Graph Representation for Multivariate Time Series Analytics

Paul Boniol, Inria, ENS, PSL University
boniol.paul@gmail.com

Context:

Massive collections of time-varying data (i.e., *time series*, or *data series* in general) are becoming a reality in virtually every scientific and social domain. Examples of fields that involve data series include finance, environmental sciences, astrophysics, neuroscience, engineering, and multimedia. What is challenging in these data is that they are mainly highly multivariate, and also, the different dimensions that compose these data may originate from different sources.

However, this high number of dimensions from different sources causes severe limitations. First, existing solutions employ one model per dimension or data type. This implies (i) a drop in accuracy because of missed correlations among important dimensions, (ii) a significant increase in execution time, because of all the independent models that are used, and (iii) a drop in interpretability, because of the multitude of embedding produced by all independent models. To reach efficient and scalable analysis without sacrificing accuracy, we need a unified data embedding that can enable multiple analytic tasks (such as anomaly detection, classification, and clustering) on multivariate and heterogeneous data series.

The objective is to move towards a unified data embedding that allows multiple analytic tasks (such as anomaly detection, classification, and clustering) on multivariate and heterogeneous data. Towards that direction, we proposed in past research Series2graph (illustrated in Figure 1), a method that summarizes univariate time series into a graph [1,2]. Even though the latter method has been proposed mainly for anomaly detection, similar graph embedding for time series has demonstrated state-of-the-art and scalable results for tasks such as clustering [3], classification, and representation learning [4]. The benefit of such time series graph representation is three-fold. (i) First, such graph representation is easy to interpret by any user. (ii) Second, it can benefit from other graph-represented data (such as ontologies and knowledge graphs and textual data represented as graphs [5]). (iii) Last, one unified embedding can significantly reduce the analysis execution time (as shown for anomaly detection [1]).

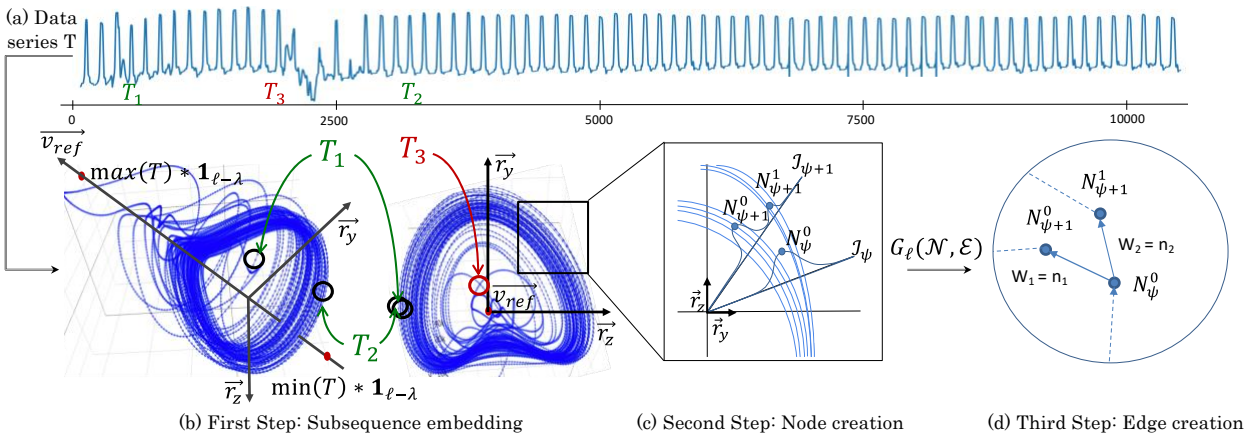

Figure 1: Series2Graph methods

Research Problem:

However, there exists no method that proposes a unified graph embedding for multivariate time series. The straightforward solution would be to build one graph embedding per dimension and then analyze them all together. However, the graph size would be linearly proportional to the number of dimensions, making it impossible to use in practice. In the case of heterogeneous multivariate time series, there exists no holistic graph representation, and we need novel approaches to address this problem.

Therefore, the objective of this internship is to propose new meaningful graph representation and transformation for multivariate time series, that could support basic analytics (classification, clustering, and anomaly detection).

Tasks:

During this internship, the student will do the following tasks:

- Acquire an exhaustive understanding of the literature on graph representation for time series, and graph-based methods for time series analytics.
- Propose and implement a new graph representation for multivariate time series
- Evaluate the proposed solution on publicly available benchmarks (UCR-Archive for classification and clustering, and equivalents of TSB-UAD [6] for multivariate time series).
- Compare the proposed solution to existing (non-graph-based) state-of-the-art methods.

Required skills:

- M2 in Data science, Computer Science
- Strong analytical and programming (Python) skills

Team and Location

This 5 months internship will take place within the computer science department at Ecole Normale Supérieure (45 rue d'Ulm, Paris 5), a member of PSL University, in the Valda team (led by Prof. Pierre Senellart) and supervised by Paul Boniol (Inria researcher at VALDA). We are interested in candidates considering the possibility of doing a PhD in the team after the internship.

If interested, please send your application to boniol.paul@gmail.com

References:

1. Boniol, P., and Palpanas, T. **Series2graph: Graph-based subsequence anomaly detection for time series**. Proc. VLDB Endow. 13, 12 (July 2020), 1821–1834.
2. Schneider, J., Wenig, P., and Papenbrock, T. **Distributed detection of sequential anomalies in univariate time series**. The VLDB Journal 30, 4 (2021), 579–602.
3. Tiano, D., Bonifati, A., and Ng, R. **Featts: Feature-based time series clustering**. In Proceedings of the 2021 International Conference on Management of Data (New York, NY, USA, 2021), SIGMOD '21, Association for Computing Machinery, p. 2784–2788.
4. Heng, Z., Yang, Y., Jiang, S., Hu, W., Ying, Z., Chai, Z., and Wang, C. **Time2graph+: Bridging time series and graph representation learning via multiple attentions**. IEEE Transactions on Knowledge and Data Engineering (2021), 1–1
5. Boniol, P., Panagopoulos, G., Xypolopoulos, C., Hamdani, R. E., Amariles, D. R., and Vazirgiannis, M. **Performance in the courtroom: Automated processing and visualization of appeal court decisions in France**. NLLP workshop of the KDD Conference (2020).
6. Paparrizos, J., Kang, Y., Boniol, P., Tsay, R. S., Palpanas, T., and Franklin, M. J. **Tsb-uad: an end-to-end benchmark suite for univariate time-series anomaly detection**. Proceedings of the VLDB Endowment 15, 8 (2022), 1697–1711.
7. Hoang Anh Dau and Anthony J. Bagnall and Kaveh Kamgar and Chin-Chia Michael Yeh and Yan Zhu and Shaghayegh Gharghabi and Chotirat Ratanamahatana and Eamonn J. Keogh: **The UCR time series archive**. IEEE/CAA Journal of Automatica Sinica. 2019