Exposé for Student Research Project: Generator of Synthetic Time Series for Motif Discovery

Rafael Moczalla[1]

[1]Humboldt University of Berlin

September 26, 2018

1 Introduction

In colloquial language the motif discovery problem donates the problem of finding interesting patterns in sequences. The application of motif discovery can be found in a variety of domains like the discovery of DNA and protein sequence motifs of Bailey et al., 2009, and the automatic pattern recognition in ECG of Sternickel, 2000. Since the problem of finding interesting patterns or motifs is subjective and domain dependent the motif discovery problem is difficult to model mathematically.

In Literature research on motif discovery splits into the time series top motif pair and the time series top motif set discovery problem as mentioned by Grabocka et al., 2015. The time series top motif pair discovery problem donates the problem of finding "[t]he most significant nearest neighbor [subsequence pair $\{S_1, S_2\}$] in a time series [. . .] such that the subsequence [S_1] has minimal distance to its non-trivial nearest neighbor [S_2]", according to Dragomir Yankov and et al., 2007. In contrast, the time series top motif set discovery donates the problem of finding "the subsequence [in a time series] that has the highest count of non-trivial matches" of very similar subsequences, according to Lin et al., 2002a. Figure 1 illustrates a time series top motif pair example and figure 2 illustrate a time series top motif set example. There are different ways to evaluate time series top motif pair and time series top motif set discovery algorithms. In the case of time series top motif pair algorithms the algorithm complexity and runtime is compared to the complexity and runtime of the brute

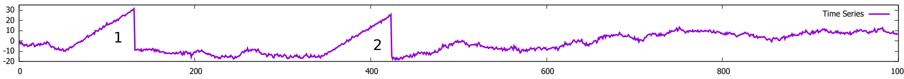

Figure 1: Synthetic time series example of length 1000. The time series contains two positive flanks of length 100 and height 50 as well as 1 % noise. The similar positive flanks are marked with the numbers 1 and 2. In our use-case is $\{1, 2\}$ the time series top motif pair.

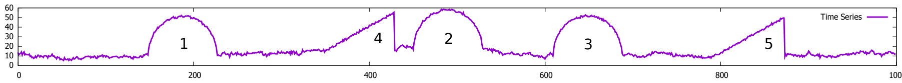

Figure 2: Synthetic time series example of length 1000. The time series contains three semicircular subsequences and two positive flanks of length 100 and height 50 as well as 1 % noise. The similar semicircular time series subsequences are marked with the numbers 1 to 3 and the similar positive flanks are marked with the numbers 4 and 5. In our use-case is $\{1, 2, 3\}$ the time series top motif set.

force time series top motif pair discovery algorithm like Mueen et al., 2009, did. The runtime experiments are conducted on different existing databases like the entomologists data archive presented by Stafford et al., 2009. In the case of time series top motif set discovery algorithms the algorithm complexity is compared to the complexity of the brute force time series top motif set discovery algorithm like Grabocka et al., 2015, Yeh et al., 2016, and Lin et al., 2002b, did as the brute force time series top motif set algorithm is computationally very expansive. Similar to the experimental evaluation of the the time series top motif pair algorithms the time series top motif set algorithm experiments are conducted on different existing databases like Chen et al., 2015, offer. Since the scientific community manually annotate the data to analyze the algorithms the experimental databases are relatively small and contain relatively small time series. We see a need for a synthetic time series generator to benchmark both, the speed of time series top motif pair and the accuracy of time series top motif set discovery algorithms.

We pursue three main goals in this student research project.

1. We categorize the approaches of evaluation methods on time series top motif pair and time series top motif set discovery algorithms in an overview.

2. We implement a synthetic time series generator. The generator calculates time series containing a defined set of time series motif pairs or time series motif sets.

3. We design and perform a benchmark for time series top motif pair and time series top motif set discovery algorithms.

2 Background

In literature, the motif discovery in time series subdivides into the time series motif pair discovery and the time series motif set discovery. The time series top motif pair discovery denotes the problem of finding the most similar pair of subsequences, given some subsequence length, using a distance-based similarity measure. Mueen et al., 2009, tackle with the time series top motif pair problem. The time series top motif set discovery denotes the problem of finding the largest set of subsequences in a time series where all subsequences in the set are within radius $2r$, according to a distance-based similarity measure, given some subsequence length and a radius r. Lin et al., 2002a, tackle with the time series top motif set problem.

