Bachelor Thesis - Exposé

Distance Measure Selection in ClaSP

Jasper Lennart Köhn
Department of Computer Science, Humboldt-Universität zu Berlin
Supervisor: Prof. Dr. Ulf Leser, Arik Ermshaus

1 Introduction

In many areas of everyday life, data is measured over time, resulting in vast collections of *time series* - sequences of temporally ordered data. Examples of this are weather and environmental measurements, such as temperature, humidity and wind speed [15], brain activity or heart rate measurements from EEG and ECG in healthcare [6][21], stock prices, network traffic and seismic activity data [18][20][29]. Time series can also be obtained by measuring static objects and ordering the data afterwards, for example by measuring the width of leaves along the midvein [3]. The abundance of time series data across a multitude of domains gives rise to many problems, including *time series segmentation*. When measuring a process, we are often interested in state transitions, e.g. the moment when a patient's heart rate becomes irregular, or distinguishing between a person walking and running in motion sensor data. Such moments of state transition are called *change points*. Time series segmentation is the problem of identifying all the change points within a time series.

Classification Score Profile (ClaSP) is a self-supervised method for segmenting time series [13]. It utilizes subsequences of time series and the *k-Nearest Neighbor Classifier (k-NN)*, a widely used tool for time series classification tasks. When predicting the class of a subsequence, the k-NN in ClaSP calculates the distance of that subsequence to every other subsequence and assigns a label/class based on the k most similar subsequences. The similarity of subsequences is determined through the use of a distance measure such as the Euclidean distance. ClaSP places a hypothetical change point, a split, at every position in a time series and uses a k-NN to determine whether the subsequences left of the split belong to the same class as the subsequences right of the split. Cross-validation is used for each split and the cross-validation scores constitute the profile. Finally, change points are extracted from the profile.

The measurement of distances is a fundamental aspect of both the k-NN and ClaSP. This distance measure is expected to behave in a certain way. When calculating the distance between two similar time series, or two similar subsequences of a time series, the distance measure should produce a small positive

real value and a larger real value when they are dissimilar.

At the time of writing, ClaSP supports three distance measure. (1) Euclidean distance [3], (2) Z-Normalized Euclidean distance [29] and (3) Complexity-Invariant distance [3]. However, it is unclear which distance measure results in the most accurate segmentation for an unseen time series. For this reason, a mechanism is needed that decides which of the distance measures should be used. A number of heuristics are employed to guide the decision. Time series are tested for stationarity and periodicity, and summary statistics are evaluated to inform the selection of a distance measure.

The goal of this bachelor thesis is to improve ClaSP by adding a distance measure selection process and by implementing another distance measure, the *Earth Mover's distance* [25]. The distance measure selection process is evaluated by comparing the covering scores of segmentations created by the version of ClaSP using the proposed method to select a distance measure, default ClaSP, that is ClaSP with the Z-Normalized Distance measure, and ClaSP with the ideal distance measure. Finally, a comparison is drawn between ClaSP and its competitors.

2 Definitions and Notation

In this section, the concepts used throughout this exposé are defined.

Definition 1 (Process)**.** A *process* $P = (Q, \Delta)$ is a tuple of a set of states Q and a set of transitions between those states $\Delta \subseteq Q \times Q$. For all $(q_i, q_j) \in \Delta$, it holds that $q_i \neq q_j$.

Definition 2 (Time Series)**.** Given a process $P = (Q, \Delta)$, a *time series* $T = (t_1, \ldots, t_N)$ is a sequence of $N \in \mathbb{N}$ real values that measures an observable output of P. Each state $q \in Q$ produces a unique observable output.

Definition 3 (Subsequence)**.** Given a time series $T = (t_1, \ldots, t_N)$, a *subsequence* $T_{s,e}$ of T is the sequence $T_{s,e} = (t_s, \ldots, t_e)$ with $1 \leq s \leq e \leq N$ and $s, e \in \mathbb{N}$.

Note. A time series that approximately repeats a subsequence after a fixed amount of time is called *periodic*.

