Change Point Detection via Synthetic Signals

Ting-Ji Huang[1], Qi-Le Zhou[1], Han-Jia Ye[1]*, and De-Chuan Zhan[1]*

National Key Laboratory for Novel Software Technology, Nanjing University, China
{huangtj,zhouql,yehj,zhandc}@lamda.nju.edu.cn

Abstract. Detecting change points in time series data is a widely acknowledged challenge with diverse applications, in which the data obtained from measured values is often characterized by complex compositions, and the availability of real data is typically limited. However, current detection algorithms often depend on domain-specific data to achieve better performance or are restricted to analyzing single variant series, limiting their applicability. In this paper, we introduce a novel approach to change point detection that eliminates the requirement for collecting supervised data. Initially, we train a discriminant model using artificially generated synthetic signals comprising a combination of intricate patterns and random noise. This discriminant model is designed to predict the number of change points, and the synthetic data set encompasses a wide range of patterns observed in real data and offers significant advantages in terms of diversity and data volume. The trained discriminant model is then applied in conjunction with the ClaSP method for change point detection. To fully exploit multivariate series information, we propose a simple yet useful weighted-merging method that improves detection performance by aggregating change point votes within each time gap. Experimental results demonstrate the superiority of our Detection Model via Synthetic Signals (DMSS) compared to the original ClaSP method, demonstrating exceptional performance on the Human Activity Segmentation dataset.

Keywords: Change Point Detection · Synthetic Signals · Multivariate Series.

1 Introduction

The exploration of time series data plays a crucial role in comprehending and predicting the intricate dynamics of real-world systems. However, the temporal nature of such data also introduces the possibility of abrupt changes or shifts in behavior, known as change points. Detecting change points in time-series data is a multifaceted and challenging problem [1]. Unlike traditional anomaly detection, which focuses on identifying outliers or deviations from a predefined norm, change-point detection aims to identify specific moments in time when the statistical properties of the data undergo a fundamental shift. These shifts can manifest as sudden spikes, dips, or changes in trend, indicating a transformation in the underlying data-generating process.

Over the following decades, numerous change point detection methods have been developed [2–8]. These methods are based on diverse concepts and have the ability to recognize various types of changes in time series, such as jumps in mean and variance, correlations among different components, and more complex dependencies. Comprehensive overviews describing these algorithms can be found in various literature sources [1, 9, 10]. To indicate the reliability of the predicted change points, change point detection methods apply various techniques to extract relevant characteristics from each segment, such as Arc Curve [4], CUSUM statistic [3], Gaussian statistics [6], and information gain [5]. ClaSP [8] trains a binary classifier for each possible split point and utilize the accuracy to generate characteristics. These characteristics capture the properties of the segment that are indicative of different semantic classes.

However, most of the existing change point detection methods implicitly assume that all data is segmentable and the specific number of segments is usually automatically identified by heuristic algorithms, making it difficult to obtain reliable predictions. On the other hand, when there are multiple variables in a time series instead of single time series, there is lack of efficient methods to combine multiple predictions. To alleviate these two issues, we propose a novel approach called DMSS (Discriminant Model via Synthetic Signals) for detecting change points in time series. Based on the observation that real signals often consist of a certain range of recognizable patterns, our method incorporates a discriminant model trained on synthetic signals and utilizes a simple merging technique, in which the synthetic data set encompasses a wide range of patterns observed in real data and offers significant advantages in terms of diversity and data volume. We introduce a model-based method to estimate the number of change points in a series more accurately, which greatly aids subsequent segmentation tasks. We present a straightforward yet effective merging method that leverages the information from multivariate time series. Experimental results demonstrate that the proposed DMSS method outperforms the original ClaSP method on the Human Activity Segmentation data set and finally ranked the third place in HAS challenge[11].

In this paper, we will begin by describing the quality metrics and the methodology we will use to develop our algorithms. We will also present the results of our experiments and evaluate the effectiveness of our approach. Our code is available at: `https://github.com/Tingji2419/MSS`.

