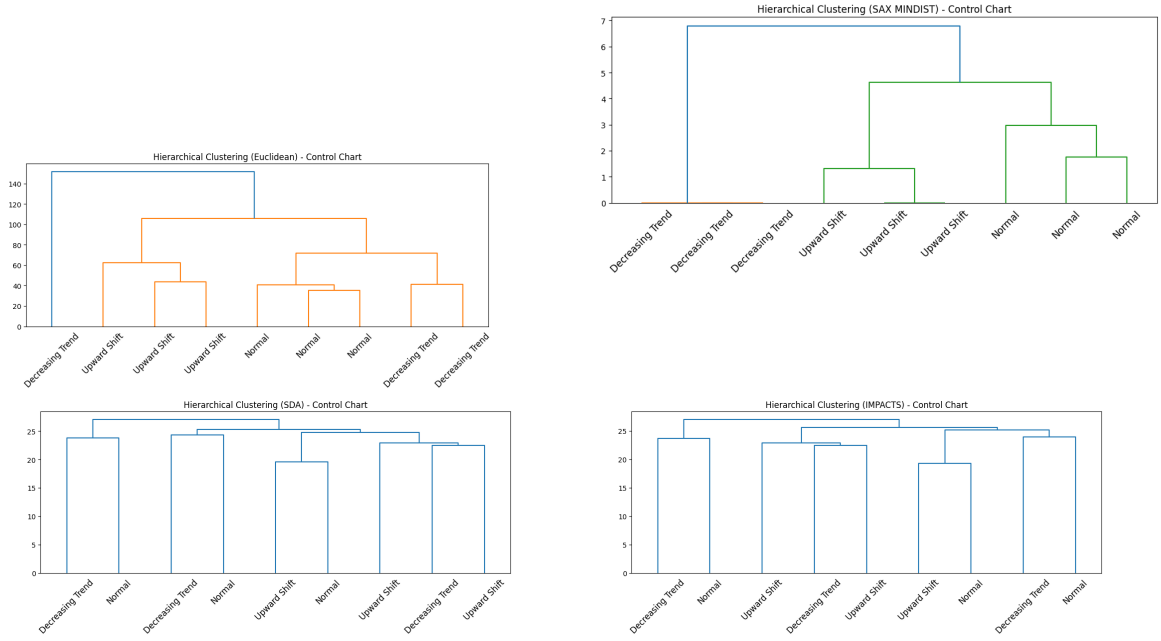

Figure 2: Dendrograms obtained from different methods

The second partitional clustering was tested using the k-means model on space telemetry data. However, since this data wasn't publicly available, I generated synthetic data following the description of the dataset used in the paper. Figure 3 shows the generated dataset. In this case, the paper only tests Euclidean distance on raw data and the SAX representation. I also tested the two other methods, SDA and Impacts. Again, the claims of the paper were confirmed: SAX outperformed Euclidean distance, while the other two methods performed poorly. This was explained by the fact that initializing the cluster centers on a low-dimensional approximation of the data can improve quality, which is what clustering with SAX implicitly does.

