Segmented Shape-Symbolic Time Series Representation

Herbert Teun Kruitbosch¹, Ioannis Giotis² and Michael Biehl¹

¹- University of Groningen - Johann Bernoulli Institute
Nijenborgh 9, 9747 AG Groningen - The Netherlands
²- Target Holding B.V., Nettelbosje 1, 9747 AJ Groningen - The Netherlands

Abstract. This paper introduces a symbolic time series representation using
monotonic sub-sequences and bottom up segmentation. The representation
minimizes the square error between the segments and their monotonic approximations.
The representation can robustly classify the direction of a segment and is scale in-
viant with respect to the time and value dimensions. This paper describes two
experiments. The first shows how accurately the monotonic functions are able to
discriminate between different segments. The second tests how well the segmenta-
tion technique recognizes segments and classifies them with correct symbols. Fi-
nally this paper illustrates the new representation on real-world data.

1 Introduction

A time series is a sequence of measurements taken at successive moments in time and
often contain a large number, \( N \), of measurements, \( T = t_1, \ldots, t_N \), with \( t_i \in \mathbb{R} \), possibly
too many to store in memory or do calculations. Two important reduced time series
representations are PLR [1] and SAX [2]. PLR segments a time series into pieces
of possibly different sizes and approximates them with linear curves. Other higher
order polynomial representations are SwiftSeg [3] and the shape space representation
[4]. These representations have been used by various researchers to support clustering
[1, 2, 5], classification [2, 5], association rule mining [2, 6], query by content [2] and
anomaly detection [1, 7, 8] in time series data. SAX splits a time series into segments
of an equal, predefined size and assigns a symbol to each piece, depending on the range
in which the mean value of the segment lies. The SAX representation is not scale-
invariant with respect to time, does not store information about the shape of a segment
and requires a parameter to state the size of an individual segment. The value of this
parameter can prevent us from detecting small or big patterns.

This paper introduces a piece-wise representation which tries to find monotonic
pieces, which may differ in length. Each piece is represented by a symbol, indicating
direction and curvature. Hence our representation is symbolic like SAX, but the sym-

2 Background and related work

Our representation uses least square estimations (LSE) to approximate and classify the
direction and shape of a segment. This section will discuss linear LSE and a relevant
Bayesian consideration regarding classification. Next we discuss bottom up segmenta-
tion. Finally, we explain the GAP statistic [10] to determine the number of segments.
Linear least squares estimation  In order to approximate and assign a symbol to a segment, we use LSE. We assume a segment $s = s_1, \ldots, s_N$, such that $s_n = \sum_{m=1}^M \theta_m f_m(n) + w_n$, where $f_1, \ldots, f_M$ is a set of basis functions and $w_n \sim \mathcal{N}(0, \sigma^2)$ is white Gaussian noise, independent and identically distributed (i.i.d.). A LSE determines the unknown parameters $\theta_1, \ldots, \theta_M$, such that the error between $s_n$ and $\sum_{m=1}^M \theta_m f_m(n)$ is minimal:

$$\hat{\theta}_{LSE} = \arg\min_{\theta} \sum_{n=1}^N (s_n - \sum_{m=1}^M \theta_m f_m(n))^2.$$  

We only consider linear estimations of the form $\alpha + \theta f(x)$ (based on $M = 1$ monotonically increasing basis function $f = f_1$ and a constant), which is monotonic for any $\alpha$ and $\theta$. We set $f$ to have zero mean: $\sum_{n=1}^N f(x_n) = \langle f(x), 1 \rangle = 0$, such that the basis $f, 1$ is orthogonal, allowing for LSE in linear time.

Bayesian classification  In order to classify a segment $s$, i.e. $cl: Segment \rightarrow f_1, \ldots, f_M$, we need to find the model with the highest posterior probability $P(f|s) = P(s|\alpha + \theta f(x))P(\alpha + \theta f(x))/P(s)$. Hence we could classify by maximizing the posterior probability: $cl(s) = \arg\max_{\alpha + \theta f} P(s|\alpha + \theta f(x))P(\alpha + \theta f(x))$. This expression classifies, whereas (1) fits, hence the meaning of $i$ in $f_i$ is different. The prior probability $P(\alpha + \theta f(x))$ is hard to determine, since $P(\alpha, \theta, f_i)$ and $P(f_i)$ are often unknown. Therefore, our classification model is simplified to maximum likelihood classification:

$$cl(s) = \arg\max_{\alpha + \theta f} P(s|\alpha + \theta f(x)) = \arg\max_{\alpha + \theta f} -\sum_{n=1}^N (\alpha + \theta f_i(x_n) - s_n)^2.$$  

