Apprentissage automatique pour une aide au codage PMSI

CDS 2.0 2016 project

project acronym	Infomed
principal investigator	Namik Taright
contact email	Namik.taright@aphp..fr
other team members	Rémi Flicoteaux
contracting partner	Polytechnique
principal team/laboratory	APHP – INSERM UMRS 1153
partner teams/laboratories	
type of support	Data challenges and RAMPs ou Standalone research engineers
amount of support	
parallel submissions	ANR 2017

Summary

Contexte Depuis la mise en œuvre effective de la Tarification à l'Activité (T2A), les modalités d'allocation de ressources des hôpitaux s'articulent autour d'un remboursement forfaitaire de chaque séjour par l'assurance maladie. Son attribution repose sur le traitement d'informations normalisées transmises par les établissements à l'issue de chaque séjour hospitalier, c'est le programme de médicalisation des systèmes d'information (PMSI). La constitution de ce recueil demande aux établissements la mise en place de procédures organisationnelles complexes en raison du haut niveau de qualité attendu. Des méthodes d'apprentissage machine supervisé à partir de données textuelles des comptes rendu d'hospitalisation ont été proposées pour aider au codage [Perrote2014,Dermouche2016]. En pratique cependant, il n'existe pas aujourd'hui de modélisation fiable permettant une extraction de connaissances au format des classifications utilisées pour le codage de l'activité (CIM-10 pour les pathologies et CCAM pour les actes) à partir des dossiers médicaux. Il s'agit ici d'une tâche d'apprentissage supervisé multilabel (en moyenne 4 à 5 codes par document) dans un espace de prédiction de grande dimension (environ 10 000 codes possibles).

Objectifs Nous souhaitons implémenter des méthodes intégrant les interactions hommes machine en tant qu'élément entrant à part entière dans le processus d'apprentissage automatisé. Ces méthodes d'apprentissage interactive (« human in the loop »), offrent de nouvelles perspectives d'utilisation des algorithmes du machine learning visant 1) à rendre plus lisible pour l'utilisateur les déterminants du choix proposés par la machine (Fairness, Accountability, and Transparency in Machine Learning), 2) à intégrer l'utilisateur dans le processus d'aide à la décision.

Positionnement Pour atteindre cet objectif il est nécessaire de construire des algorithmes et des interfaces qui donneront à l'expert l'opportunité d'intervenir/interagir/agir dans le processus d'apprentissage pour l'orienter dans le sens qu'il souhaite. Plusieurs méthodes algorithmiques peuvent être développées. Les méthodes d'optimisation bayésiennes peuvent être utilisées dans cette perspective pour de nombreux algorithmes d'apprentissage où les expériences des utilisateurs sont utilisées pour adapter les paramètres des modèles de prédiction [Shahriari2016]. Des approches telles que celles que nous avons testées reposant sur des modèles thématiques [Dermouche 2014 ; Dermouche 2016]

permettent en outre à l'utilisateur de comprendre les décisions du classificateur et d'avoir accès à des moyens d'action, ce qui en soi peut conduire à améliorer la performance des codeurs [Dinakar2014]. Nous envisageons également de recourir à des approches couplant la classification de textes et le plongement de mots (*word embedding*), et d'apprentissage de réseaux profonds (*deep learning*). Cependant, notre projet ne se limite pas aux données textuelles des comptes-rendus médicaux et nous souhaitons enrichir les données d'apprentissage à d'autres données hétérogènes du dossier patient informatisé. Les données quantitatives ou qualitatives issues d'examens médicaux biologique (biologie, hématologie, anatomopathologie,…), radiologiques, ou autres peuvent en effet être intégrées aux modèles comme des données fonctionnelles. Les modèles graphiques probabilistes peuvent être un formalisme adapté pour réaliser l'intégration de ces différentes modalités [Wainwright2008 et couplées à des méthodes d'optimisation bayésiennes permettant d'intégrer le retour des utilisateurs [Blei2014, Dinakar 2015].

Etapes

L'objectif sera découpé en plusieurs tâches :

Les données Ce travail s'appuie sur un entrepôt de données développé à l'AP-HP et regroupant l'ensemble des données informatisées du dossier patient. Une plateforme de traitement des données devra être développée. Elle permettra de réaliser une chaine de traitement dont les principales étapes sont : 1) mettre à disposition les données historiques pour l'apprentissage, 2) réaliser l'apprentissage sur les données 3) recueillir les données temps réel à utiliser pour la prédiction et la visualisation, 4) réaliser le feedback utilisateur. La question de l'accès et du partage de ces données sensibles est une problématique en soit et des méthodes devront être mise en place pour garantir leur sécurité. Des méthodes d'anonymisation et ou de simulation pourraient être utilisées pour la partie apprentissage supervisé.

Apprentissage supervisé [*Data challenges and RAMPs*] Une tâche vise à améliorer les performances des prédicteurs actuels en particulier par l'intégration de données hétérogènes non textuelles du dossier patient (aucune publication dans ce domaine). Entre autre des prédicteurs à bases de réseaux de neuronaux profond ou de modèles graphiques pourraient servir de comparateurs pour mesurer l'évolution des performances des modèles utilisés dans les étapes ultérieures.

Apprentissage interactif [*Standalone research engineers*] Un travail exploratoire sur les algorithmes d'apprentissage interactif disponibles sera réalisé. Des algorithmes seront développés pour les modèles répondant le mieux aux objectifs du projet qui outre les aspects de performance privilégie également une compréhension par l'utilisateur des éléments sur lesquelles la tâche de classification est basée.

Développement logiciels : Les logiciels de codage et de visualisation des données des dossiers patients ont été développés en interne. Ils sont disponibles sous licence open source. Ils constitueront une architecture de base pour le développement des interfaces utilisateurs.