2 Related Works

Change Points Detection (CPD) has been extensively studied over the last several decades in the fields of data mining, statistics, and computer science, as it addresses a wide range of real-world problems. There are three main groups of approaches for time series segmentation: dynamic programming, heuristic, and probabilistic [12].

Dynamic programming is utilized as an optimization method in conjunction with a cost function [13–15]. The fundamental technique for dynamic program-

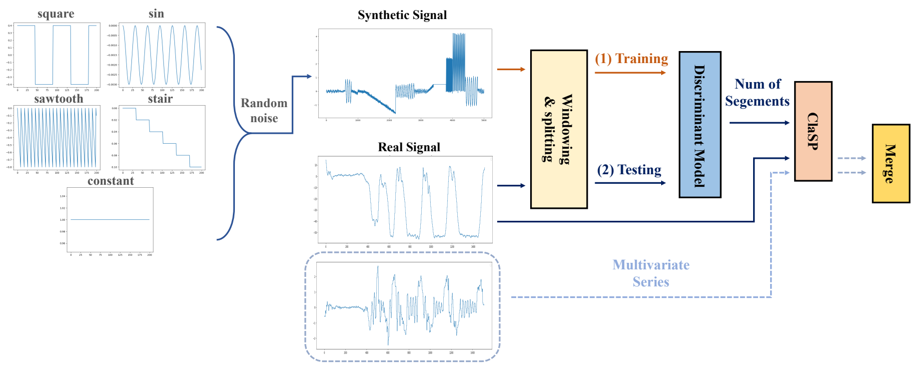

Fig. 1. The process of Detection Model via Synthetic Signals (DMSS). We first generate synthetic signals with complex patterns. The synthetic data gets multiple discrimination problems after being split by sliding windows. We train a discriminant model based on synthetic data set, which will be used to predict the number of segments of each time series in advance. After combining the discriminant model with ClaSP, we re-weight the segmentation results of each individual variable and get the final merged result, making full use of the semantic information of multiple variables.

ming segmentation is called k-segmentation, which focuses on minimizing the variance within the segments. Heuristic approaches can be categorized into three groups: sliding window, TopDown, and BottomUp. Sliding window approaches involve sliding a window over the time series and initiating a new segment when a specified error criterion is met [16]. TopDown approaches start with a single segment and recursively partition the time series until a specific error criterion is satisfied at each step [2, 17, 3, 18]. IGTS [5] proposes both TopDown and dynamic programming as optimization methods. On the other hand, BottomUp approaches begin with the maximum number of segments and merge them iteratively until a predefined error criterion is fulfilled. Probabilistic-based segmentation algorithms take into account the data distribution and transition times, employing methods such as Bayesian distribution [19], Hidden Markov Models [20], Gumbel distribution [7], and multivariate Gaussian distribution [6].

Apart from the aforementioned primary categories, other methods have been proposed. For example, FLUSS [4] leverages the assumption that a high probability of semantic change exists when only a few arcs intersect at a given index point. ClaSP [8] takes a unique approach by enriching a time series with a customized classification score profile using the self-supervision concepts [8].

However, many check point detection challenges, such as human activity segmentation, involve time series composed of heterogeneous data from different types of sensors. Most existing temporal segmentation methods are designed for single time series analysis. In contrast, our method extends the capabilities of ClaSP by enabling the analysis of multi-series data through a simple yet efficient merge strategy. This extension allows for more comprehensive and accurate segmentation results, taking into account the diverse information from multiple sensor streams.

Furthermore, traditional approaches for estimating the number of segments typically rely on comparing evaluation metrics on real data [21, 22, 5]. However, our observation is that real signals often exhibit a certain range of recognizable patterns. To address this, we generate synthetic signals that simulate real data, encompassing a wide range of patterns observed in real data sets. This synthetic data set offers significant advantages in terms of diversity and data volume.