Bottom up Segmentation  Bottom up segmentation [1] is one of many segmentation techniques, like sliding window and top down segmentation. None of them consider all possible segmentations and hence have to deal with avoiding local optima. In general bottom up outperforms other methods [11] and can use of the GAP statistic to determine the number of segments. The bottom up approach starts with a fine grained segmentation, with small equal sized segments. Based on the introduced error, adjacent segments are merged greedily and iteratively until some condition is met (Figure 2).

A local optimum occurs when merging 3 adjacent segments is beneficial, but merging any two of them is expensive. Assuming that merging two same-direction monotonic segments is never expensive and that merging two different-direction monotonic segments always is, such a local optimum will not occur.

GAP statistic  We use the GAP statistic [10] to determine the optimal number of segments. However, we use a different error function than the original GAP statistic, since we are segmenting instead of clustering:

$$\epsilon_{segment}(s) = \|s - \hat{f}(s)\|^2, \quad \epsilon_{segmentation}(S) = \sum_{s \in S} \epsilon_{segment}(s).$$

**Legend of (e):** $-\epsilon_{segmentation}$ of $t$ $-\epsilon_{segmentation}$ of $u$ $\cdots$ gap$(k|t)$

Fig 1. Bottom up segmentation of a 4-piece piece-wise linear time series $t$ into $k$ segments. The right-most figure shows the GAP statistic as discussed in section 2.
where \( fit \) approximates a segment \( s \) and \( S \) is the set of segments. The GAP statistic is based on the error of a segmentation of a time series, and the expected error of a uniform distribution in the same range. Therefore we define a reference signal \( u = u_1, \ldots, u_N \) where \( u_i \sim \mathcal{U}(0,1) \): uniformly randomly distributed between 0 and 1. We also define a time series \( t = t_1, \ldots, t_N \) with values normalized between 0 and 1. The optimal number of segments is determined from:

\[
\text{gap}(k|t) = \log \left( e_{\text{segmentation}}(S_{t,k}) \right) - \log \left( e_{\text{segmentation}}(S_{u,k}) \right),
\]

where \( S_{t,k} \) and \( S_{u,k} \) respectively are segmentations of \( t \) and \( u \) of \( k \) segments. We estimated \( e_{\text{segmentation}}(S_{U,k}) \) by taking the mean of \( Z = 10 \) instances of \( u, \hat{u} = \hat{u}_1, \ldots, \hat{u}_Z \), and corrected it with the standard deviation.

The elbow of \( \text{gap}(k|t) \) (Fig. 2e) is the point where adding more segments will only result in fitting to noise. This elbow may not be located at \( \text{argmin}_k \text{gap}(k|y) \), due to fitting to noise or numerical instability. We selected the \( k_{\text{elbow}} \) such that the second order derivative of \( \text{gap}(k|t) - \text{gap}(k+1|t) \) is maximal.

## 3 Method

Our representation is created by first segmenting a time series using an approximation as described in algorithm 1. The GAP statistic determines the amount of segments and each final segment is classified according to algorithm 1. The result is a symbolic string which can then be used by various machine learning algorithms.

**Assigning Shapes to Segments** To detect non-monotonic structures in time series, this paper focuses on finding monotonic segments. This way a concatenation of monotonic segments defines a more general non-monotonic structure. We classify a segment as one of four shapes, shown in figure 2. Except for the first, each shape is monotonically increasing. However, we also classify whether the fit of such a shape goes up or down. This results in a classification tree for shapes:

![Classification Tree](image)

**Detecting flat segments** In order to classify the flat shape of figure 2, we could fit a segment \( f_0(x) = 0 \). However, for \( \theta_0 = 0 \), each of \( f_1, f_2, f_3 \) will have the same likelihood as \( f_0 \). This disallows us to distinguish between \( f_0 \) and any of \( f_1, f_2, f_3 \). We could solve this by adding a prior, but like stated in section 2 this can not always be determined.