To be precise, we present the definitions of time series motifs in this section. We distinguish between

1. the time series top motif pair,
2. the time series k^{th} motif pair,
3. the time series top motif set and finally,
4. the time series k^{th} motif set.

First, we need to define the terms time series, subsequence, similarity measure and matching subsequences.

Definition 2.1 (Time Series) *The time series $T = (t_1, t_2, \ldots, t_n)$ is a ordered sequence of n real valued numbers.*

Definition 2.2 (Subsequence) *A time series $T_{i,l} = (t_i, t_{i+1}, \ldots, t_{i+l-1})$ of length l and at offset i is a subsequence of another time series $T = (t_1, t_2, \ldots, t_n)$ if the sequence $T_{i,l}$ is contained in T starting at the offset i in T where $1 \leq i < i + l - 1 \leq n$.*

Definition 2.3 (Similarity Measure) *Let $T_{i,l}$ and $T_{j,l}$ be two subsequences of length l. The similarity of $T_{i,l}$ and $T_{j,l}$ is measured using a distance metric $d(T_{i,l}, T_{j,l}) \in \mathbb{R}$, such as the Euclidean metric.*

A similarity measure can be interpreted as the inverse of the distance measure. High values indicate low similarity and low values indicate high similarity. We present the definition of matching subsequences as follows.

Definition 2.4 (Matching Subsequences) *Let $T_{i,l}$ and $T_{j,l}$ be two time series of length l, $d(\cdot, \cdot)$ be a similarity measure and t be a value called range. $T_{i,l}$ and $T_{j,l}$ are matching if their distance is within t, $d(T_{i,l}, T_{j,l}) \leq t$.*

Furthermore, we define the notion of non-self matching subsequences to avoid trivial matches of overlapping subsequences.

Definition 2.5 (Non-Self Matching Subsequences) *Let $T_{i,l}$ and $T_{j,l}$ be two subsequences of length l of the same time series and $d(\cdot, \cdot)$ be a similarity measure. $T_{i,l}$ and $T_{j,l}$ are non-self matching if they are matching but not overlapping, i.e. $i + l \leq j$.*

We are finally in a position to define time series motifs. We define the term of time series top motif pair as follows.

Definition 2.6 (Time Series Top Motif Pair) *Given a time series T and a similarity measure $d(\cdot, \cdot)$. The time series top motif pair is the unordered non-overlapping pair $\{T_{i,l}, T_{j,l}\}$ of subsequences of length l in T with the smallest distance $d(T_{i,l}, T_{j,l})$ among all non-self matching subsequence pairs in T.*

Next, by ranking the subsequence pairs according to their distance, we can determine the time series k^{th} motif pairs.

Definition 2.7 (Time Series k^{th} Motif Pair) *Given a time series T and a similarity measure $d(\cdot,\cdot)$. The time series k^{th} motif pair is the unordered pair $\{T_{i,l}, T_{j,l}\}$ of subsequences in T with k^{th} smallest distance $d(T_{i,l}, T_{j,l})$ among all non-self matching subsequences in T.*

To move from pairs of subsequences (time series motif pairs) to frequently repeated subsequence (time series motifs sets), we define the term of a time series top motif set as follows.

Definition 2.8 (Time Series Top Motif Set) *Given a time series T, a similarity measure $d(\cdot,\cdot)$ and a radius r. The time series top motif set is the largest set of non-self matching subsequences M in T within distance $2r$, i.e. $\forall T_{i,l}, T_{j,l} \in M : d(T_{i,l}, T_{j,l}) \leq 2r$ as well as $\forall R_{k,l} \notin M : \exists S \in M$ such that $d(R_{k,l}, S) > 2r$.*

The time series top motif set can be interpreted as the most frequently repeated subsequence in the time series. When ranking the sets according to their size, we define the time series k^{th} motif set, the generalization of the time series top motif set.

