The Journal of Open Source Software

STUMPY: A Powerful and Scalable Python Library for Time Series Data Mining

Sean M. Law[1]

1 TD Ameritrade

DOI: 10.21105/joss.01504

Software

- Review
- Repository
- Archive

Submitted: 03 June 2019
Published: 18 July 2019

Summary

Direct visualization, summary statistics (i.e., minimum, maximum, mean, standard deviation), ARIMA models, anomaly detection, forecasting, clustering, and deep learning are all popular techniques for analyzing and understanding time series data. However, the simplest and most intuitive approach of comparing all of the pairwise distances between each subsequence within a time series (i.e., a self-similarity join) has not seen much progress due to its inherent computational complexities. For a time series with length n and a subsequence comparison length m, the brute force self-similarity join for this sqeuence would have a computational complexity of $O(n^2m)$. To put this into perspective, assuming that each distance calculation took 0.0000001 seconds, a time series of length $n = 100{,}000{,}000$ would require roughly 1,585.49 years to compute all of the pairwise distances in a brute force manner. The ability to accurately and efficiently compute the exact similarity join would enable, amongst other things, time series motif and time series discord discovery. While approximate methods exist, they are often inexact, lead to false positives or false dismissals, and do not generalize well to other time series data. Novel research for computing the exact similarity join has significantly improved the scalability for exploring larger datasets without compromise.

Leveraging this work, we present STUMPY, a powerful and scalable library that efficiently computes something called the matrix profile (a vector that represents the distances between all subsequences within a time series and their nearest neighbors) (Yeh et al., 2016), (Zhu et al., 2016), which can be used for a variety of time series data mining tasks such as:

- pattern/motif (approximately repeated subsequences within a longer time series) discovery
- anomaly/novelty (discord) discovery
- shapelet discovery
- semantic segmentation
- density estimation
- time series chains (temporally ordered set of subsequence patterns)
- and more ...

The library also includes support for parallel and distributed computation, multi-dimensional motif discovery (Yeh, Kavantzas, & Keogh, 2017), and time series chains (Zhu, Imamura, Nikovski, & Keogh, 2017). Whether you are an academic, data scientist, software developer,

Figure 1: STUMPY Logo

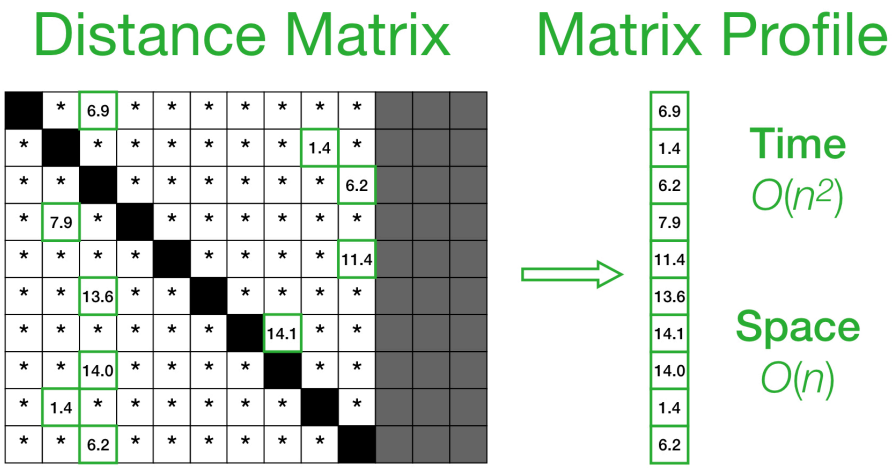

Figure 2: Matrix Profile

or time series enthusiast, STUMPY is straightforward to install and allows you to compute the matrix profile in the most efficient way. The goal of STUMPY is to allow you to get to your time series insights faster.

References

Yeh, M. C.-C., Kavantzas, N., & Keogh, E. (2017). Matrix Profile VI: Meaningful Multidimensional Motif Discovery. In *International Conference on Data Mining (ICDM)* (pp. 565–574). IEEE. doi:10.1109/ICDM.2017.66

Yeh, M. C.-C., Zhu, Y., Ulanova, L., Nurjahan, B., Ding, Y., Dau, H. A., Silva, D. F., et al. (2016). Matrix Profile I: All Pairs Similarity Joins for Time Series: A Unifying View That Includes Motifs, Discords and Shapelets. In *International Conference on Data Mining (ICDM)* (pp. 1317–1322). IEEE. doi:10.1109/ICDM.2016.0179

Zhu, Y., Imamura, M., Nikovski, D., & Keogh, E. (2017). Matrix Profile VII: Time Series Chains: A New Primitive for Time Series Data Mining. In *International Conference on Data Mining (ICDM)* (pp. 695–704). IEEE. doi:10.1109/ICDM.2017.79

Zhu, Y., Zimmerman, Z., Senobari, N. S., Yeh, M. C.-C., Funning, G., Mueen, A., Brisk, P., et al. (2016). Matrix Profile II: Exploiting a Novel Algorithm and GPUs to Break the One Hundred Million Barrier for Time Series Motifs and Joins. In *International Conference on Data Mining (ICDM)* (pp. 739–748). IEEE. doi:10.1109/ICDM.2016.0085