### Algorithm 1 Classification and approximation of a segmented time series

**Require**: Segment \( S = (s_1, \ldots, s_N) \) with mean 0 and std dev 1, Threshold \( T \) (Sec 3)

**Ensure**: Classification \( C \in \{f_0, f_1, f_2, f_3\} \times \{\text{none, up, down}\} \) and approximation of \( S \).

Set \( x_1, \ldots, x_N \) equidistant on \([0,1]\)

for \( i = 1, 2, 3 \) set \( \theta_i = \frac{1}{N} \sum_{n=1}^N f_i(x_n) \cdot s_n / \sum_{n=1}^N f_i^2(x_n) \) \( \triangleright \) Linear LSE (for \( f_i \) see Fig. 2)

\( i = \arg\max \theta_i = \sum_{n=1}^N \left( \theta_i f_i(x_n) - s_n \right)^2 \) \( \triangleright \) Likelihood based classification

if \( |\theta_i| < T \) then \( i = 0; \text{dir} = \text{none}; \theta_0 = 0; f_0(x) = 0 \) \( \triangleright \) Flat segment

else if \( \theta_0 < 0 \) then \( \text{dir} = \text{down} \) else \( \text{dir} = \text{up} \) \( \triangleright \) Down / Up

\( C = (f_i, \text{dir}) \) \( \triangleright \) Symbol of the segment

Set \( A = (a_1, \ldots, a_N) \) such that \( a_n = \theta_i f_i(x_n) \) \( \triangleright \) Approximation of the segment
Therefore, we introduce a heuristic based on the signal to noise ratio to classify flat segments. We consider a noisy, flat segment $\hat{s} = x_1, \ldots, x_N$, $s_n = \pi_n + \sigma_n (\pi_n \sim \mathcal{N}(0, \sigma^2))$, shown in figure 3a, for which we want to determine whether it is flat or has some non-flat $f$ shape. We consider a simple case where $f$ has the first half of the points equal to $-1$ and the second half equal to $1$, hence $\text{Var}(f) = 1$. Then we normalize $\hat{s}$ to have variance $1$. Due to $f$’s shape, estimation of $\theta$ will result in the average of the right half’s average and the left half’s negative average. Therefore the distribution of the LSE $|\theta|$ for $f$ depends on the distribution of the average value of $\pi$: $P(|\theta|) = \mathcal{N}(1/|\theta|; 0, 1/N)$.

Fig. 3. In (a) the fit of $f$ on the data of a flat model $s_n = \pi_n$ (gray) and a non-flat model $\hat{s}_n = \theta f(x_n) + \sigma_n (n = \text{black})$ are shown. See Sec 3. In (c) the probability distribution of $\theta$ for both models for $\theta = 1, \sigma = 1$ and $N = 5$ is shown.

Legend of (a): $\bigcirc / \bigcdot$ data, $-f / \cdots$ fit of $f$, $\cdots y = 0$

Legend of (b): $\bigcirc$ flat shape, $\bigcdot$ non-flat

Now consider a noisy, non-flat segment $\hat{s}_n = \theta f(x_n) + \sigma_n (\hat{\pi}_n \sim \mathcal{N}(0, \sigma^2))$ (Fig. 3a). We have that $\text{Var}(\hat{s}_n) = \theta^2 + \sigma^2$. Hence if we normalize $\hat{s}$ to variance $1$, we get $\hat{\pi}_n \sim \mathcal{N}(0, \sigma^2/\sqrt{\theta^2 + \sigma^2})$, and the probability of a LSE $|\theta|$ for $f$ is:

$$P(|\theta|) \approx \mathcal{N}(1/|\theta|; \sqrt{\theta^2 + \sigma^2}, 2/(N \sqrt{\theta^2 + \sigma^2})).$$

Notice that $\sqrt{\theta^2 + \sigma^2}$ is a normalized $\theta$. The variance of $\theta$ for the flat shape and the $f$ shape depends on the number of samples by a factor $1/N$ and on the signal to noise ratio $\sigma^2/\sqrt{\theta^2 + \sigma^2}$ in case of a non-flat signal. Hence more noise moves the means closer to each other; decreasing the discriminative power of $|\theta|$. Figure 3b shows an example of both distributions and suggests a threshold $|\theta| = 0.26$, such that $|\theta| < 0.26$ would classify as flat and $|\theta| \geq 0.26$ as not flat. This threshold $T$ is used in 1. We assumed a block-shaped $f$. Assuming other shapes of $f$ makes it harder to determine the distribution of $|\theta|$, because the variance $\text{Var}(f(x) + \sigma_i)$ is not necessarily $\text{Var}(f(x)) + \text{Var}(\sigma_i)$, the idea is however similar.