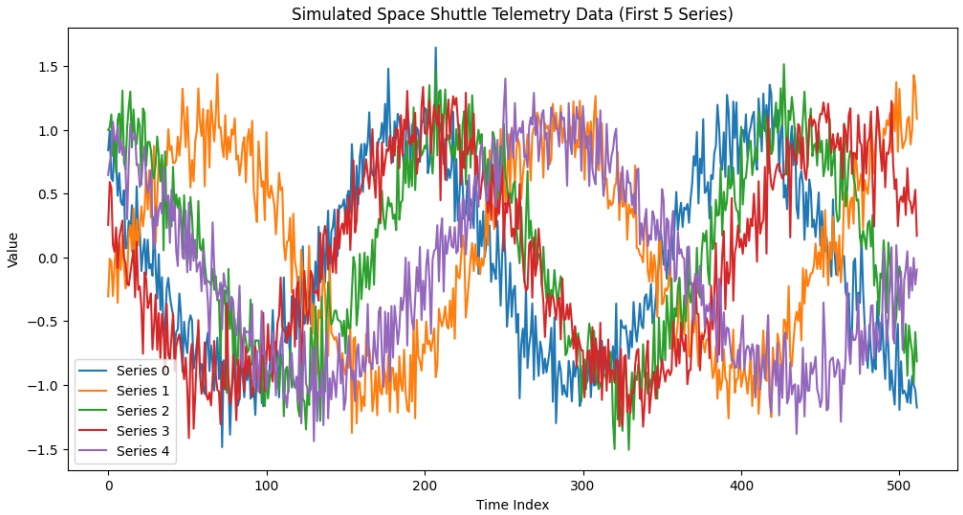

Figure 3: Generated synthetic dataset for clustering

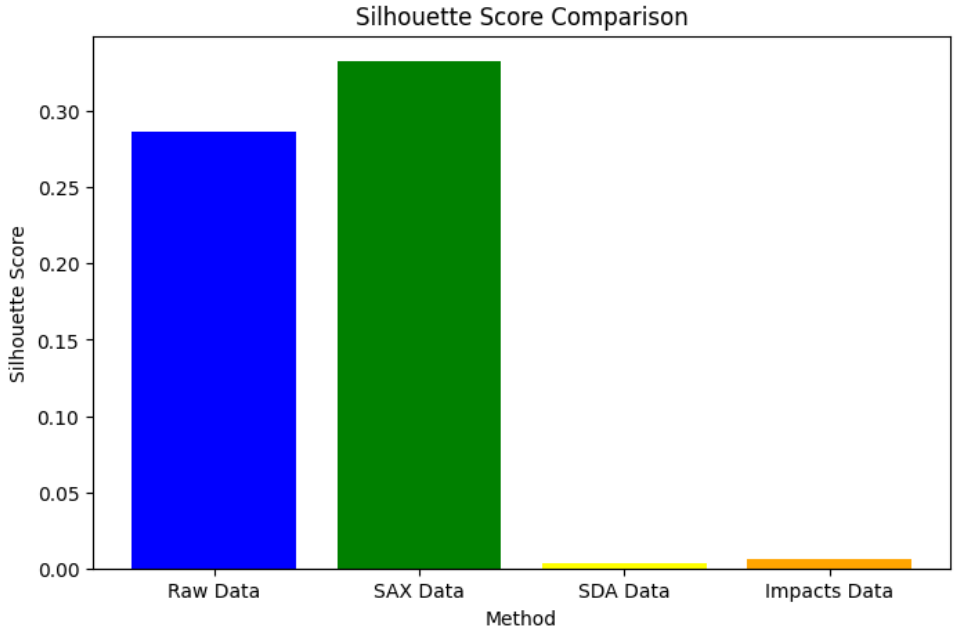

Figure 4: Silhouette Scores for Kmeans on the Different Methods

5.3 1-NN Classification Experiment

The third experiment tested is 1-NN classification using leave-one-out on the CBF dataset, comparing Euclidean distance, SDA with a fixed alphabet size of 5, LPinfinity, Impacts, and SAX with different alphabet sizes. The experiments yielded results consistent with those in the paper, showing that SAX's ability to outperform Euclidean distance is likely due to the smoothing effect of dimensionality reduction.

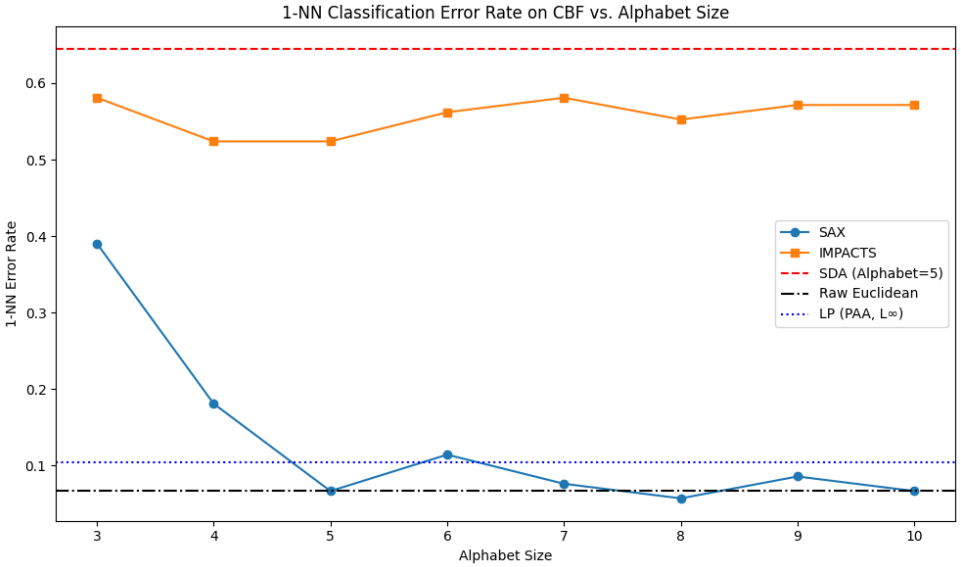

Figure 5: 1NN Classification Results

5.4 Decision Tree Classification

I also tested decision tree classification between SAX and regressor trees on raw data, using the CBF dataset. The results I obtained were similar to those in the paper. I found that the decision tree on the regressor tree representation of the data slightly outperformed SAX representation.