Definition 2.9 (Time Series k^{th} Motif Set) *Given a time series T and a similarity measure $d(\cdot,\cdot)$. The time series k^{th} motif set is the k^{th} largest set of non-self matching subsequences in T within distance $2r$.*

As a single subsequence may be contained in several time series motif sets, motif sets can overlap.

3 Related Work

An extensive overview of the general field of time series data mining is given by Esling et al., 2012, including definitions, tasks, implementation components and a categorization of the existing literature till 2012.

As defined by Mueen et al., 2009, a time series top motif pair is the most similar pair of time series in a database of time series. The database maybe considered as the subsequences of a time series. With this definition Mueen et al., 2009, present a tractable exact algorithm to discover k^{th} motif pairs with runtime $O(m \log m)$ in average and $O(m^2)$ in worst case where m is the time series length.

Yeh et al., 2016, revealed another approach for finding the time series top motif pair in a time series. They define the concepts of a matrix profile and a matrix profile index. The matrix profile is the distance and the matrix profile index is the location of all subsequences nearest neighbors of one time series in another time series according to a distance-based similarity measure. First, the STAMP algorithm calculates the matrix profile as well as the matrix profile index of two time series. Second, these two meta data objects contain the information to discover the time series top motif pair. Overall runtime of $O(m^2 \log m)$ where m is the time series length. In there experiments the growth rate of there algorithm was roughly $O(m^2)$ instead of $O(m^2 \log m)$.

Grabocka et al., 2015, introduce an approximate time series k^{th} motif set discovery approach. They translate the time series k^{th} motif set discovery problem into a principled optimization problem. The frequency of a time series

k^{th} motif set is the size of the time series k^{th} motif set. There solution is based on maximizing a differentiable frequency function.

An algorithm for approximate time series motif set discovery is proposed by Senin et al., 2014. First, the tool GrammarViz 2.0 discretizes the real-valued time series. SAX (Patel et al., 2002) performs z-normalization and divides the time series into segments of equal size for further processing. Afterwards, the resulting symbolic series is parsed and decomposed into a context free grammar. Finally, time series k^{th} motif sets are discovered by evaluation of the grammar rule hierarchy and grammar rule counts. The assumption is that frequently used rules are likely to correspond to frequent subsequences in the time series.

Some scientist use synthetic time series. Bagnall et al., 2017, introduce a generator of simulated time series classification problems. Since, the time series classification is the problem of identifying to which set of categories a time series belong the generator of simulated time series classification problems produces only time series for a specific classification problem.

4 Objective

As part of the research project we work out a summary of known methods, like datasets, generators, and so forth, that are already used in the evaluation processes of time series motif pair and time series motif set discovery algorithms.

Also, we will set up a synthetic time series generator designed for the evaluation of time series motif pair and time series motif set discovery algorithms. The generated synthetic time series will contain clearly defined time series motif pairs and time series motif sets. The main criteria for the generator are

1. a pipeline in `C++`,
2. time series output in form of a `CSV` file,
3. time series motif locations output in form of a `CSV` file,
4. a graphical representation in `PDF` format and
5. a set of default arguments.

Finally, we will generate and perform a benchmark with use-cases for time series motif pair and time series motif set discovery algorithms. The benchmark consists of `CSV` files.

5 Approach

To accomplish our aims we study previous scientific publications containing an evaluation of time series motif pair and time series motif set discovery algorithms.

Evaluation Overview

The summary of evaluation methods on time series motif pair and time series motif set discovery algorithms will include both a taxonomy, e.g. real world data, synthetic data and so forth, as well as pros and cons of each evaluation method.

Time Series Generator

The synthetic time series generator will produce the time series by adding continuously and randomly a time series value, a time series motif pair or time series motif set subsequence. The generator also stores the corresponding locations in a separate file while inserting a time series motif pair or time series motif set subsequence.

