

Stage et thèse
“Apprentissage de modèles à base de règles pour
la prédiction de séries temporelles”

Aurore Lomet, Jean-Philippe Poli

September 2021



1 Objectif

Apprendre une base de règles pour la prédiction de séries temporelles

2 Résumé

La réduction des coûts liés à l'instrumentation et au stockage permet de disposer aujourd'hui de nombreuses bases de données permettant la mise en œuvre de modèles prédictifs (c'est-à-dire capables de prédire des événements). Le traitement des séries temporelles a fait l'objet de nombreux travaux dans ces différents domaines de l'IA. Nous nous intéressons ici aux algorithmes d'IA symbolique.

Contrairement aux réseaux de neurones, les algorithmes d'intelligence artificielle symbolique permettent d'exploiter des connaissances et de simuler un raisonnement humain. Ces algorithmes permettent d'obtenir des modèles interprétables, c'est-à-dire facilement compréhensibles par un humain. Dans certains domaines dans lesquels l'humain est le sujet principal, comme en médecine ou dans la sécurité, de tels modèles sont nécessaires pour obtenir une IA de confiance. Dans ce cadre, le CEA List développe sa propre plate-forme

logicielle ExpressIF® (<https://expressif.cea.fr>) qui est capable d'exploiter des connaissances humaines, d'extraire automatiquement des connaissances depuis des données, et de justifier ses décisions.

Les méthodes de l'état de l'art proposent souvent des cas d'étude simplifiés. D'une part, l'analyse des variables prédictives et à prédire repose souvent sur une analyse des corrélations : la corrélation n'étant pas systématiquement une causalité, le lien peut être quelque peu erroné. D'autre part, la temporalité est souvent simplifiée : on considère souvent les valeurs des variables aux mêmes temps.

Pourtant, de nombreuses applications ne correspondent pas à ces simplifications. A titre d'exemple :

- Dans le domaine du manufacturing, on cherche souvent à prédire la défaillance d'un produit par rapport aux différentes étapes du procédé. Or la défaillance peut être causée par des événements ayant eu lieu à différents moments du procédé.
- Dans le domaine de la cybersécurité (et plus généralement en sécurité), une attaque peut être détectée par différentes tentatives ou différents événements au fil du temps.
- Dans le cadre de certains capteurs chimiques, les gaz sont transportés par fluide et présentés à différents capteurs. Les réponses de ces capteurs sont donc décalées dans le temps.

Dans ce contexte, cette thèse a pour objectif de développer une solution innovante pour l'apprentissage d'une base de règles pour la prédiction de séries temporelles. Pour ce faire, il est nécessaire dans un premier temps d'avoir une méthode automatique pour déduire les relations de causalité entre les séries temporelles. Dans un second temps, il sera nécessaire d'appliquer ou adapter ou développer des méthodes d'extraction de vocabulaire pour la définition des règles à partir des relations causales établies. De nouveaux opérateurs spécifiques aux séries temporelles pourront être définis et enrichir la plateforme ExpressIF®. Ainsi, ce travail devra prendre en compte la finalité d'apprentissage de la base de règles. Plusieurs enjeux seront donc étudiés dans cette thèse :

1. l'ordre temporel devra être pris en compte pour définir les relations causales. On cherchera à développer ou adapter une méthode pour des relations non linéaires et non gaussiennes qui sont les plus fréquentes dans les données réelles. La causalité au sens de Granger et ses extensions pour le cas non linéaire ne pourront pas être directement appliquées du fait qu'elles peuvent exprimer une corrélation avec antécédent. En effet, ces méthodes reposent sur le principe d'amélioration de la prévision d'une série par l'intégration d'information exogène qui est possible avec une corrélation. Dans le but de déduire des bases de règles, l'objectif de ce travail est d'obtenir des relations causales. Ainsi, les graphes causaux bayésiens acycliques peuvent apporter une solution en les adaptant pour les séries temporelles. Du fait de la complexité de ces méthodes, il sera alors nécessaire de proposer une méthode adaptée pour les jeux de données de grande dimension comme dans le domaine du monitoring, par

exemple.