Notably, to the best of our knowledge, there is currently no existing method for estimating the number of segments through an artificial dataset. In this regard, our proposed method stands out. By constructing a diverse synthetic dataset and training a dedicated discriminant model, we can predict the number of segments more accurately and robustly. This capability not only provides valuable insights into segment estimation but also assists subsequent methods in achieving improved performance.

In summary, while existing methods such as FLUSS and ClaSP have made notable contributions, our method expands the scope of ClaSP to analyze multi-series data and introduces a novel approach for estimating the number of segments using synthetic data. These advancements enhance the accuracy, flexibility, and applicability of segment analysis techniques, paving the way for further improvements in various domains.

3 Background

3.1 Change-Point Detection

Consider a multivariate time series $T = \{t_1, t_2, ..., t_l\}$ consisting of l observations, where each observation for a moment t is represented by a d-dimensional value $t_i \in \mathbb{R}^d$. The time series changes its behaviour at multiple moments $c_1, c_2, \ldots, c_N$.

$$\underbrace{t_1, t_2, ..., t_{c_1-1}}_{\text{Segment } g_1},\quad \underbrace{t_{c_1}, ..., t_{c_2-1}}_{\text{Segment } g_2},\quad t_{c_2}, ...,\quad \underbrace{t_{c_N}, ..., t_l}_{\text{Segment } g_N}$$

The change-point detection algorithm recognizes m change-points at moments $\hat{c}_1, \hat{c}_2, \ldots, \hat{c}_M$. Let $G = \{g_1, g_2, \ldots, g_N\}$ represent the set of ground truth segments split by $c_1, c_2, \ldots, c_N$, and $P = \{p_1, p_2, \ldots, p_M\}$ denote the set of predicted segments split by $\hat{c}_1, \hat{c}_2, \ldots, \hat{c}_M$.

3.2 Quality Metrics

To evaluate the change-point detection algorithm, we computes the F1-score by comparing the predicted and ground truth segments.

We first computes the intersection over union (IoU) between two segments. Given a predicted segment p_i and a ground truth segment g_j, when the intersection between p_i and g_j is $\emptyset$, the IoU is 0, otherwise it is calculated as:

$$\text{IoU}(p_i, g_j) = \frac{\min(\text{end}(p_i), \text{end}(g_j)) - \max(\text{start}(p_i), \text{start}(g_j))}{\max(\text{end}(p_i), \text{end}(g_j)) - \min(\text{start}(p_i), \text{start}(g_j))},$$

where $\text{start}(p_i)$ and $\text{end}(p_i)$ represent the starting and ending moments of the predicted segment p_i, respectively. Similarly, $\text{start}(g_j)$ and $\text{end}(g_j)$ represent the starting and ending moments of the ground truth segment g_j.

All IoU values populate a confusion matrix for each pair of predicted and ground truth segments. The confusion matrix is a $N \times M$ matrix, where N is the number of predicted segments and M is the number of ground truth segments. Each element of the matrix represents the IoU between a predicted segment and a ground truth segment. Next, the maximum IoU value for each predicted segment is determined by taking the maximum along the rows of the confusion matrix, resulting in a vector, $\boldsymbol{V}$ of length N. $\boldsymbol{V}$ contains the highest IoU value among ground truth segment for each predicted segment.

For each predicted segment, we iterate over the thresholds in the range 0.5 to 0.95 with a step size of 0.05. Let t represent a threshold value within this range. We compare the corresponding intersection over union (IoU) value in $\boldsymbol{V}$, denoted as $\boldsymbol{V}_j$ for the j-th predicted segment, to the threshold t. If the IoU value $\boldsymbol{V}_j$ is greater than or equal to the threshold t, the predicted segment is considered a true positive (TP). Similarly, if the IoU value $\boldsymbol{V}_j$ is less than the threshold t, the predicted segment is considered a false positive (FP):

$$TP = \sum_{j=1}^{N} \mathbb{I}(\boldsymbol{V}_j \geq t), \quad FP = \sum_{j=1}^{N} \mathbb{I}(\boldsymbol{V}_j < t),$$

where $\mathbb{I}(\cdot)$ represent the indicator function. To calculate the number of false negatives (FN), we subtract the true positives (TP) from the total number M of ground truth segments. We have:

$$FN = M - TP.$$

Once the values of TP, FP, and FN are computed for each threshold t, the F1-score can be calculated as:

$$\text{F1-score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}},$$

where precision is the ratio of true positives to the sum of true positives and false positives, and recall is the ratio of true positives to the sum of true positives and false negatives.

