Offre n°2024-07668

PhD Position F/M [Campagne Doctorant] Graph Representation for Multivariate Time Series Analytics

Le descriptif de l'offre ci-dessous est en Anglais

Type de contrat : CDD
Niveau de diplôme exigé : Bac + 5 ou équivalent
Fonction : Doctorant

Contexte et atouts du poste

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 about these data is that they are mainly highly multivariate, and the different dimensions that compose them 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, 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]).

However, no method exists 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, no holistic graph representation exists, and we need novel approaches to address this problem.

References:


Supervision:

The thesis will be co-supervised by Paul Boniol (Valda team, DI ENS & Inria Paris) and Michaël Thomazo (Valda team, DI ENS & Inria Paris). The PhD student will be part of the Valda team within the Computer Science Department of the Ecole normale supérieure. Registration for the thesis will be carried out at Université PSL, via Ecole doctorale 386 (Sciences mathématiques de Paris Centre). The doctoral student will benefit from the environment and resources of the VALDA team, the DIENS, the Inria Paris Center,
and the PRAIRIE Institute, including local computing clusters. In addition, the PhD student will have access to the IDRIS Jean Zay supercomputer for GPU-intensive tasks.

**Mission confiée**

**Research objective**

The objective of this Ph.D. is to propose new meaningful graph representation and transformation for multivariate time series that could support basic analytics (classification, clustering, and anomaly detection). Overall, the research questions tackled are the following:

- Can a unique graph embedding method be more accurate on multiple analytical tasks than specific methods for each task?
- Can a unique graph embedding be constructed for large heterogeneous multivariate time series that preserves accuracy while remaining scalable?
- How can such embedding be used to interpret and explain analytical tasks (e.g., classification, clustering, anomaly detection)?

**Application and use cases:**

Time series analysis is a very important task for electricity production relevant applications. Indeed, the desire to analyze a large quantity of data efficiently and be able to express complex queries (i.e., anomalies discovery) can be crucial for industrial actors like EDF. For instance, one crucial goal for EDF is to improve the safety and availability of its electrical power plants by detecting anomalies that could occur. As massive gains are expected from reducing maintenance volumes, there is thus a serious need to have accurate and efficient algorithms to detect anomalies and understand their origins. Moreover, EDF has collected sensor data in every nuclear power plant for decades (at least 20 years). With a total of 58 nuclear power plants and more than 2000 sensors per unit, it represents a database of approximately 500 TeraBytes. Considering that the electrical power plants were built 20 to 30 years ago, we can expect that recent maintenance and new power plants will see their number of sensors and acquisition rate significantly increase, resulting in an exponential increase of new data. Moreover, half of the 2000 EDF electrical power plant sensors are boolean sensors, and the remaining half measure either water flow, pressure, temperature, or water level from very different parts of the plant, making each sensor almost unique. In addition to these already highly heterogeneous data series, the EDF database contains multiple logs (i.e., textual data) and structural knowledge (i.e., knowledge graphs representing the structure of the plant and the link between sensors). The context described above is highly related to the problems that will be tackled in this PhD. Thus, benefiting from previous collaborations of Paul Boniol with the research department of EDF, the PhD candidates might have the opportunity to apply the research conducted on such use cases.

**Principales activités**

**Main 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 for multivariate time series).
- Study the impact of heterogeneous (i.e., various acquisition rates, types of time series, stationarity, etc.) multivariate time series on unified graph embedding.
- Propose interpretable solutions based on the graph representation of time series for specific analytical tasks.

**Additional tasks:**

- Write scientific research papers with the objective to publish them on top data analytics and data management conferences and journals.

**Avantages**

- Subsidized meals
- Partial reimbursement of public transport costs
- Leave: 7 weeks of annual leave + 10 extra days off due to RTT (statutory reduction in working hours) + possibility of exceptional leave (sick children, moving home, etc.)
- Possibility of teleworking
- Flexible organization of working hours (after 12 months of employment)
- Professional equipment available (videoconferencing, loan of computer equipment, etc.)
- Social, cultural and sports events and activities
- Access to vocational training

**Rémunération**

Monthly gross salary: 2100 € during the first and second years. 2190 € the last year.
Informations générales

- Thème/Domaine : Représentation et traitement des données et des connaissances
- Ville : Paris
- Centre Inria : Centre Inria de Paris
- Date de prise de fonction souhaitée : 2024-10-01
- Durée de contrat : 3 ans
- Date limite pour postuler : 2024-05-19

Contacts

- Équipe Inria : VALDA
- Directeur de thèse : Boniol Paul / paul.boniol@inria.fr

A propos d'Inria

Inria est l'institut national de recherche dédié aux sciences et technologies du numérique. Il emploie 2600 personnes. Ses 215 équipes-projets agiles, en général communes avec des partenaires académiques, impliquent plus de 3900 scientifiques pour relever les défis du numérique, souvent à l'interface d'autres disciplines. L'institut fait appel à de nombreux talents dans plus d'une quarantaine de métiers différents. 900 personnels d'appui à la recherche et à l'innovation contribuent à faire émerger et grandir des projets scientifiques ou entrepreneuriaux qui impactent le monde. Inria travaille avec de nombreuses entreprises et a accompagné la création de plus de 200 start-up. L'institut s'efforce ainsi de répondre aux enjeux de la transformation numérique de la science, de la société et de l'économie.

L'essentiel pour réussir

The doctoral student must have obtained a Master's degree or equivalent in computer science or mathematics. He or she should have had courses and initial research experience in one of the following fields: artificial intelligence, data management, statistical learning, information retrieval. He or she should be comfortable with large-scale data processing and the use of modern artificial intelligence techniques, particularly deep learning. He or she should be able to read and write research articles in English.

Attention : Les candidatures doivent être déposées en ligne sur le site Inria. Le traitement des candidatures adressées par d'autres canaux n'est pas garanti.

Consignes pour postuler

To apply, please include in your application:

- Covering letter highlighting the relevance of the candidate's training to the topic.
- CV
- Master's grades

Sécurité défense :

Ce poste est susceptible d'être affecté dans une zone à régime restrictif (ZRR), telle que définie dans le décret n°2011-1425 relatif à la protection du potentiel scientifique et technique de la nation (PPST). L'autorisation d'accès à une zone est délivrée par le chef d'établissement, après avis ministériel favorable, tel que défini dans l'arrêté du 03 juillet 2012, relatif à la PPST. Un avis ministériel défavorable pour un poste affecté dans une ZRR aurait pour conséquence l'annulation du recrutement.

Politique de recrutement : 

Dans le cadre de sa politique diversité, tous les postes Inria sont accessibles aux personnes en situation de handicap.