Definition 4 (Change Point)**.** Given a process $P = (Q, \Delta)$ and a corresponding time series T of length N, a *change point* is an offset $i \in [1, 2, \ldots, N]$ in T that corresponds to a state transition $(q_k, q_l) \in \Delta$ in P.

Definition 5 (Segmentation)**.** Given a process P and a corresponding time series T, a segmentation is an ordered set of change points C.

Definition 6 (Segment)**.** Given a segmentation $SEG = \{1, c_1, c_2, \ldots, c_m, N\}$, $S = \{x \in \mathbb{N} \mid c_i \leq x < c_{i+1}\}$ is a *segment* for any $i \in \mathbb{N}$ with $0 \leq i \leq m$.

Note. $c_0 = 1$ and $c_{m+1} = N$ for notational convenience.

3 Related Work and Background

This section provides an overview over the different kinds of distance measures, it contains a detailed explanation of the distance measures used in ClaSP and explains the inner workings of ClaSP.

3.1 Distance Measures

Calculating distances between different subsequences of a time series is a central part of ClaSP. There is an abundance of different distance measures available to perform the aforemetioned task.

Time series distance measures can be divided into four groups [19]. Shape-based distances compare the shape of time series by measuring the similarity of the raw-values of the time series. Shape-based distances can either be (i) lock-step measures, which compare the i-th value of a time series to the i-th value of another e.g. the Euclidean distance [3], or (ii) elastic measures, that allow one-to-many and one-to-none value matchings e.g. Dynamic Time Warping [24]. Feature-based distances compare features that are extracted from the time series, such as Fourier or wavelet coefficients [2]. Structure-based distances can be divided into (i) model-based distances, where a model is fit to each time series and the comparison of the models results in a distance, e.g. Piccolo distance [22] and (ii) compression-based distance [16]. Finally, prediction-based distances calculate the similarity of forecasts for different time series [28].

ClaSP supports three distance measures that all belong to the category of lock-step measures. These distance measures can be implemented efficiently using the *Scalable Time series Ordered-search Matrix Profile (STOMP)* algorithm [29].

STOMP solves the *all-pairs-similarity-search* problem for time series subsequences. Given a time series T and a positive integer m, STOMP computes the *distance profile* for each subsequence of T with length m. A distance profile is a vector of distances between a specified query subsequence and each subsequence of equal length in T. The distance between subsequences can be computed with their dot product. Once a dot product between two subsequences T_{s_1,e_1} and T_{s_2,e_2} has been computed, the dot product of T_{s_1,e_1} and T_{s_2+1,e_2+1} can be computed in constant time by exploiting the overlap between neighboring subsequences. The first dot products are pre-computed using FFT [29].

3.1.1 Euclidean Distance

Let $T = (t_1, \ldots, t_N)$ be a time series of length N and $T_{s_1,e_1} = A = (a_1, \ldots, a_m)$ and $T_{s_2,e_2} = B = (b_1, \ldots, b_m)$ subsequences of T with length m. The *Euclidean distance* (ED) between A and B is defined as [3]:

$$ED(A, B) = \sqrt{\sum_{i=1}^{m} (a_i - b_i)^2} \tag{1}$$

By applying the *binomial theorem*, Eq. 1 can be expanded into three sums. The dot product of A and B, denoted as $\langle A, B \rangle$, and the cumulative sum of squares for A and B, denoted as $\#^2(A)$ and $\#^2(B)$:

$$ED(A, B) = \sqrt{-2 \sum_{i=1}^{m} a_i b_i + \sum_{i=1}^{m} a_i^2 + \sum_{i=1}^{m} b_i^2} \tag{2}$$

Using the notation:

$$ED(A, B) = \sqrt{-2 \langle A, B \rangle + \#^2(A) + \#^2(B)} \tag{3}$$

As demonstrated by [29], $\langle A, B \rangle$ can be computed in constant time from an existing dot product using a sliding window approach. The same is true for $\#^2(A)$ and an existing cumulative sum of squares $\#^2(\hat{A})$ for the previous subsequence $\hat{A}$. $\#^2(A)$ can be computed from $\#^2(\hat{A})$ with the subtraction of the first squared value in $\hat{A}$ and the addition of the last squared value in A.

$$\#^2(A) = \#^2(\hat{A}) - \hat{a}_1^2 + a_m^2 \tag{4}$$

3.1.2 *Z*-Normalized Euclidean Distance

In order to calculate the *Z-Normalized Euclidean distance* (ZED), it is first necessary to normalize the subsequences by calculating Z-scores. The mean μ and the standard deviation σ of the respective subsequences are required. In a similar manner to the sliding cumulative sum of squares and the sliding dot product, both the mean and the standard deviation of the subsequences in T can be calculated by employing a sliding window approach [23].