After calculating the F1-score at each threshold, the normed score is obtained by averaging the average F1-scores across all thresholds and all time series. This normed score provides an overall measure of the change point detection algorithm's performance, with higher values indicating better performance. In summary, the metrics provide a robust evaluation of the quality of change point detection algorithms applied to the time series.

4 Proposed Methods

We first introduce a model-based method using a discriminant model to find the optimal number of change points C of a TS by training on a synthetic signals data set in advance (see Fig. 1). We assume that a TS consists of a variety of continuous signals. This assumption comes from the fact that many signal series are easily distinguishable by the human. We furthermore assume that there is only one source of the signal in the same time period. Under these two assumptions, the task of change point detection is transformed into a multiple discriminant task. Therefore, a simple idea is to first train a discriminant model to determine whether a certain subsegment has a change point.

4.1 Synthetic Signals Generation

Our research is driven by the observation that real signals frequently exhibit distinct patterns that can be recognized and analyzed. To capture the essence of these patterns, we generate synthetic signals that closely simulate real data. This synthetic data set encompasses a wide range of patterns observed in real data sets, providing significant advantages in terms of diversity and data volume. Consequently, we employ these artificially generated synthetic signals to train our discriminant model.

The utilization of synthetic signals in our approach serves two important purposes. Firstly, it alleviates the need for collecting large-scale supervised data sets, effectively minimizing the associated overhead and resource requirements. By leveraging synthetic signals, we can generate an extensive set of training samples that represent various patterns and scenarios, augmenting the effectiveness of our discriminant model. This approach contributes to the reduction of manual data labeling efforts and facilitates more efficient model training.

Furthermore, the availability of a large number of training samples enhances the classification ability of our discriminant model. The diverse nature of the synthetic data set allows the model to learn and generalize from a wide spectrum of patterns and variations, enabling it to accurately classify and distinguish between different segments in real data.

We consider the following five basic signals to compose our analog signal training set: square wave signal, sinusoidal signal, sawtooth signal, stair signal and constant signal, as shown in Figure 1. At the same time, we also consider a variety of combination ways to construct complex signals, and add some noise to make the synthetic data set more realistic.

4.2 Discriminant Model for Change Point Detection

DMSS, the algorithm we propose in this paper, is based on change point discrimination and ClaSP [23] method, *i.e.*, it solves the change point detection problem by splitting a series into multiple discrimination problems, and then using a linear discriminant model to detect whether there exists a change point, as shown in Figure 1.

Let $T = \{t_1, t_2, ..., t_l\}$ be a time series consisting of l observations, we first computes $l - w + 1$ overlapping windows of width w with each being split into several sub segments $x_i = \{t_i, t_{i+1}, ..., t_{i+w-1}\}$, where $i \in [0, \lfloor l/w \rfloor]$. To detect

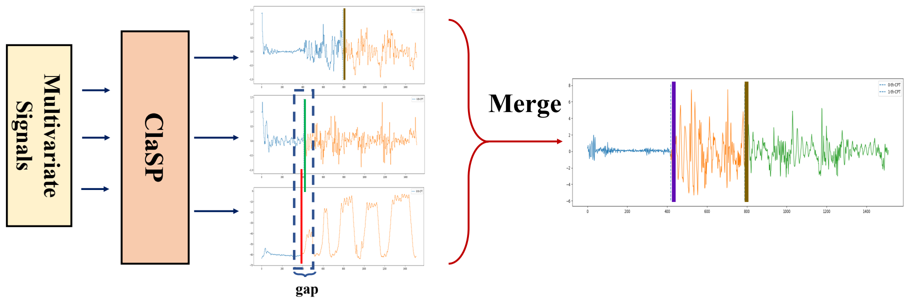

Fig. 2. Merging detected change points using three axises in Human Activity Segmentation dataset. Points within a gap will be merged by predefined weights.

