Symbolic vs Numeric Time–Series Clustering Using SAX and nSDL

Universidad Politécnica de Madrid — Escuela Técnica Superior de Ingenieros Informáticos

Ottavia Biagi, Leandro Duarte, Emanuele Alberti
Master's Computer Engineering and Data Science
Data Mining
Universidad Politécnica de Madrid (UPM)

***Abstract*—This work extends the previous clustering assignment by applying symbolic representations to synthetic time series from two domains: FAANG stock-like trajectories and City Hotel ADR (Average Daily Rate). Using the DMonTS tool, we preprocess the sequences using SAX (Symbolic Aggregate approXimation) and nSDL (Normalized Segment Difference Labels), and analyse their suitability for domain separation. Additionally, we implement a real symbolic clustering experiment using SAX and K-Means in Python to quantify separability. Results reveal different strengths: SAX best captures the smooth seasonal pattern of Hotel ADR, while nSDL is more expressive for financial-like series dominated by trend and local fluctuations. The study highlights the trade-offs between numeric and symbolic encodings and how segmentation and alphabet choices affect robustness.**

I. Introduction

This assignment focuses on symbolic preprocessing of time series and its impact on clustering performance. We reuse the synthetic dataset generated in the previous task: 15 FAANG-like time series and 15 City Hotel ADR time series, each with length 100. Following the feedback from the previous assignment, we increased the difficulty by injecting higher variance and stronger noise, making domain separation more challenging.

The two domains exhibit fundamentally different structures:

- **FAANG**: a monotonic decreasing trend, a mid-series valley, and irregular high–frequency fluctuations;
- **Hotel ADR**: a smooth cosine-like seasonal cycle with low noise and mild trend.

Symbolic methods compress numeric sequences into discrete words. This representation can highlight global shape (SAX) or local directional behaviour (nSDL). The goal is to determine which symbolic representation better preserves domain structure and leads to clearer cluster separation.

II. Symbolic Representations

A. SAX

SAX applies three steps: z-normalisation, PAA segmentation, and symbol assignment using Gaussian breakpoints that divide the standard normal distribution into equiprobable regions. Small alphabets produce coarse symbols; large alphabets increase precision but risk overfitting noise.

We tested segment sizes **20, 40, 60, 80** and alphabet sizes **4 and 10**.

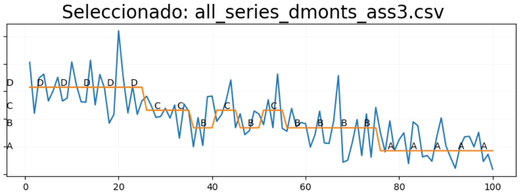

Fig. 1: FAANG — SAX (20 segments, alphabet 4).

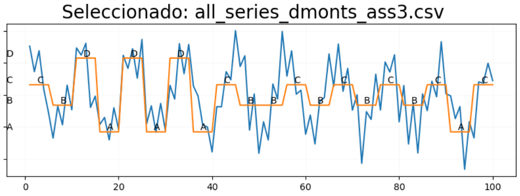

Fig. 2: Hotel ADR — SAX (20 segments, alphabet 4).

SAX(20,4) provides the clearest global structure for both domains: FAANG's trend and valley remain visible, while Hotel ADR's seasonality is well preserved.

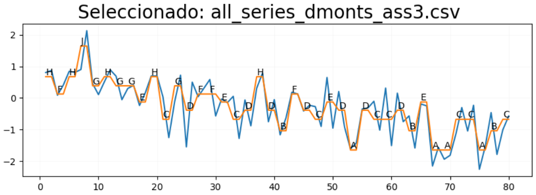

Fig. 3: FAANG — SAX (40 segments, alphabet 10). Overfitting begins.

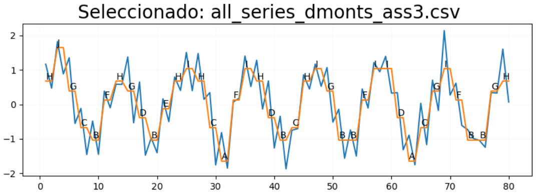

Fig. 4: Hotel ADR — SAX (40 segments, alphabet 10). Seasonal structure remains.

B. *nSDL*

nSDL assigns symbols based on local segment differences. Each segment is labelled as `up`, `down`, or `stable`, optionally with graded magnitudes (e.g., `up0`–`up10`). This makes nSDL highly expressive for series dominated by directional changes.

We tested segment sizes **8, 20, 40** and symbolic alphabets **2, 4, 20**.

