Sujet de Master 2 Recherche

Modèles graphiques probabilistes pour la détection d'anomalie

Introduction

L'équipe DUKe (Data User Knowledge) du LS2N, UMR CNRS 6004, est l'une des principales équipes du laboratoire dans le thème « science des données et de la décision », forte de ses compétences en manipulation de données, en fouille de données et en interaction. Dans ce cadre, l'équipe a développé de nombreux algorithmes d'apprentissage et de manipulation de modèles graphiques probabilistes (réseaux bayésiens, réseaux bayésiens dynamiques, réseaux bayésiens relationnels).
L'équipe DUKe travaille en collaboration avec Talend, leader mondial des solutions d'intégration big data et cloud, sur l'utilisation de modèles graphiques pour détecter et corriger des anomalies dans les données.
Nous avons ainsi proposé une approche centrée autour de l'apprentissage de réseaux bayésiens permettant de découvrir automatiquement des anomalies dans des données tabulaires mixtes (discrètes et continues) [1].

Objectif du stage

Nous avons proposé une architecture basée sur l'utilisation de réseaux bayésiens pour l'apprentissage de dépendances probabilistes et la prise en compte de dépendances fonctionnelles, et l'identification de valeurs anormales dans un jeu de données. L'objectif du stage est d'étendre l'architecture réalisée dans un contexte incrémental, où les données peuvent arriver par lot, où des variables peuvent être ajoutées et/ou enlevées, et où des propositions de correction d'anomalies par l'utilisateur peuvent faire évoluer le modèle existant.

Travail à réaliser

- Familiarisation avec le formalisme des réseaux bayésiens et avec la librairie C++ PILGRIM General
- Etude de la solution existante proposée dans [1]
- Etude de la prise en compte de corrections d'anomalies par l'utilisateur pour faire évoluer le modèle.
- Implémentation de l'approche avec la librairie PILGRIM
- Illustration des résultats obtenus sur des cas d'utilisations fournis par l'entreprise
- Prise en compte ensuite de l'aspect incrémental où des données peuvent arriver par lor., avec le même cycle : étude, implémentation et tests.

Ce travail sera supervisé par P. Leray (LS2N / DUKe, Nantes). Le stagiaire sera intégré à une équipe de plusieurs stagiaires, doctorants et ingénieur travaillant sur les réseaux bayésiens et PILGRIM, et sera régulièrement en contact avec les experts de Talend.

Période : Février-Juillet 2021
Indemnité de stage : approx. 554 € / mois
Compétences

- Concepts de probabilité, statistiques
- Programmation C++

Candidature

CV + lettre de motivation + résultats académiques (format PDF) à philippe.leray@univ-nantes.fr

Références

[1] Evan Dufraisse, Philippe Leray, Raphaël Nedellec and Tarek Benkhelif. Interactive Anomaly Detection in Mixed tabular Data using Bayesian Networks, The 10th International Conference on Probabilistic Graphical Models, Aalborg, September 23-25, 2020