change points using a logistic model, we define a binary response variable y_i, which indicates whether a change has occurred at sub segment i or not.

$$\log\left(\frac{y_i}{1-y_i}\right) = \beta_0 + \beta_1 \cdot x_i, \tag{1}$$

$$y_i = \frac{\exp(\beta_0 + \beta_1 \cdot x_i)}{1 + \exp(\beta_0 + \beta_1 \cdot x_i)}. \tag{2}$$

To obtain the number of change points in the entire time series, we can sum up the binary classification results mentioned above and take the average across all windows:

$$C = \frac{1}{L} \sum_{j\in[0,L]} \sum_{i\in[0,\lfloor l/w \rfloor]} y_i, \tag{3}$$

where $L = l - w + 1$.

The logistic regression model assumes a logistic relationship between the predictor variables and the log-odds of the binary response variable. And one important aspect to mention is that in this step, it involves, but is not limited to, the use of logistic regression. Any other binary classification or discriminant model can also be utilized. In our experiments, we employed eXtreme gradient boosting (XGBoost)[24], a boosting algorithm based on logistic regression, as an alternative approach.

4.3 Merging Multivariate Series

The original ClaSP method was primarily designed to address univariate time series problems, which limits its ability for multivariate sequences. This constraint becomes evident when considering scenarios involving multiple spatial signals associated with human body postures. For example, if the motion is confined to a single plane, relying solely on information from the y-axis would fail to capture comprehensive insights. In such cases, the incorporation of signals from the

Algorithm 1 Merge and Combine Change Points

```
Require:
S: Number of time sequences
Seq[1 : S]: Time sequences with change points
Weights[1 : S]: Weights corresponding to change points in each sequence
g: Time gap threshold for combining points
Ensure:
Merged_Seq: Merged time sequence
procedure MERGE_AND_COMBINE(N, Seq, Weights, g)
    Merged_Seq ← []
    for i ← 1 to S do
        combined_point ← Seq[i][1]
        combined_weight ← Weights[i][1]
        for j ← 2 to length(Seq[i]) do
            if Seq[i][j] − combined_point > g then
                Merged_Seq.append(combined_point)
                combined_point ← Seq[i][j]
                combined_weight ← Weights[i][j]
            else
                combined_weight ← combined_weight + Weights[i][j]
            end if
        end for
        Merged_Seq.append(combined_point)
    end for
    return Merged_Seq
end procedure
```

x-axis and z-axis becomes critical to accurately detect change points. Therefore, the integration of information from multiple sensors assumes paramount importance in the context of change point detection, enabling a more comprehensive understanding of complex multivariate data.

To tackle this challenge, we propose a straightforward merging method based on interval weights, akin to a voting approach, as depicted in Figure 2. Given S sequences with a time gap, denoted as g, between them, and assuming their mutual independence, we assign weights, denoted as w_i, to each sequence. Initially, we apply an individual change point detection method to each sequence. Subsequently, utilizing the assigned weights, we merge the detected change points within each segment by considering the respective weights within an interval surrounding the split points. The detailed algorithm for this merging process is defined in Algorithm 1.

5 Experiments

To evaluate the performance of various methods accurately, we conducted a series of experiments using our self-labeled Human Activity Segmentation dataset [11]. In this section, we describe the setup of our experiments, including the evaluation metric, the choice of the discriminant model, and the design of weights for merging points.

