English version following the French version.

TITRE DU PROJET : Evaluation des méthodes d'apprentissage automatique de type PU learning pour le phénotypage des maladies dans les bases de données médico-administratives : une proposition d'étude sur la cohorte Constances

RESPONSABLE : Karim Bounebache, Inserm, Unité US10 Centre d'épidémiologie sur les causes médicales de décès (CEPIDC), Le Kremlin-Bicêtre

RESUME

Contexte

Les études sur les bases de données médico-administratives (BDMA) requièrent fréquemment l'identification de certaines pathologies. L'identification de ces pathologies repose sur des algorithmes de phénotypage basés sur des règles élaborées par des experts. Le recours aux méthodes de machine learning de type PU learning (positive and unlabeled learning) pourrait aider à définir des algorithmes de phénotypage. Alors que la plupart des familles de machine learning sont fondées sur de la classification supervisée et s'appliquent à des ensembles de données étiquetées (malades, non malades), les méthodes de PU learning ne s'appuient que sur des données labélisées positivement (malades) et ne nécessitent pas un ensemble exhaustif d'observations. Il existe 3 grandes familles d'algorithmes de PU learning (approches heuristiques, approches par pondération, et méthodes d'ensemble) qui n'ont pas été évaluées et comparées dans le cadre de la construction d'algorithmes de phénotypage.

Objectifs

L'objectif principal est d'évaluer l'intérêt des méthodes de PU learning pour l'aide à la construction d'algorithmes de phénotypage dans les BDMA. Nous évaluerons ces méthodes dans le cadre du statut diabétique déterminé via les algorithmes du Redsiam et en comparaison avec le gold-standard que constitue le statut diabétique renseigné dans la cohorte Constances. Le diabète étant déjà exploré par les études de Fosse-Edorh et al., il constitue un cas d'usage solide pour évaluer les méthodes de PU learning.

Méthodes

Tous les patients, diabétiques et non diabétiques (groupe témoin), de la cohorte Constances seront inclus. Toutes les données du SNDS comprises entre le 01/01/2009 et le 31/12/2017 seront nécessaires. L'étude se déroule en 3 étapes. L'étape 1 de feature engineering consistera à essayer plusieurs stratégies de construction de matrices d'apprentissages. L'étape 2 repose sur des modèles de PU learning à partir des matrices d'apprentissage construites lors de l'étape 1. Pour lancer un apprentissage avec un algorithme de PU learning, il faut un ensemble initial d'observations étiquetées « diabétiques ». Nous évaluerons 3 ensembles initiaux : les traitements médicamenteux de ville spécifiques du diabète, les ALD pour diabète ou les hospitalisations pour diabète. A la suite des différents apprentissages, il en résultera des ensembles estimés d'individus diabétiques et des ensembles estimés d'individus non diabétiques. L'étape 3 consiste : (1) à effectuer des mesures de concordance, de type kappa de Cohen, entre les résultats des méthodes de PU learning et les différents algorithmes du Redsiam et le gold-standard, (2) à effectuer des mesures de performances

de classification des méthodes de PU learning comparativement au gold-standard (Précision, Rappel, Sensibilité, Spécificité, VPP, VPN…), et (3) à comparer qualitativement les variables sélectionnées par les méthodes de PU learning aux variables sélectionnées par les 3 algorithmes du Redsiam.

Perspectives

Nous faisons l'hypothèse que les méthodes guidées par les données de type PU learning sont adaptées à la construction d'algorithmes de phénotypage dans les BDMA et pourrait accompagner les experts dans l'élaboration d'algorithmes validés.

TITLE OF THE PROJECT: Evaluation of machine learning methods such as PU learning for building phenotyping algorithms in administrative databases: a study proposal on the Constances cohort

HEAD OF THE TEAM: Karim Bounebache, Inserm - US10 Epidemiological center on medical causes of death (CEPIDC), Le Kremlin-Bicêtre (France)

SUMMARY

Background

Epidemiological studies on administrative databases often require the identification of specific diseases. These identifications are based on rules-based phenotyping algorithms built by experts. Machine learning methods such as PU learning (positive and unlabeled learning) could improve the conception of phenotyping algorithms. While most machine learning families are based on supervised classification using fully labeled datasets, PU learning methods are based on partially and positively labeled data and do not require a complete annotated training set. There are 3 families of PU learning algorithms (heuristic approaches, biased methods, and ensemble methods) that have not been yet evaluated and compared for the purpose of the building of phenotyping algorithms.

Objectives

The aim of this study is to assess the PU learning methods for building phenotyping algorithms in administrative databases. We compare them to the diabetic status determined with the Redsiam algorithms and with the gold standard of diabetic status provided in the Constances cohort. As diabetes is already being explored by the studies, this use case is appropriate to assess a new methodology.

Methods

All patients, both diabetic and nondiabetic (control group), in the Constances cohort will be included. All SNDS data from 2009/01/01 to 2017/31/12 are required. The study is conducted in 3 steps. In Step 1 (feature engineering) we try several strategies for building a learning matrix. In step 2, we train PU learning models based on the learning matrix developed in step 1. To initiate the learning with a PU learning algorithm, an initial set of observations labeled diabetic patients is required. We evaluate 3 initial sets: specific antidiabetic drug treatments, Long Term Disease status for diabetes, or hospitalizations for diabetes. As a result of the different PU learning outcomes, we obtain a set of patients identified as diabetic and a set of patients identified as non-diabetic. In step 3 we compute: (1) the Cohen kappa between the results of PU learning methods and the Redsiam's algorithms (2) the performance of PU learning methods in comparison with the Gold-Standard (precision, recall, sensitivity, specificity, PPV, NPV), and (3) a qualitative comparison between the selection of variables performed by the PU learning methods and the predictors selected by the Redsiam's algorithms.

Perspectives

We hypothesize that data driven PU learning methods are suitable for building phenotyping algorithms in administrative databases and could help the experts for the construction of phenotyping algorithms.