ZED can be calculated using the following formula [29]:

$$ZED(A, B) = \sqrt{2m \left(1 - \frac{\langle A, B \rangle - m\mu_A\mu_B}{m\sigma_A\sigma_B}\right)} \tag{5}$$

ZED is the default distance measure for ClaSP. The Z-normalization ensures that all subsequences are compared with a mean of 0 and a standard deviation of 1 [8]. Consequently, ZED between subsequences is less prone to noise and more dependent on the general shape of each subsequence than on the raw values of the subsequences themselves.

3.1.3 Complexity-Invariant Distance

The *Complexity-Invariant distance* (CID) uses information about the complexity of a time series as a correction factor (CF) for the used distance measure.

In ClaSP the correction is applied to the Euclidean distance but it can be used with any other distance measure [3].

$$CID(A, B) = ED(A, B) \cdot CF(A, B) \tag{6}$$

CF is defined in terms of a complexity estimate (CE):

$$CF(A, B) = \frac{\max(CE(A), CE(B))}{\min(CE(A), CE(B))}. \tag{7}$$

There are many ways to quantify the complexity of a time series. Batista et al. suggest the following [3]:

$$CE(A) = \sqrt{\sum_{i=1}^{n-1}(a_i - a_{i+1})^2} \tag{8}$$

This definition of CE is based on the intuition that if we could stretch a time series until it became a straight line, a complex time series would result in a longer line than a simple time series.

3.1.4 Earth Mover's Distance

The *Earth Mover's distance* (EMD) was originally designed to measure the distance between two probability distributions [25]. Intuitively, one distribution can be seen as a mass of earth and the other as a collection of holes in the same space. A unit of work corresponds to transporting a unit of earth by a unit of distance. The EMD measures the least amount of work needed to fill the holes with earth, hence the name.

Computing the EMD is based on a solution of the transportation problem [7]. This is a bipartite network flow problem and it can be formalized by the following linear program: Let $\mathcal{I}$ be a set of suppliers, $\mathcal{J}$ a set of consumers and c_{ij} the cost to transport a unit from $i \in \mathcal{I}$ to $j \in \mathcal{J}$. The objective is to find a set of flows that minimizes the cost

$$\sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{J}} c_{ij} f_{ij}, \tag{9}$$

subject to the constraints:

$$f_{ij} \geq 0 \qquad i \in \mathcal{I}, j \in \mathcal{J} \tag{10}$$

$$\sum_{i \in \mathcal{I}} f_{ij} = y_j \qquad j \in \mathcal{J} \tag{11}$$

$$\sum_{j \in \mathcal{J}} f_{ij} \leq x_i \qquad i \in \mathcal{I}, \tag{12}$$

where x_i is the supply of supplier i and y_j the capacity of consumer j.

The EMD can be applied to time series subsequences with some minor adjustments. Given two subsequences $T_{s_1,e_1} = A = (a_1, \ldots, a_m)$ and $T_{s_2,e_2} = B = (b_1, \ldots, b_m)$ and their cumulative sums $\#(A)$ and $\#(B)$, let $\#(A) \geq \#(B)$, then A will be the set of suppliers and B the set of consumers. A and B can be understood as arrays of suppliers and consumers where supplier i has a supply of $|a_i|$ units of earth and consumer j has a capacity of $|b_j|$. We have to use absolute values because negative supply/capacity would reverse the roles of supplier and consumer. The cost c_{ij} to move a unit of earth from supplier i to consumer j will be

$$c_{ij} = |s_1 + i - 1 - (s_2 + j - 1)| \tag{13}$$

or simply the distance between supplier and consumer on the time axis. The $EMD(A, B)$ will be the minimum cost to transform A into B.

