Proposition de thèse CIFRE
Réalisée par le laboratoire I3S et l'entreprise Himydata
2022-2025

Intitulé de la thèse : Nettoyage automatisé de données avec des réseaux de neurones profonds

Résumé du projet de thèse :

L'industrie 4.0 s'appuie sur la récupération et l'analyse de données provenant de nombreux capteurs et machines. La détection automatique non-supervisée d'anomalies dans ces données, suivie d'un nettoyage adéquat, est essentiel pour garantir le développement de cette nouvelle industrie. Ce projet de thèse souhaite développer un réseau de neurones profond pour effectuer ces détections et le nettoyage associé. L'architecture de ce réseau doit permettre de comprendre pourquoi le réseau nettoie, et donc modifie, une donnée. De cette façon, ce nettoyage automatisé sera explicable pour les utilisateurs impliqués.

La thèse sera réalisée le laboratoire I3S rattaché à l'Université Côte d'Azur et au CNRS et l'entreprise Himydata.

Contexte de la thèse :

Le passage d'une industrie dite 3.0 à une industrie 4.0 se fait par la récupération de données en masse concernant les machines à l'aide de capteurs (positionnés à chaque étape du processus de production) qui sont les éléments de base des systèmes de contrôle et d'acquisition de données en temps réel. Cette quatrième révolution industrielle se caractérise aussi par la connexion des objets (ou machines) entre eux. Le but est d'obtenir une nouvelle génération d'usines connectées, robotisées et intelligentes afin de pouvoir surveiller sa production et donc d'améliorer la qualité et le temps de création des produits ainsi que d'optimiser les procédés. Cela va permettre aussi de réduire les temps d'arrêt en étant averti au préalable de l'état des machines.

Le but de l'entreprise Himydata associée à cette thèse est de rendre toutes ces données accessibles, compréhensibles et analysables par le plus grand nombre. Ce sont des bases de données complexes avec des données provenant de sources hétérogènes, possédant de nombreux attributs. Les règles sur ces données sont inconnues ou, dans le meilleur cas, très peu connues. De plus, les données provenant du monde réel sont bruitées et souvent corrompues. Ces effets peuvent avoir un impact important sur la compréhension des données, leur modélisation et sur les prises de décisions qui en découlent [1,2]. Ainsi, l'étape cruciale dans l'utilisation des données est la détection et le nettoyage des erreurs dans les données. En effet, en identifiant et en nettoyant les données « sales », il est possible d'obtenir une plus grande compréhension des données, d'améliorer les modèles se servant de ces dernières ou encore de prendre de meilleures décisions.

Objectifs de la thèse :

Le principal objectif de cette thèse est de développer une méthode d'apprentissage non-supervisé qui analyse des données, notamment des séries temporelles, pour détecter des anomalies puis corriger ces anomalies. Dans ce but, le doctorant concevra un réseau de neurones profonds novateur.

L'apprentissage profond représente une approche prometteuse dans le sens où il permet de réaliser toutes les étapes nécessaires à la détection et au nettoyage. Les modèles génératifs sont un très bon moyen d'apprendre la distribution de données en modélisant les probabilités conditionnelles d'un jeu de données de manière non-supervisée. Le but d'un modèle génératif est d'apprendre la véritable distribution mais, comme cela n'est généralement pas faisable à cause de la malédiction de la dimension, son but va être de s'approcher au mieux de cette véritable distribution en optimisant un critère d'apprentissage. De nos jours, ce travail d'apprentissage de fonction est fait à l'aide des réseaux neuronaux profonds [3].

Un des points importants pour Himydata est l'explicabilité des résultats obtenus et, par conséquent, l'identification des mécanismes internes au modèle. Par conséquent, le modèle qui sera développer dans la thèse doit rester explicable. L'apprentissage d'un réseau profond dans le cas non-supervisé (lorsque les étiquettes sont totalement inconnues) conduit à représentation emmêlée [4]. Ceci signifie que les liens entre les variables au sein du réseau de neurones et les valeurs en entrée sont obscurs et inexplicables. Il s'agira donc de construire un réseau de neurones pertinent et de proposer un algorithme d'entrainement du réseau qui permet de créer des liens explicites entre les variables impliquées dans les représentations internes du réseau.

Principales tâches de la thèse :

Les principales tâches de la thèse seront les suivants :

- Rédaction d'un état de l'art détaillé et implémentation des méthodes qui seront privilégiées pendant la thèse. Les méthodes de l'état de l'art seront testées sur des données publiques et des données propres de l'entreprise.
- Conception d'un réseau de neurones profonds génératif sur des données propres uniquement. Il nous faudra incorporer dans ce modèle profond un modèle graphique bayésien qui permet de démêler les représentations apprises.
- Conception d'un algorithme basé sur la programmation probabiliste afin de pouvoir entrainer facilement et efficacement les réseaux de neurones proposés.
- Implémentation efficace des algorithmes proposées avec une analyse statistique détaillée des résultats obtenus sur plusieurs jeux de données.

Compétences souhaitées : Apprentissage automatique (machine learning), réseau de neurones, probabilités et statistiques, Python (ou langage équivalent), autonomie, curiosité, esprit de synthèse.

Le candidat devra être titulaire d'un M2 ou grade équivalent au moment du recrutement.

Comment candidater ?

Le dossier de candidature sera composé d'un CV détaillé, d'un relevé de notes (si possible de la 1ère année post-bac jusqu'au master/diplôme d'ingénieur), d'une lettre de motivation et des coordonnées d'au moins une personne référente (encadrant de stage, professeur...).

Contact :

- Lionel Fillatre, Université Côte d'Azur, Laboratoire I3S
  E-mail : lionel.fillatre@i3s.unice.fr

- David Bessoudo, Himydata
  E-mail : david@himydata.com

Références bibliographiques :

[1] Y. Hu, S. De, Y. Chen, S. Kambhampati. Bayesian Data Cleaning for Web Data, 2012.

[2] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning, 2nd ed. Springer-Verlag New York, 2009.

[3] S. Eduardo, A. Nazábal, C. K. I. Williams, C. Sutton. Robust variational autoencoders for outlier detection and repair of mixed-type data, 2020.

[4] N. Siddharth, Brooks Paige, J-W. van de Meent, A. Desmaison, N. D. Goodman, P. Kohli, F. Wood, and P. Torr. Learning Disentangled Representations with Semi-Supervised Deep Generative Models. In Advances in Neural Information Processing Systems. 5927–5937, 2017.