We offer the user a variety of options to setup the synthetic time series generator. The user chooses the time series length and optional the variance of the time series length for random length generation as well as a base value for the random generation of the time series values. The time series values are generated randomly around the base value. Also, the user sets the noise ratio for the base time series and extra the noise ratio for the time series motif pair or time series motif set subsequences. Moreover, the user hands over a list of time series motif pair or time series motif set types containing in addition the corresponding amount, length and height for each time series motif pair or time series motif set type. More precisely, we parameterize the generator by arguments like

1. the base value of the time series values,
2. the length and variance of the time series,
3. the length and a warp factor of the inserted time series motif pair or time series motif set subsequences,
4. the height and a height scale factor of the inserted time series motif pair or time series motif set subsequences,
5. the number of different time series motif pair or time series motif set types,
6. the number of each time series motif pair or time series motif set subsequences for each type,
7. the amount of noise, which cover the base time series as well as
8. the amount of noise, which cover the time series motif pair or time series motif set subsequences.

A visualization of the arguments is presented in figure 3.

Benchmark

We use our synthetic time series generator to design a benchmark, a set of files, containing the synthetic time series as well as the synthetic time series motif pairs or synthetic time series motif sets locations with regard to the synthetic time series.

To evaluate the runtime of a time series motif pair discovery algorithm we calculate the runtime of the MK time series motif pair discovery by Mueen et al., 2009. To evaluate the accuracy of a time series motif set discovery algorithm we run various time series motif set algorithms on the synthetic times series.

All in all, the benchmark will be tested and performed with

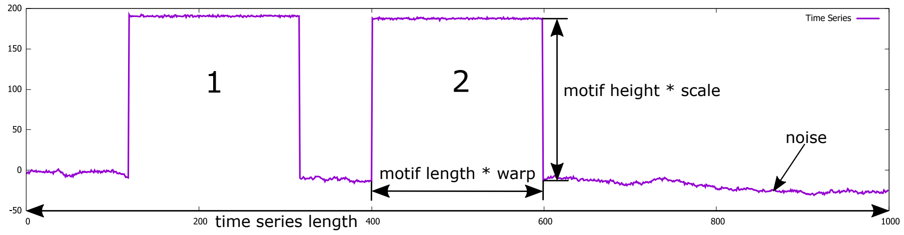

Figure 3: Synthetic time series example with labels visualizing the arguments of the synthetic time series generator. The time series of length 1000 has the base value 0.0 and contains a time series motif of length 200 and height 200 as well as 1 % noise. The box subsequence repeats itself two times and is marked with the numbers 1 and 2.

1. the MK time series motif pair discovery by Mueen et al., 2009,
2. the time series latent motif discovery by Grabocka et al., 2015,
3. the GrammarViz time series motif set discovery by Senin et al., 2014, and
4. the Matrix Profile time series motif discovery by Yeh et al., 2016.

We record the runtime of the MK time series motif pair algorithm and we record the runtime and the located time series motif sets of each time series motif set discovery algorithm run.

Our work will be presented in a scientific paper and we offer the source code on a git repository. The whole project will be documented in Doxygen style providing an HTML.

References

Bagnall, Anthony, Aaron Bostrom, James Large, and Jason Lines (Mar. 2017). “Simulated Data Experiments for Time Series Classification Part 1: Accuracy Comparison with Default Settings”. In: *arXiv* 1703.09480, pp. 1–28. URL: https://arxiv.org/pdf/1703.09480.

Bailey, Timothy L, Mikael Boden, Fabian A Buske, Martin Frith, Charles E Grant, Luca Clementi, Jingyuan Ren, Wilfred W Li, and William S Noble (July 2009). “MEME SUITE: Tools for Motif Discovery and Searching”. In: *Nucleic acids research* 37.2, pp. 202–208. URL: https://academic.oup.com/nar/article/37/suppl_2/W202/1135092.

Bianco, Vincenzo, Oronzio Manca, and Sergio Nardini (Sept. 2009). “Electricity Consumption Forecasting in Italy using Linear Regression Models”. In: *Energy* 34.9, pp. 1413–1421. URL: https://s3.amazonaws.com/academia.edu.documents/46988248/Electricity_consumption_forecasting_in_I20160703-23011-1opbod.pdf?AWSAccessKeyId=AKIAIWOWYYGZ2Y53UL3A&Expires=1537958014&Signature=YUkuCwX4CJlFegJ%2FVwdaAriwYwc%3D&response-content-disposition=inline%3B%20filename%3DElectricity_consumption_forecasting_in_I.pdf.