3.2 ClaSP

Classification Score Profile (ClaSP) [13] is a self-supervised method for time series segmentation. For an unseen time series, ClaSP creates a classification score profile. The profile is obtained by inserting hypothetical change points into the time series at every possible position. The resulting splits are then viewed as binary classification problems where the subsequences to the left of a hypothetical change point are classified as 0 and the subsequences to the right are classified as 1. Finally, cross-validation is performed and the cross-validation scores constitute the profile. The magnitude of the score indicates the dissimilarity between the sequences on either side of the hypothetical change point. Therefore, the local maxima of the profile are candidates for change points. The candidates are extracted recursively from the profile and subjected to a validation process.

4 Method

Given that ClaSP is unsupervised and operates on a single time series, the distance measure selection process must be designed to align with these operational constraints.

We plan to base the selection of the distance measure on properties of the time series. If the time series is stationary, meaning that it does not change its statistical properties over time, comparing the raw values of subsequences may prove more effective in identifying discriminatory features than transforming the data in some way. Therefore, ED could be an effective distance measure. Similarly, if the time series is non-stationary ZED or CID may prove more effective. When periodicity is detected in a time series, ZED may be preferred over ED and CID because the general shape of a subsequence is more relevant when distinguishing between the different states of the underlying process than its raw values. In cases when many subsequences show a high mean or variance ZED may be used to avoid exagerated differences between subsequences caused by the magnitude of the values. On the other hand if some subsequences show a low variance and some show a high variance, we consider using CID as it amplifies the differences between subsequences with differing complexity estimates and has only a negligible effect if the subsequences have similar complexity estimates (see Eq. 7).

A number of tests can be conducted to gain insight into the properties of the time series. The remainder of this section provides an explanation of these test.

The time series under consideration is tested for *stationarity*. Statistical tests exist to determine whether a time series is likely stationary or not, e.g. the *Augmented Dickey-Fuller (ADF) Test* [10][11] and the *Kwiatkowski-Phillips-Schmidt-Shin (KPSS) Test* [17].

The time series is also analyzed for periodicity. Detecting periodicity can be approached through the use of the *autocorrelation function (ACF)*, which quantifies the correlation between the time series and its lagged version [4]. Peaks in the ACF at regular intervals indicate a periodic pattern.

In order to further characterize the time series, summary statistics of its subsequences are computed using the aforementioned sliding window approach [23].

5 Evaluation

The *Time Series Segmentation Benchmark (TSSB)* [12][27] is used to evaluate the proposed heuristics. The TSSB consists of 75 annotated time series with 1 to 9 segments. Each time series is created by grouping time series from the UEA and UCR time series classification datasets by their labels and concatenating them. The offsets at which the time series were concatenated are annotated as change points.

We employ a covering metric Cov to evaluate the segmentation [5].

Let $T = (t_1, \ldots, t_N)$ be a time series of length N, $Y = \{1, y_1, \ldots, y_k, N\}$ be the set of true change points and $P = \{1, p_1, \ldots, p_l, N\}$ be the set of predicted change points for $k, l \in \mathbb{N}$. Note that $\forall i : y_i < y_{i+1}$ and $p_i < p_{i+1}$.

Then S_Y (S_P) is the set of segments, seperated by the true (predicted) change points:

$$S_Y = \{\{y_i, y_i + 1, \ldots, y_{i+1} - 1\} \mid i \in \{0, 1, ..., k\}\} \tag{14}$$
$$S_P = \{\{p_i, p_i + 1, \ldots, p_{i+1} - 1\} \mid i \in \{0, 1, ..., l\}\} \tag{15}$$

The covering score Cov is then defined as:

$$Cov(S_Y, S_P) = \frac{1}{N} \sum_{s \in S_Y} |s| \cdot \max_{s' \in S_P} \frac{|s \cap s'|}{|s \cup s'|} \tag{16}$$

We compare the covering scores of segementations made by the version of ClaSP, that utilizes the proposed distance measure selection process, default ClaSP, as well as ClaSP employing the optimal distance measure. ClaSP is then compared to other state-of-the-art time series segmentation algorithms, including FLOSS [14], ESPRESSO [9], BOCD [1] and BinSeg [26].

