Projet InvolvD : Elicitation interactive de contraintes pour la fouille de données supervisée et semi-supervisée

Christel Vrain*†[1]

[1]LIFO, Université d'Orléans – Université d'Orléans – France

Résumé

Les recherches recentes en Apprentissage Automatique et Fouille de Donnees cherchent à automatiser le processus de decouverte de connaissances et à reduire les interactions avec l'expert avec de bonnes raisons comme la difficulte à traiter des volumes importants de donnees (d'autant plus en grandes dimensions) ainsi que les progrès techniques qui ont permis d'alleger les taches chronophages. Cela a conduit à l'emergence d'offres telles que AutoML service (Google) envisageable dans un cadre supervise ou les etiquettes des objets peuvent etre exploitees pour regler des paramètres ou selectionner des modèles. En revanche, si les donnees sont partiellement etiquetees (apprentissage semi-supervise) ou n'ont pas d'etiquette (non supervise), la demarche inverse est necessaire : mettre l'expert dans la boucle d'apprentissage et integrer ses retours sur les resultats pour ameliorer le processus, autrement dit rendre le processus interactif. Cela pose de nouveaux defis comme presenter les resultats pour permettre des retours informes de l'expert, etre capable de les expliquer, interagir frequemment avec l'utilisateur alors qu'AutoML a la possibilite de tourner pendant des heures. Resoudre ces defis non seulement ameliore les resultats mais offre un autre avantage : un utilisateur est plus enclin à accepter un resultat si le processus qui a conduit à son emergence est explique.Ceci est d'autant plus vrai dans des applications ou les investissements (en argent, temps, vies humaines) reposent sur la justesse des resultats. De plus, les reglementations recentes en Europe et aux Etats-Unis donnent des droits aux citoyens concernes par des decisions algorithmiques et imposent que les decisions soient expliquees. Ces exigences ont ainsi motive des recherches sur l'interpretabilite des methodes de type boite noire (e.g. apprentissage profond). Pour obtenir des resultats explicables en fouille de donnees non supervisee ou semi-supervisee, le projet InvolvD traite des questions posees par le developpement de processus interactif de fouille de donnees : identification automatique de visualisations faisant sens, explications pour des retours informes, transformation en contraintes operationnelles et developpement de nouveaux systèmes d'apprentissage integrant ces contraintes. A contrario d'approches de type boite noire, nous nous fonderons sur le clustering et la recherche de motifs symboliques. Le cas d'usage en chemo- informatique, qui servira de guide tout au long du projet, est un cas typique d'illustration de cette problematique. En conception de medicaments, l'analyse exploratoire de donnees est capitale : les molecules doivent etre comprises en termes de structures et/ou de proprietes chimiques, et les experts ont des connaissances qu'ils ne peuvent expliciter qu'au vu de resultats preliminaires.

*Intervenant

†Auteur correspondant: Christel.Vrain@univ-orleans.fr