Chen, Yanping, Eamonn Keogh, Bing Hu, Nurjahan Begum, Anthony Bagnall, Abdullah Mueen, and Gustavo Batista (July 2015). “The UCR Time Series Classification Archive”. In: URL: http://www.cs.ucr.edu/~eamonn/time_series_data/.

Dragomir Yankov and, Eamonn Keogh, Jose Medina, Bill Chiu, and Victor Zordan (Aug. 2007). “Detecting time series motifs under uniform scaling”. In: pp. 844–853. URL: https://www.cs.ucr.edu/~eamonn/motifs_under_scaling.pdf.

Esling, Philippe and Carlos Agon (Nov. 2012). “Time-Series Data Mining”. In: *ACM Computing Surveys (CSUR)* 45.1, 12:1–12:34. URL: https://hal.archives-ouvertes.fr/hal-01577883/document.

Grabocka, Josif, Nicolas Schilling, and Lars Schmidt-Thieme (May 2015). “Latent Time-Series Motifs”. In: *ACM Transactions on Knowledge Discovery from Data (TKDD)* 11, 6:1–6:20. URL: https://www.ismll.uni-hildesheim.de/pub/pdfs/grabocka2016a-tkdd.pdf.

Lin, Jessica, Eamonn Keogh, Stefano Lonardi, and Pranav Patel (2002a). “Finding Motifs in Time Series”. In: *Proceedings of the Second Workshop on Temporal Data Mining*, pp. 53–68. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.19.6629&rep=rep1&type=pdf.

— (Oct. 2002b). “Finding Motifs in Time Series”. In: *CiteSeerX* 10.1.1.19.6629, pp. 1–11. URL: https://cs.gmu.edu/~jessica/Lin_motif.pdf.

Mueen, Abdullah, Eamonn Keogh, Qiang Zhu, Sydney Cash, and Brandon Westover (Oct. 2009). “Exact Discovery of Time Series Motifs”. In: *Proceedings of the 2009 SIAM International Conference on Data Mining* 9781611972795.41, pp. 473–484. URL: http://alumni.cs.ucr.edu/~mueen/pdf/EM.pdf.

Patel, Pranav, Eamonn Keogh, Jessica Lin, and Stefano Lonardi (Dec. 2002). “Mining Motifs in Massive Time Series Databases”. In: *Proceedings of the 2002 IEEE International Conference on Data Mining* 1, pp. 370–377. URL: https://cs.gmu.edu/~jessica/publications/motif_icdm02.pdf.

Senin, Pavel, Jessica Lin, Xing Wang, Tim Oates, Sunil Gandhi, Arnold P. Boedihardjo, Crystal Chen, Susan Frankenstein, and Manfred Lerner (Sept. 2014). “GrammarViz 2.0: A Tool for Grammar-Based Pattern Discovery in Time Series”. In: *Machine Learning and Knowledge Discovery in Databases* 8726, pp. 468–472. URL: https://cs.gmu.edu/~xwang24/papers/grammarviz2.pdf.

Stafford, C. and G. Walker (Jan. 2009). “Characterization and Correlation of DC Electrical Penetration Graph Waveforms with Feeding Behavior of Beet Leafhopper, Circulifer Tenellus”. In: *Entomologia Experimentalis et Applicata* 130.2, pp. 113–129. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1570-7458.2008.00812.x.

Sternickel, Karsten (June 2000). “Automatic pattern recognition in ECG time series”. In: *Computer Methods and Programs in Biomedicine* 68.2, pp. 109–115. URL: http://www.computing.northampton.ac.uk/~scott/csy3025/sdarticle6.pdf.

Yeh, Chin-Chia Michael, Yan Zhu, Liudmila Ulanova, Nurjahan Begum, Yifei Ding, Hoang Anh Dau, Diego Furtado Silva, Abdullah Mueen, and Eamonn Keogh (Dec. 2016). “Matrix Profile I: All Pairs Similarity Joins for Time Series: A Unifying View That Includes Motifs, Discords and Shapelets”.

In: *2016 IEEE 16th International Conference on Data Mining (ICDM)* 1, pp. 1317–1322. URL: http://www.cs.ucr.edu/~eamonn/PID4481997_extend_Matrix%20Profile_I.pdf.