2. En lien avec le point précédant, une discrétisation des séries pourra être envisagée pour définir les relations causales suivant les valeurs et les instants des séries. Cette discrétisation pourra être mise en relation avec l'extraction de vocabulaire des règles de décision. Par exemple, on considère les variables X_1, X_2 et Y . On souhaite obtenir une règle telle que : si X_1 est grand à l'instant $t - k$ et X_2 est petit à l'instant $t - l$ alors Y est solide à l'instant t . A partir de cet exemple, on peut noter que les règles pourront considérer des décalages temporels différents suivant les causes.
3. Enfin, les effets joints des causes seront également considérés. Il existe plusieurs types d'effets joints : additifs, multiplicatifs... Ces relations devront être prises en compte dans l'établissement des règles.

La thèse sera précédée d'un stage de MASTER2. Dans ce contexte, le premier objectif du stage est d'étudier les différents tests possibles pour l'indépendance/l'influence de séries temporelles (notamment les tests d'indépendance basés sur l'entropie de transfert). Cette étude considérera également les indépendances conditionnelles pour tester une catégorie des effets joints des causes. D'autres types d'effets joints pourront être considérés. On se concentrera également sur la discrétisation des séries temporelles pour tester l'influence par morceau. Enfin, on s'intéressera à la mise en application de réseaux bayésiens pour la causalité existants dans la littérature.

3 Environnement

Situé à Saclay, en Ile-de-France sud, le CEA LIST est un institut de recherche scientifique et technologique dédié au développement de logiciels, de systèmes embarqués et de capteurs pour des applications destinées à la défense, la sécurité, l'énergie, le nucléaire, l'environnement et la santé. Le CEA LIST compte plus de 700 chercheurs travaillant sur les systèmes numériques intelligents, centrés autour de l'intelligence artificielle, l'usine du futur, l'instrumentation innovante, les systèmes cyberphysiques et la santé numérique. Au sein de cet institut, les laboratoires du Service d'Intelligence des Données (SID) travaillent au développement et au transfert industriel de technologies de pointe en IA. Le périmètre technique de nos 50 ingénieurs et chercheurs concerne l'analyse des signaux (c.à.d. des séries temporelles, mais aussi des spectres) produits par les équipements développés en interne, par les équipes du CEA, ou par des sociétés externes. L'exploitation de ces données repose sur un large spectre de méthodes d'apprentissage machine, relevant de l'IA numérique (réseaux de neurones profonds, forêts aléatoires, SVM) et de l'IA symbolique (systèmes à base de règles).

4 Profil de la candidate ou du candidat

MASTER 2 en IA, datascience, statistique, ingénierie statistique.

La capacité à travailler en équipe est nécessaire, tout en faisant preuve d'autonomie dans les tâches au quotidien. Les développements en IA avançant rapidement, il est nécessaire que la candidate ou le candidat ait la capacité de renouveler et enrichir ses compétences de manière continue. Enfin, d'excellentes capacités d'expression, orale et écrite, en anglais, vous seront nécessaires pour échanger avec les nombreux partenaires du projet.

Sur le plan scientifique et technique, vous pouvez justifier :

- de solides compétences en statistique (traitement de séries temporelles, signal) et en machine learning (data science),
- des connaissances et/ou une expérience en étude de causalité sont un plus,
- des connaissances en IA symbolique, en particulier (mais non obligatoire)
- d'une expérience significative en développement logiciel, s'appuyant idéalement sur :
 - La maîtrise d'un langage de prototypage (idéalement Python 3 ou R),
 - L'utilisation d'outils de développement et de partage de code tels que Git ou SVN sera également appréciée.

5 Dates

- Démarrage souhaité du stage : avant la fin du 1er semestre 2022
- Démarrage souhaité de la thèse : dernier trimestre 2022

6 Contact :

Merci d'envoyer vos CV et lettre de motivation à :

- aurore.lomet@cea.fr
- jean-philippe.poli@cea.fr