5.1 Dataset

We utilized the Human Activity Segmentation dataset [11], which is a collection of labeled human activity sequences. The challenge involved collecting and annotating 10.7 hours of real-world multi-dimensional time series (TS) data. The dataset consists of 250 TS, each comprising twelve dimensions and sampled at a frequency of 50 Hertz (Hz). These TS were recorded using various smartphone sensors and captured the performance of 100 different human activities. Sixteen bachelor students participated in the data collection, showcasing diverse motion sequences during the activities. The TS data ranges from seven seconds to fourteen minutes in duration, with a median duration of 100 seconds. Within each TS, a varying number of potentially recurring activities are present, and each activity has its own variable time duration. The main challenge task is to predict the precise locations of activity changes without availability of ground truth labels.

5.2 Discriminant Model

For the discriminant model, we employed XGBoost [25] as a straightforward implementation. XGBoost is a popular gradient boosting algorithm known for its robustness and effectiveness in various machine learning tasks. We used the default parameters of XGBoost to ensure a fair comparison among different methods.

5.3 Merging of Points

To merge neighboring points and obtain coherent activity segments, we designed weights based on a predefined time gap, denoted as g. We set g to be 120 units of time, representing a reasonable duration for consecutive activities. The weight vector for merging points was defined as $1, 0, ..., 0$, where the first element has a weight of 1 and the remaining elements have weights of 0. This design choice ensured that only the first point within the time gap was selected, effectively merging subsequent points.

5.4 Result

Table 1 illustrates the results of our experiments, comparing the performance of our proposed method, DMSS, with the original ClaSP method [23]. DMSS achieved the highest F1-score of 0.411, outperforming the performance of ClaSP and ranked the third place in HAS challenge. These results highlight the effectiveness of our approach in accurately segmenting human activities.

Table 1. Performance on Human Activity Segmentation dataset.

Method	FLUSS	BinSeg	GSS	IGTS	STRAY	ClaSP	DMSS(ours)
F1-Score	0.214	0.263	0.152	0.141	0.227	0.395	**0.411**

6 Discussion

In this work, we propose a method for change point detection in time series based on the ClaSP method. Additionally, we train an additional discriminant model to accurately determine the number of segmentation points. To ensure sufficient training of the discriminant model, we create and utilize synthetic simulated data. Furthermore, in order to fully leverage the information from multiple sequences, we present a simple yet effective merging method based on weights and time intervals, which provides a more robust and efficient approach to segmenting time series from multiple perspectives.

Experimental results on the Human activity segmentation dataset demonstrate that our proposed DMSS (discriminant Model for Change Point Detection with Sequence Merging) method outperforms the original ClaSP method, achieving higher F1-Score. However, it should be noted that the current synthetic data used in our experiments has a relatively fixed composition of basic signal components. Future work will explore the incorporation of statistical characteristics of target signals into the generation process of the synthetic data set, aiming to enhance the realism and versatility of the simulated data.

7 Acknowledgments

This work is supported by National Key R&D Program of China (2022ZD0114805), NSFC (61773198, 61921006, 62006112), Collaborative Innovation Center of Novel Software Technology and Industrialization, NSF of Jiangsu Province (BK20200313).

References

1. Michele Basseville, Igor V Nikiforov, et al. *Detection of abrupt changes: theory and application*, volume 104. prentice Hall Englewood Cliffs, 1993.
2. Jushan Bai. Estimating multiple breaks one at a time. *Econometric theory*, 13(3):315–352, 1997.
3. Piotr Fryzlewicz. Wild binary segmentation for multiple change-point detection. *The Annals of Statistics*, 42(6):2243–2281, 2014.
4. Shaghayegh Gharghabi, Yifei Ding, Chin-Chia Michael Yeh, Kaveh Kamgar, Liudmila Ulanova, and Eamonn Keogh. Matrix profile viii: domain agnostic online semantic segmentation at superhuman performance levels. In *IEEE international conference on data mining*, pages 117–126, 2017.
5. Amin Sadri, Yongli Ren, and Flora D Salim. Information gain-based metric for recognizing transitions in human activities. *Pervasive and Mobile Computing*, 38:92–109, 2017.