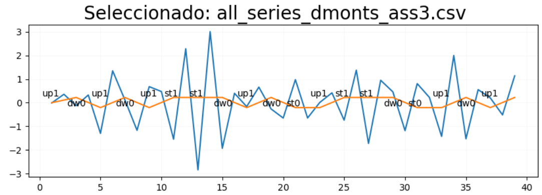

Fig. 5: FAANG — nSDL (20 segments, 2 symbols). Directional structure preserved.

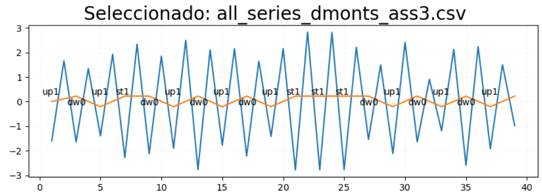

Fig. 6: Hotel ADR — nSDL (20 segments, 2 symbols). Smooth oscillation visible.

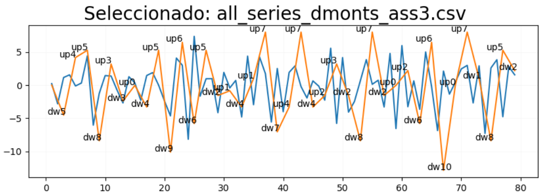

Fig. 7: FAANG — nSDL (40 segments, 20 symbols). Highly sensitive to noise.

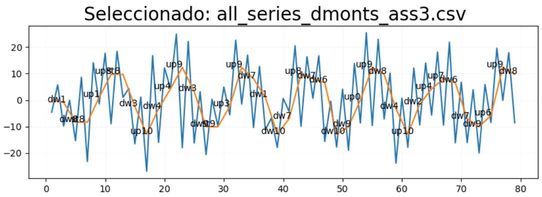

Fig. 8: Hotel ADR — nSDL (40 segments, 20 symbols). Symbol inflation amplifies noise.

III. Symbolic Clustering Experiment (Python Results)

To complement the qualitative DMonTS evaluation, we implemented a real symbolic clustering experiment in Python using SAX (20 segments, alphabet 4) and K-Means with $K = 2$. Each symbolic string was converted into a bag-of-symbols feature vector.

The confusion matrix obtained is shown in Table I.

TABLE I: SAX clustering (20 segments, alphabet 4), $K = 2$.

	Cluster C0	Cluster C1
FAANG	4	11
Hotel	11	4

Overall and per-domain purities:

- **Overall purity:** 0.733
- **FAANG purity:** 0.733
- **Hotel purity:** 0.733

The SAX(20,4) representation does not cleanly separate the two domains numerically. This is expected: bag-of-symbols ignores ordering and cannot capture seasonality or trend direction as effectively as numeric Fourier/PAA.

IV. Discussion

A. *SAX vs nSDL*

SAX excels on smooth and periodic sequences such as Hotel ADR. However, it struggles on FAANG because Gaussian breakpoints are sensitive to trend shifts and heavy-tailed noise.

nSDL, instead, captures local directionality and is naturally suited for financial-like series, but with many symbols it becomes extremely sensitive to noise and loses interpretability.

B. *Effect of Segments and Alphabet Size*

The experiments reveal consistent behaviours:

- **Increasing segments increases resolution but amplifies noise.** This is especially visible in the SAX(60,10) and SAX(80,10) figures, where nearly every small fluctuation becomes a new symbol.
- **Large alphabets cause symbolic overfitting.** This is particularly evident in nSDL with 20 symbols: the representation explodes into `up4`, `up5`, `dw7`, `dw9`, etc., masking the true domain differences.
- **Peaks and extremes are often lost.** Both SAX and nSDL reduce each segment to a single statistic (mean or slope), so very narrow anomalies disappear.
- **Moderate configurations perform best.** Both methods achieve their best interpretability at segment$=$ 20 and alphabet$=$ 4–6.

V. Conclusion

Symbolic encodings provide a complementary perspective to numeric representations. SAX effectively captures global seasonal behaviour and is ideal for Hotel ADR, while nSDL captures directional information well suited for financial-like FAANG series. However, both methods degrade when segments or alphabet sizes are too large, leading to noise amplification and symbolic overfitting.

The real SAX+KMeans experiment confirms that symbolic clustering is more fragile than numeric Fourier or PAA.

Overall, symbolic methods are useful interpretability tools but require careful tuning.

In practical applications, we recommend:

- using a numeric baseline (Fourier or PAA) for robust clustering,
- applying SAX for seasonal data,
- using nSDL for directional / trend–dominated time series.