Summary In contrast with PLR, the introduced method is symbolic and allows for processing using discrete algorithms and data structures. In contrast to SAX, the proposed representation uses segmentation and can identify two similar segments with a different number of samples as the same, allowing for scale invariance. Whereas SAX symbols are based on mean values, the introduced symbols are based on shapes. In summary, the introduced method is based on a symbolic representation like SAX but retains the scale invariance of the shape based PLR.
4 Results

The accuracy of our symbolic representation is quantified using two measures. These two measures are used to first determine how often the classification of a synthetic segment matches the label in section 4; this tests the accuracy of algorithm 1 and illustrates how well the classifier is able to discriminate between shapes. Section 4 illustrates the accuracy of our representation including segmentation. Finally an example of real world data from the IJkdijk project is represented using our method.

The two accuracy measures are based on the symbol tree of section 3. The first measure quantifies the fraction of segments or samples for which the direction of the shape is correctly classified: the direction accuracy, the second quantifies the fraction for which both the direction and the shape are correct: the shape accuracy.

Classification accuracy To measure the discriminating ability of our classification described by algorithm 1, we created 100 synthetic shapes with variance 1 for each direction of the 4 template functions (Figure 2). Then we tested the accuracy as the fraction of correctly classified segments, with respect to the amount of i.i.d. white Gaussian noise, the number of samples in a segment and both measures. The results are shown in figure 4. The direction accuracy stays above 80% up to $\sigma = 0.7$ and the direction accuracy up to $\sigma = 0.4$. Note that the curvature gets lost with more noise, but the direction does not. For $\sigma \geq 1$ both accuracies drop as all segments are classified as flat.

Segmentation and classification accuracy To further evaluate the accuracy and robustness of the proposed technique we measured the fraction of correctly classified samples of a segmented time series. We created 100 time series of 1001 samples by concatenating 4 monotonic segments of random size with at least 50 samples based on the shapes in figure 2. The time series have variance 1. We label each sample as the shape it was created from. The segments are concatenated such that there are no gaps in the y direction. Finally the synthetic time series is segmented and each sample is classified by classifying the encapsulating segment using algorithm 1. The fraction of matching classified and labelled samples is shown in figure 5. The direction accuracy is robust with respect to noise; it remains above 0.8 for up to noise $\sigma = 1$, whereas the shape accuracy scores lower, showing that shape information is lost when noise is added.

IJkdijk data example Finally, some subsequences from the LiveDijk Eemshaven[9] measurements were taken and translated into symbolic strings. The results are shown in figure 6. The plots show that indeed monotonic structures can be found, the first and third figure have smaller patterns and the middle figure has larger ones. This illustrates the scale invariance of our representation. This behaviour, i.e. finding small or

Fig. 4: Accuracy of identifying the segment Fig. 5: Segmentation and segment classification of one synthetic shape for respectively 501 cation accuracy as described in section 4. and 1001 samples. The MC classification The gray area is half of the standard deviation accuracy is of a classifier which randomly assigns a direction and shape.

Accuracy legend: — direction, — shape, - - MC direction, - - MC shape

263
large patterns is influenced by the estimation of the number of segments. Changing the trade-off between the amount of segments and the amount of error will change how the representation deals with smaller, less defined monotonic segments.

5 Conclusion and future work

We introduced a scale-invariant, symbolic representation of a time series, based on monotonic segments of possibly different sizes, which captures direction and curvature. The representation enables machine learning algorithms to efficiently find anomalies or structures in a time series based on the shape of monotonic sub-sequences. However, selecting the number of segments is still an ill-defined problem. Future work could define a hybrid representation, representing both smaller and larger, encapsulating segments, for example by finding a tree instead of a sequence of segments. Clustering patterns generated by our representation on the IJkdijk data may give insight in the general behaviour and hence help define and determine anomalous sub-sequences.

References