Futhermore, we compare the covering scores of ClaSP with EMD to those of ClaSP with the other distance measures. The objective is to determine whether the segmentation quality can be significantly improved by compromising computational speed in favour of a more complex distance measure.

References

[1] Ryan Prescott Adams and David J. C. MacKay. *Bayesian Online Changepoint Detection*. 2007. DOI: 10.48550/arXiv.0710.3742.

[2] R. Agrawal, C. Faloutsos, and A Swami. "Efficient similarity search in sequence databases". In: *4th International Conference of Foundations of Data Organization and Algorithms*. 1993, pp. 69–84. DOI: 10.1007/3-540-57301-1_5.

[3] Gustavo E. A. P. A. Batista et al. "CID: an efficient complexity-invariant distance for time series". In: *Data Mining and Knowledge Discovery* 28.3 (May 2014), pp. 634–669. ISSN: 1573-756X. DOI: 10.1007/s10618-013-0312-3.

[4] George E. P. Box et al. *Time Series Analysis - Forecasting and Control*. Vol. 5. Wiley Series in Probability and Statistics, 2015, pp. 21–44.

[5] Gerrit J.J. van den Burg and Christopher K.I. Williams. *An Evaluation of Change Point Detection Algorithms*. Feb. 2022. DOI: 10.48550/arXiv.2003.06222.

[6] Wanpracha Art Chaovalitwongse, Ya-Ju Fan, and Rajesh C. Sachdeo. "On the Time Series K-Nearest Neighbor Classification of Abnormal Brain Activity". In: *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* 37.6 (2007), pp. 1005–1016. DOI: 10.1109/TSMCA.2007.897589.

[7] G. B. Dantzig. "Application of the Simplex Method to a Transportation Problem". In: *Activity Analysis of Production and Allocation*. 1951, pp. 359–373.

[8] Dieter De Paepe, Diego Nieves Avendano, and Sofie Van Hoecke. "Implications of Z-Normalization in the Matrix Profile". In: *Pattern Recognition Applications and Methods*. Ed. by Maria De Marsico, Gabriella Sanniti di Baja, and Ana Fred. Cham: Springer International Publishing, Jan. 2020, pp. 95–118. ISBN: 978-3-030-40014-9. DOI: 10.1007/978-3-030-40014-9_5.

[9] Shohreh Deldari et al. "ESPRESSO: Entropy and ShaPe awaRe timE-Series SegmentatiOn for Processing Heterogeneous Sensor Data". In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4.3 (Sept. 2020), pp. 1–24. ISSN: 2474-9567. DOI: 10.1145/3411832.

[10] David A. Dickey and Wayne A. Fuller. "Distribution of the Estimators for Autoregressive Time Series With a Unit Root". In: *Journal of the American Statistical Association* 74.366 (1979), pp. 427–431. ISSN: 01621459, 1537274X. DOI: 10.2307/2286348.

[11] David A. Dickey and Said E. Said. "Testing for unit roots in autoregressive-moving average models of unknown order". In: *Biometrika* 71.3 (Dec. 1984), pp. 599–607. ISSN: 0006-3444. DOI: 10.1093/biomet/71.3.599.

[12] Arik Ermshaus, Patrick Schäfer, and Ulf Leser. "ClaSP - Time Series Segmentation". In: *CIKM*. 2021.

[13] Arik Ermshaus, Patrick Schäfer, and Ulf Leser. "ClaSP: parameter-free time series segmentation". In: *Data Mining and Knowledge Discovery* 37.3 (May 2023), pp. 1262–1300. ISSN: 1573-756X. DOI: 10.1007/s10618-023-00923-x.

[14] Shaghayegh Gharghabi et al. "Domain agnostic online semantic segmentation for multi-dimensional time series". In: *Data Mining and Knowledge Discovery* 33.1 (Jan. 2019), pp. 96–130. ISSN: 1573-756X. DOI: 10.1007/s10618-018-0589-3.

