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

Lukas Vollenweider

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

Zürich, February 18, 2022

MSc Thesis
Datenbanktechnologie

Topic: Optimization of Lempel-Ziv Entropy Rate Estimator for Numerical Time series

Time series entropy rate estimation is an important problem which is crucial for many applications such as time series similarity, time series complexity estimation and data compression. The Lempel-Ziv Estimator is one of the most used estimators since it converges to the real entropy rate with probability 1 if the time series is stationary and ergodic [2].

As described by Thomas and Cover [1] and Kontoyiannis et al. [2] the Lempel-Ziv algorithm is used when the series is symbolic. The computational complexity of the naive implementation of the Lempel-Ziv algorithm is $O(n^2)$, where n is the length of the series. An optimized implementation of the algorithm for symbolic series has a time complexity of $O(n)$ as mentioned in Lang et al. [3]. This is because it uses a dictionary based look-up to compute the Lempel-Ziv coefficients which are necessary for the computation of the entropy rate. Lang et al. [3] also adapt the optimal Lempel-Ziv algorithm to numerical time series.

The main goal of this Master thesis is to understand, design and implement a Lempel-Ziv entropy rate estimator based on the approach described in Lang et al. [3] for numerical time series and investigate its run time complexity with respect to the brute-force implementation.

Tasks:

1. **Literature Review:**
   - Study relevant literature [1] sections 13.4 and 13.5, [3] to understand the algorithm and have a background about its computational complexity, implementation and applications.

2. **Implement an interval based brute force Lempel-Ziv entropy estimator for numeric time series**
    - Describe the worst case run time complexity
    - Compute Entropy rate on real data using the expression described in [2]
    - Analyse how the estimated entropy rate behaves with changing the length of the series and the number of distinct values that the series takes.
3. **Implement the approach in [3] for entropy rate estimation.**
    - Describe the run time complexity of the implementation
    - Experiment it with respect to the behavior of different publicly available datasets.
4. **Evaluation:** using task 2 as a baseline, assess the run time complexity and the goodness of the estimated entropy rates of the implementation of task 3
5. **Optional:**
    - Improve the run time complexity of the approach proposed in [3]
6. **Write a concise document of your work**
    - Describe the implementations, results, evaluations and findings.
    - Present and defend your thesis in a DBTG meeting.

Supervisor: Jamal Mohammed (mjamal@ifi.uzh.ch)
Start date: 01.03.2022
End date: 31.08.2022

References

[1] Thomas M. Cover, Joy A. Thomas. Elements of Information Theory, 2nd Edition. Book July,2006.

[2] Kontoyiannis I. and Algoet P. H. and Suhov Yu. M. and Wyner A. J. Nonparametric Entropy Estimation for Stationary Processes and Random Fields, with Applications to English Text. IEEE Transactions on Information Theory, vol. 44,pp.1319-1327,1998.

[3] W. Lang, M. Morse and J. M. Patel, "Dictionary-Based Compression for Long Time-Series Similarity," in IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 11, pp. 1609-1622, Nov. 2010, doi: 10.1109/TKDE.2009.201

Department of Informatics, University of Zurich

Prof. Dr. Michael Böhlen