6. David Hallac, Peter Nystrup, and Stephen Boyd. Greedy gaussian segmentation of multivariate time series. *Advances in Data Analysis and Classification*, 13(3):727–751, 2019.
7. Priyanga Dilini Talagala, Rob J Hyndman, and Kate Smith-Miles. Anomaly detection in high-dimensional data. *Journal of Computational and Graphical Statistics*, 30(2):360–374, 2021.
8. Patrick Schäfer, Arik Ermshaus, and Ulf Leser. Clasp-time series segmentation. In *Proceedings of the 30th ACM international conference on information & knowledge management*, pages 1578–1587, 2021.
9. Samaneh Aminikhanghahi and Diane J Cook. A survey of methods for time series change point detection. *Knowledge and information systems*, 51(2):339–367, 2017.
10. Charles Truong, Laurent Oudre, and Nicolas Vayatis. Selective review of offline change point detection methods. *Signal Processing*, 167:107299, 2020.
11. Arik Ermshaus, Patrick Schäfer, Ulf Leser, Anthony Bagnall, Romain Tavenard, Colin Leverger, Vincent Lemaire, Simon Malinowski, Thomas Guyet, and Georgiana Ifrim. Human activity segmentation challenge. ECML/PKDD 2023 Discovery Challenge, 2023.
12. Vana Panagiotou. *Blind segmentation of time-series: A two-level approach*. PhD thesis, Delft University of Technology, 2015.
13. Johan Himberg, Kalle Korpiaho, Heikki Mannila, Johanna Tikanmaki, and Hannu TT Toivonen. Time series segmentation for context recognition in mobile devices. In *Proceedings of IEEE international conference on data mining*, pages 203–210, 2001.
14. Heli Hiisilä. *Segmentation of Time Series and Sequences Using Basic Representations*. PhD thesis, Helsinki University of Technology, 2007.
15. Ath Kehagias, Ev Nidelkou, and V Petridis. A dynamic programming segmentation procedure for hydrological and environmental time series. *Stochastic Environmental Research and Risk Assessment*, 20:77–94, 2006.
16. Oresti Banos, Juan-Manuel Galvez, Miguel Damas, Hector Pomares, and Ignacio Rojas. Window size impact in human activity recognition. *Sensors*, 14(4):6474–6499, 2014.
17. Jing Yuan, Yu Zheng, Xing Xie, and Guangzhong Sun. T-drive: Enhancing driving directions with taxi drivers' intelligence. *IEEE Transactions on Knowledge and Data Engineering*, 25(1):220–232, 2011.
18. Wei Cheng, Xiang Zhang, Feng Pan, and Wei Wang. Hicc: an entropy splitting-based framework for hierarchical co-clustering. *Knowledge and Information Systems*, 46:343–367, 2016.
19. Ryan Prescott Adams and David JC MacKay. Bayesian online changepoint detection. *CoRR*, 2007.
20. Taketoshi Mori, Yu Nejigane, Masamichi Shimosaka, Yushi Segawa, Tatsuya Harada, and Tomomasa Sato. Online recognition and segmentation for time-series motion with hmm and conceptual relation of actions. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3864–3870, 2005.
21. Andrew Foss and Osmar R Zaïane. A parameterless method for efficiently discovering clusters of arbitrary shape in large datasets. In *IEEE International Conference on Data Mining*, pages 179–186, 2002.
22. R Scott Harris, Dean R Hess, and José G Venegas. An objective analysis of the pressure-volume curve in the acute respiratory distress syndrome. *American journal of respiratory and critical care medicine*, 161(2):432–439, 2000.
23. Arik Ermshaus, Patrick Schäfer, and Ulf Leser. Clasp: parameter-free time series segmentation. *Data Mining and Knowledge Discovery*, 37(3):1262–1300, 2023.

24. Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 785–794, 2016.
25. Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.