University of Zurich[UZH]

Department of Informatics

University of Zürich
Department of Informatics
Binzmühlestr. 14
CH-8050 Zürich
Phone. +41 44 635 43 11
Fax +41 44 635 68 09
www.ifi.uzh.ch/dbtg

UZH, Dept. of Informatics, Binzmühlestr. 14, CH-8050 Zürich

Prof. Dr. Michael Böhlen
Professor
Phone +41 44 635 43 33
Fax +41 44 635 68 09
boehlen@ifi.uzh.ch

Zürich, September 7, 2021

MSc Thesis
Exhaustive enumeration of variable-length motifs in time series

Time series motifs are equal-length subsequences of time series that are similar to each other. An example motif in a single time series is illustrated in Figure 1 where the two highlighted patterns are similar to each other. Detecting a priori unknown similar patterns has been used extensively for time series analytics and the analysis of genomic sequence datasets.

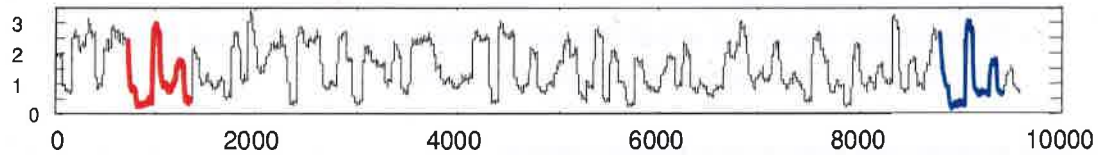

Figure 1. Illustration of Motifs from Mueen at al. [4, p. 1]

Existing algorithms find motifs for a given length. An obvious brute force approach to enumerate motifs of all lengths is to run a motif discovery algorithm for each possible length. Clearly, such an algorithm is computationally very expensive and prohibitive for large real-world time series. Mueen proposed the exact MOEN algorithm that uses a lower bound on the normalized Euclidean distance to not re-discover the same motif at different lengths. The performance of the MOEN algorithm makes it possible to determine motifs of all possible lengths.

The goal of this MSc thesis is to precisely define the problem of discovering motifs of all lengths and to implement and evaluate a scalable exact algorithm to solve this problem. An optional task is apply the implemented techniques to develop a solution that makes it possible to find a good window length for determining all-pair cross-correlations.

Tasks

1. **Task 1: Literature review and data preparation**
   - Study the relevant research work on motif discovery [3, 2, 1].
   - Acquire and prepare a dataset for the continuous validation of your solution. For example, use time series from `https://opendata.swiss` with temperature, pressure, humidity, and rain duration values.
   - Implement Algorithm 1 from Mueen [3] and run it on your data. Implement a solution to visualize and explore motifs. Describe your solution in a report.

2. **Task 2: Implement Mueen's enumeration algorithm to find all variable length motifs**
   - Precisely define the problem of enumerating all variable-length motifs. Strive for a solution that is parameter-free and minimizes required human interventions.
   - Implement your version of the MOEN algorithm [3] that enumerates all variable-length motifs as defined above. Focus on a careful design and implementation of your data structures and algorithms to get a scalable solution.
   - Analytically and empirically evaluate the properties and scalability of your solution. Use appropriate synthetic and real-world time series to evaluate the precision of your solution.

3. **Task 3 [optional]: Determine all-pair (cross-)correlations in a set of time series**
   - A problem that is related to the detection of all-lengths motifs is the detection of all-pair correlations in a set of time series. The correlation is usually measured over a window and this window length is not known a priori. One possibility to determine the window length is to compute the correlation between windows of all lengths and select the window length based on the observed correlations.
   - Compute and display the all-pair Pearson correlation and let the user interactively explore the correlations for different window lengths.

4. **Task 4: Write and defend your Master's thesis**
   - Describe the implementations, results and evaluations in your Master's thesis.
   - Present and defend your Master's thesis in the DBTG group meeting.

References

[1] Nuno Filipe Castro and Paulo J. Azevedo. Time series motifs statistical significance. In *Proceedings of the Eleventh SIAM International Conference on Data Mining, SDM 2011, April 28-30, 2011, Mesa, Arizona, USA*, pages 687–698. SIAM / Omnipress, 2011. `doi:10.1137/1.9781611972818.59`.

[2] Yifeng Gao and Jessica Lin. Discovering subdimensional motifs of different lengths in large-scale multivariate time series. In Jianyong Wang, Kyuseok Shim, and Xindong Wu, editors, *2019 IEEE International Conference on Data Mining, ICDM 2019, Beijing, China, November 8-11, 2019*, pages 220–229. IEEE, 2019. doi:10.1109/ICDM.2019.00032.

[3] Abdullah Mueen. Enumeration of time series motifs of all lengths. In *2013 IEEE 13th International Conference on Data Mining*, pages 547–556, 2013. doi:10.1109/ICDM.2013.27.

[4] Abdullah Mueen, Eamonn Keogh, Qiang Zhu, Sydney Cash, and M Brandon Westover. Exact discovery of time series motifs. volume 2009, pages 473–484, 04 2009. doi:10.1137/1.9781611972795.41.

Supervisors:

- Prof. Dr. Michael Böhlen (boehlen@ifi.uzh.ch)

Start Date: September 6, 2021

End Date: March 6, 2022

University of Zurich
Department of Informatics

Prof. Dr. Michael Böhlen
Professor