EDF R&D – Département PRISME A. Ajenjo, E. Remy, P. Stephan

Sujet de stage de fin d'études / M2 – Année universitaire 2024/2025
Détection non-supervisée d'anomalies dans les séries temporelles : revue et comparaison d'algorithmes

MOTS-CLÉS :

Apprentissage non-supervisé, machine learning, anomalies, séries temporelles, surveillance en fonctionnement, capteurs – instrumentation

CONTEXTE GÉNÉRAL ET PROBLÉMATIQUE INDUSTRIELLE :

Avec le développement des technologies numériques (dont l'internet des objets et les architectures de données Big Data) et l'explosion des capacités de stockage des données informatiques, **les larges collections de séries temporelles deviennent une réalité dans un grand nombre de domaines**, comme la finance, les sciences de l'environnement, la médecine, les métiers du numérique, l'ingénierie ou l'industrie. Il y a donc un intérêt et un besoin croissants de développer des techniques efficaces pour traiter et analyser ce type de données.

Une série temporelle est une séquence ordonnée dans le temps de points ou de valeurs, par exemple des mesures à différents instants d'un paramètre physique (débit, pression…) issues d'un capteur de surveillance installé sur un système industriel. Une fois une série collectée, enregistrée, nettoyée (débruitage, synchronisation, complétion des données manquantes…) et mise à disposition de l'utilisateur, **celui-ci souhaite généralement l'étudier pour en extraire de la connaissance et de la valeur**. Cette analyse peut être simple, comme sélectionner une fenêtre temporelle pour visualiser la série et calculer des statistiques sur les valeurs afin de résumer l'information (valeur moyenne par exemple). Elle peut aussi être complexe, comme rechercher des similarités entre plusieurs séries temporelles pour réaliser des regroupements, prévoir les prochaines valeurs de la série à partir de l'historique des mesures, identifier des motifs récurrents, ou détecter des anomalies ou des ruptures associées à des changements de régime dans les valeurs de la série, synonymes d'évolutions soudaines et inhabituelles possiblement non souhaitées.

Dans le contexte de la surveillance et de la maintenance des matériels industriels, comme ceux qui équipent les installations de production d'électricité d'EDF (turbines, pompes…), **la détection d'anomalies dans les séries temporelles représente un enjeu crucial** : plus elle est précoce et rapide, plus on est en mesure de réagir tôt pour tenter d'atténuer les impacts, voire d'éviter la survenue, de tout événement potentiellement critique, comme un dysfonctionnement ou une défaillance d'un système. Ce même objectif est visé dans un grand nombre d'autres domaines, comme en sismologie, en volcanologie, ou en sécurité informatique (cyberattaque) ou bancaire (fraude massive).

Parmi les différentes familles de méthodes de détection d'anomalies dans les séries temporelles, **les algorithmes non-supervisés**, qui n'ont pas besoin d'exemples caractérisant précisément ce qu'est une anomalie et ce qu'est un comportement normal (ou sain) des matériels, **sont particulièrement attractifs d'un point de vue industriel**. En effet, même s'ils peuvent être moins efficaces que les approches supervisées ou semi-supervisées, **ils ne requièrent pas de travail d'annotation des sous-séquences** (souvent chronophage et coûteux, voire impossible dans certains cas) **et sont bien adaptés à la découverte d'anomalies nouvelles encore non observées**.

OBJECTIFS DU STAGE ET PLANNING PRÉVISIONNEL :

L'analyse des séries temporelles fait l'objet de travaux de recherche intensifs, le nombre annuel de publications sur le sujet référencées sur la base Scopus ayant été multiplié par plus de 100 entre 1985 et 2020. Il est donc essentiel d'assurer une **veille bibliographique** continue et **d'identifier, parmi toute la production scientifique, les développements les plus prometteurs, pour ensuite les tester et évaluer leurs performances sur des jeux de données de référence**. Le stage s'inscrit dans cet objectif de veille active et de benchmark.

À cette fin, il s'articulera selon plusieurs phases :

1) **Recherche de bases de données annotées (simulées ou réelles)** mises à disposition de la communauté, comme Exathlon, ODDS ou TSB-UAD.
2) **Revue de littérature des algorithmes de détection non-supervisée d'anomalies dans les séries temporelles** (**préférentiellement multivariées**, c'est-à-dire sur plusieurs dimensions, comme des mesures issues de multiples capteurs installés sur différents équipements) et construction d'une taxonomie tenant compte, entre autres, des deux critères suivants : principe de l'approche (par densité, par distance, par reconstruction…) et capacité à gérer des flux continus.
3) **Récupération des codes existants et/ou implémentation informatique des algorithmes identifiés comme à fort potentiel** suite à l'état de l'art réalisé à l'étape 2), puis mise en œuvre et évaluation de leurs performances à partir des données annotées recensées lors de la phase 1).
4) Comparaison, analyse critique et synthèse des résultats obtenus.
5) Rédaction du rapport de stage.

À noter qu'en perspective un **sujet de thèse pourrait être ouvert à la suite du stage**. Il élargira son périmètre dans l'objectif de lever les limites des algorithmes actuels, en particulier dans un **objectif de détection non-supervisée d'anomalies dans des flux continus de séries temporelles multivariées**.

PROFILS :

Étudiant.e de M2 (mathématiques appliquées / probabilités-statistiques-machine learning) ou d'écoles d'ingénieur.e.s (généralistes avec majeure en mathématiques appliquées / probabilités-statistiques-machine learning).

COMPÉTENCES REQUISES :

- Solides compétences en probabilités, statistiques, machine learning et analyse numérique
- Goût pour la programmation (Python, R)
- Aisance dans la communication, orale et écrite, en français et/ou en anglais

APTITUDES PERSONNELLES SOUHAITÉES :

- Goût pour la recherche (méthodologies, concepts mathématiques, applications industrielles)
- Ouverture d'esprit, curiosité et capacité à être autonome

CONTACTS (ENCADRANTS INDUSTRIELS) :

Antoine Ajenjo (antoine.ajenjo@edf.fr), Emmanuel Remy (emmanuel.remy@edf.fr) et Pierre Stephan (pierre.stephan@edf.fr)

CONTACT (ENCADRANT ACADÉMIQUE) :

Paul Boniol (boniol.paul@gmail.com)

DURÉE ENVISAGÉE :

6 mois à compter de septembre 2024

LIEU :

EDF R&D – Lab Chatou
Département PRISME
Groupes P13 « Modèles Système, Surveillance et Qualité de la Mesure » et P17 « Gestion d'Actifs, Incertitudes et Apprentissage Statistique »
6 quai Watier, 78401 Chatou, France