Model	**Accuracy**
Regressor Tree	86.67%
Decision Tree on SAX	84.44%

Table 1: Accuracy of Models

5.5 Query by Content Experiment

I was unable to reproduce the experiment on query by content, as the code for the Haar wavelet approach did not work for me. I likely made an error, and I also couldn't correctly implement the vector approximation file indexing algorithm.

5.6 Anomaly Detection Experiment

The last experiment I conducted was for anomaly detection. I generated synthetic sine wave time series data and perturbed it as discussed in the paper to obtain abnormal data. I trained a Markov model on SAX, SDA, and Impacts to test whether the anomalies were detected. This experiment was challenging, as I found the code for the two other approaches used in comparison (TSA tree wavelet and IMM) difficult to implement correctly. The paper did not explicitly mention the parameters used for each method to obtain the reported results, so I had to perform hyperparameter tuning. The results I obtained were not optimal, likely due to my lack of knowledge regarding the scale of hyperparameters (e.g., threshold for the Markov model, alphabet sizes, word sizes) and possibly also due to the implementation of the Markov model itself. Consequently, the results differed significantly from those in the paper, as all methods failed to capture the anomalies correctly, with SDA and Impacts performing the worst.

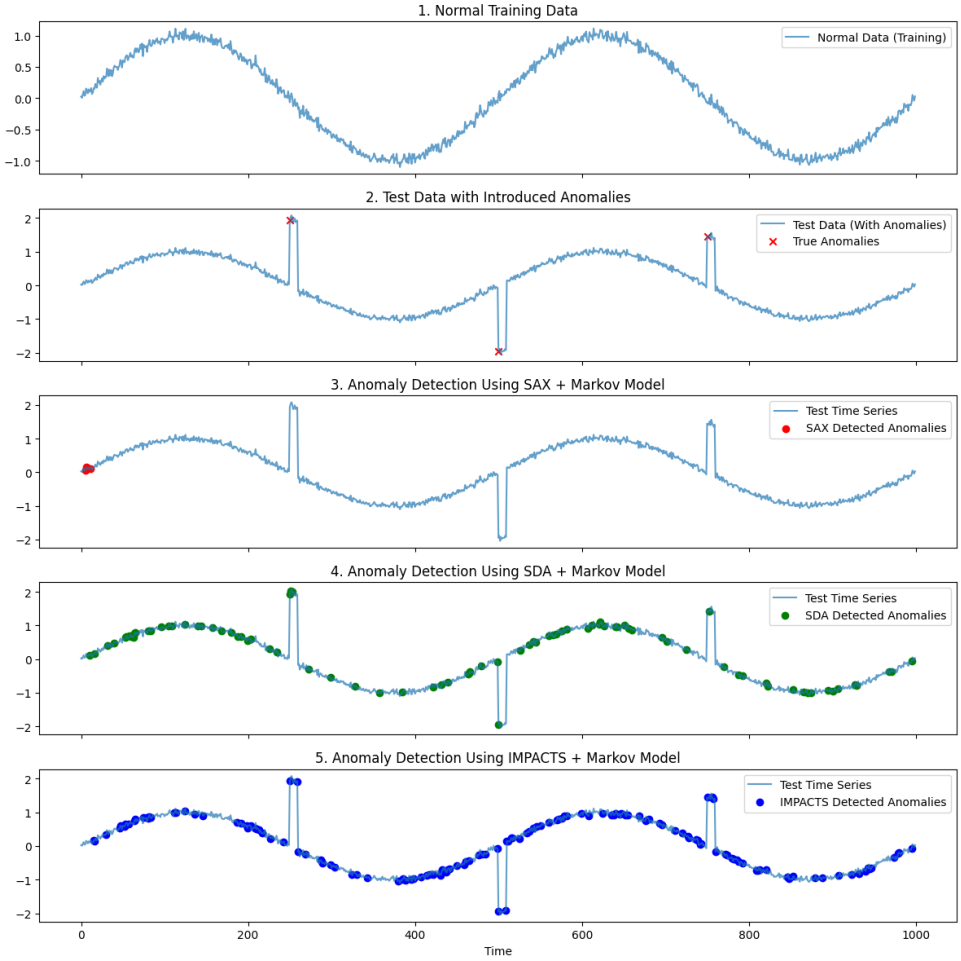

Figure 6: Anomaly Detection Results

5.7 Summary of Findings

Overall, while I was able to reproduce many aspects of the experiments outlined in the SAX paper, there were discrepancies in certain results, particularly in the anomaly detection experiments. This highlights the challenges of reproducibility in experimental research.