[15] Zahra Karevan and Johan A.K. Suykens. "Transductive LSTM for time-series prediction: An application to weather forecasting". In: *Neural Networks* 125 (May 2020), pp. 1–9. ISSN: 0893-6080. DOI: 10.1016/j.neunet.2019.12.030.

[16] Eamonn Keogh, Stefano Lonardi, and Chotirat Ratanamahatana. "Towards parameter-free data mining". In: *KDD-2004 - Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Aug. 2004, pp. 206–215. DOI: 10.1145/1014052.1014077.

[17] Denis Kwiatkowski et al. "Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root?" In: *Journal of Econometrics* 54.1 (1992), pp. 159–178. ISSN: 0304-4076. DOI: 10.1016/0304-4076(92)90104-Y.

[18] Sidra Mehtab and Jaydip Sen. "A time series analysis-based stock price prediction using machine learning and deep learning models". In: *International Journal of Business Forecasting and Marketing Intelligence* 6.4 (2020), pp. 272–335. DOI: 10.1504/IJBFMI.2020.115691.

[19] Usue Mori, Alexander Mendiburu, and Jose A. Lozano. "Distance Measures for Time Series in R: The TSdist Package". In: *The R Journal* (2016).

[20] Mbulelo Brenwen Ntlangu and Alireza Baghai-Wadji. "Modelling Network Traffic Using Time Series Analysis: A Review". In: *Proceedings of the International Conference on Big Data and Internet of Thing*. BDIOT '17. London, United Kingdom: Association for Computing Machinery, 2017, pp. 209–215. ISBN: 9781450354301. DOI: 10.1145/3175684.3175725.

[21] Robert Thomas Olszewski, Roy Maxion, and Dan Siewiorek. "Generalized feature extraction for structural pattern recognition in time-series data". PhD thesis. USA: Carnegie Mellon University, 2001. ISBN: 0493538712.

[22] Domenico Piccolo. "A Distance Measure for classifying ARIMA Models". In: *Journal of Time Series Analysis* 11.2 (1990), pp. 153–164. DOI: https://doi.org/10.1111/j.1467-9892.1990.tb00048.x.

[23] Thanawin Rakthanmanon et al. "Addressing Big Data Time Series: Mining Trillions of Time Series Subsequences Under Dynamic Time Warping". In: *ACM Trans. Knowl. Discov. Data* 7.3 (Sept. 2013). ISSN: 1556-4681. DOI: 10.1145/2500489.

[24] Chotirat Ann Ratanamahatana and Eamonn Keogh. "Everything you know about dynamic time warping is wrong". In: *Third workshop on mining temporal and sequential data*. Vol. 32. Citeseer. 2004.

[25] Y. Rubner, C. Tomasi, and L.J. Guibas. "A metric for distributions with applications to image databases". In: *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)*. 1998, pp. 59–66. DOI: 10.1109/ICCV.1998.710701.

[26] A. J. Scott and M. Knott. "A Cluster Analysis Method for Grouping Means in the Analysis of Variance". In: *Biometrics*. Vol. 30. Sept. 1974, pp. 507–512. DOI: 10.2307/2529204.

[27] *Time Series Segmentation Benchmark*. https://github.com/ermshaua/time-series-segmentation-benchmark.

[28] J.A. Vilar, A.M. Alonso, and J.M. Vilar. "Non-linear time series clustering based on non-parametric forecast densities". In: *Computational Statistics and Data Analysis* 54.11 (2010). The Fifth Special Issue on Computational Econometrics, pp. 2850–2865. ISSN: 0167-9473. DOI: https://doi.org/10.1016/j.csda.2009.02.015.

[29] Yan Zhu et al. "Matrix Profile II: Exploiting a Novel Algorithm and GPUs to Break the One Hundred Million Barrier for Time Series Motifs and Joins". In: *2016 IEEE 16th International Conference on Data Mining (ICDM)*. 2016, pp. 739–748. DOI: 10.1109/ICDM.2016.0085.