Avis de Soutenance

Monsieur David ING

Informatique et applications

Soutiendra publiquement ses travaux de thèse intitulés

De l'Apprentissage Automatique aux Humanités Numériques

dirigés par Monsieur Lakhdar SAÏS et Monsieur Said JABBOUR

Soutenance prévue le ***vendredi 13 décembre 2024*** à 14h00
Lieu : Faculté des Sciences Jean Perrin, rue Jean Souvraz SP18 62300 Lens Cedex
Salle : des thèses

Composition du jury proposé

M. Lakhdar SAÏS	Université d'Artois	Directeur de thèse
M. Nadjib LAZAAR	Université Paris Saclay	Rapporteur
M. Engelbert MEPHU NGUIFO	Université Clermont Auvergne	Rapporteur
M. Frédéric LARDEUX	Université d'Angers	Examinateur
M. Jabbour SAID	Université d'Artois	Co-directeur de thèse
Mme Nelly ROBIN	Université Paris Cité - IRD	Examinatrice
M. Fabien DELORME	Université d'Artois	Invité

Résumé :

La réduction de la dimensionalité ou la sélection de caractéristiques pertinentes est une étape importante dans la conception de modèles d'apprentissage automatique, ainsi que pour la génération d'explications concises et compréhensibles dans de nombreuses applications sensibles. Pour les algorithmes de classification, tels que les arbres de décision et les forêts aléatoires, de nombreux travaux se sont concentrés sur la construction d'arbres optimaux en termes de caractéristiques impliquées, ainsi que sur la génération d'explications les plus pertinentes et les plus justes. Cette thèse intitulée De l'apprentissage automatique aux humanités numériques propose des améliorations significatives aux algorithmes de classification, à la fois sur la sélection de caractéristiques via l'analyse logique des données (LAD), un cadre d'apprentissage de motifs qui combine optimisation, fonctions booléennes et la combinatoire, une approche issue des mathématiques discrètes souvent ignorée par la communauté IA. Nous avons également proposé une nouvelle mesure d'optimalité des arbres de décision, définie par le nombre minimal de caractéristiques nécessaires à leur construction. Nos expérimentations, sur une large base de tests, montrent une réduction sensible du nombre de caractéristiques nécessaires pour construire des arbres de décision et des forêts aléatoires, améliorant ainsi l'étape de génération d'explications. La seconde partie de cette thèse s'inscrit dans une tendance de recherche émergente, combinant Intelligence Artificielle (IA) et humanités numériques (HN). Notre première contribution utilise de manière originale des bases de données juridiques pour identifier les réseaux de traite d'êtres humains (HTN), impliquant à la fois des victimes d'abus sexuels et des exploiteurs. Nos modèles de classification améliorés sont utilisés non seulement pour déterminer la classe d'un réseau donné Non suspect, Suspect, ou Probablement suspect — mais aussi pour fournir des explications plus courtes qui pourraient aider le juge à prendre la bonne décision. Notre dernière contribution présente une première étape vers une approche fouille de texte visant à exploiter les récits de migrants, collectés lors d'entretiens avec des migrants sur les routes, dans deux langues, dont l'anglais et le français. Après un dialogue approfondi avec des experts, nous avons identifié les concepts essentiels du domaine, y compris le concept lieux, villes ou villages traversés par les migrants. Notre approche, basée sur la fouille de texte et le traitement automatique du langage naturel (TAL), extrait automatiquement ces termes liés aux lieux intégrés dans ces récits, en utilisant une adaptation d'un algorithme d'expansion d'ensemble de manière faiblement supervisée, avec un petit ensemble de termes annotés. Enfin, nous avons conçu un outil permettant de visualiser ces itinéraires sur une carte, facilitant ainsi l'observation des routes migratoires.