Évaluation des améliorations de prédiction d'hospitalisation par l'ajout de connaissances métier aux dossiers médicaux

Raphaël Gazzotti[*,****], Catherine Faron-Zucker[*], Fabien Gandon[**],
Virginie Lacroix-Hugues[***], David Darmon[***]

[*]Université Côte d'Azur, Inria, CNRS, I3S, France, prénom.nom@unice.fr
[**]Inria, Université Côte d'Azur, CNRS, I3S, France, prénom.nom@inria.fr
[***]Université Côte d'Azur, Département de Médecine Générale,
vhugues@outlook.fr, david.darmon@unice.fr
[****]SynchroNext, France

Résumé. Les dossiers médicaux électroniques (DME) contiennent des informations essentielles sur les différents épisodes symptomatiques qu'un patient a subis. Cependant, les connaissances disponibles à travers ces enregistrements restent limitées : les attributs extractibles à partir de ces textes pour un algorithme d'apprentissage ne contiennent pas toutes les informations implicites connues par un expert. Afin d'évaluer et de pallier ce problème, nous avons étudié l'impact de l'augmentation des textes et des informations textuelles en provenance des DMEs par des annotations ontologiques générées automatiquement à partir de leur analyse afin d'enrichir en amont les représentations vectorielles utilisées ensuite par des algorithmes d'apprentissage.

1 Introduction

Les dossiers médicaux électroniques (DME) contiennent des informations essentielles sur les différents épisodes symptomatiques traversés par un patient. Ils possèdent le potentiel d'améliorer le bien-être des patients et constituent une source de données précieuse pour les approches d'intelligence artificielle. Dans cet article, nous augmentons avec des connaissances ontologiques les champs textuels issus des DMEs et évaluons leur impact sur la tâche de prédiction d'une hospitalisation. Notre évaluation se base sur un jeu de données réelles extrait de la base PRIMEGE PACA (Lacroix-Hugues et al. (2017)) qui contient plus de 350 000 consultations par 16 médecins généralistes. Dans cette base, les descriptions textuelles rédigées par les généralistes sont disponibles avec les codes de classification internationaux des médicaments prescrits, antécédents personnels, familiaux, facteurs environnementaux, pathologies et raisons de consultation, ainsi que les valeurs numériques des différents résultats d'examens médicaux. Les connaissances disponibles dans un DME restent cependant limitées aux spécificités de chaque patient et les textes qui s'y trouvent reposent sur un certain nombre d'informations implicites connues des experts médicaux et à des niveaux de détail variables. Aussi, un algorithme d'apprentissage automatique reposant sur ces seules informations ne pourra pas exploiter de connaissances spécifiques implicites dans les documents à analyser ou devra les réapprendre, possiblement de façon incomplète et coûteuse.

Notre principale question de recherche est alors la suivante : *Des apports de connaissances ontologiques dans les représentations destinées à l'apprentissage peuvent-ils améliorer la prédiction d'un événement ?* dans notre cas d'étude, nous cherchons à améliorer la prédiction de l'hospitalisation d'un patient à l'aide de connaissances provenant de différentes ontologies du domaine médical. Dans cet article, nous nous concentrons sur les sous-questions suivantes :

— *Comment intégrer des connaissances du domaine dans une représentation vectorielle destinée à un algorithme d'apprentissage ?*
— *L'ajout de connaissances du domaine améliore-t-il la prédiction d'hospitalisation ?*
— *Quelles connaissances du domaine combinées à quelles méthodes d'apprentissage fournissent une meilleure prédiction de l'hospitalisation d'un patient ?*

Pour répondre à ces questions, nous introduisons la méthode proposée pour l'annotation sémantique et l'extraction de connaissances à partir de textes puis nous précisons la façon dont les connaissances ontologiques sont injectées dans la représentation vectorielle des DMEs (section 2). Ensuite, nous présentons le protocole expérimental et les résultats obtenus (section 3), pour terminer avec la conclusion et les perspectives de cette étude (section 4).

2 Utilisation de connaissances ontologiques

2.1 Extraction de connaissances ontologiques

Afin d'extraire les connaissances du domaine sous-jacentes aux termes employés dans les descriptions textuelles rédigées par les médecins, nous recherchons au sein des textes les entités du domaine médical et les liens aux concepts auxquels ils correspondent dans Wikidata, DBpedia et des ontologies spécifiques au secteur de la santé. Wikidata et DBpedia ont été choisis du fait que des concepts généraux ne sont identifiables qu'avec des référentiels généraux. Notre étude vise à analyser et comparer l'impact apporté par ces connaissances issues de diverses sources sur la prédiction de l'hospitalisation.

Extraction de connaissances de DBpedia Pour détecter dans un DME les concepts du domaine médical dans DBpedia, nous avons utilisé l'annotateur sémantique DBpedia Spotlight. Nous avons procédé à une analyse manuelle des entités nommées détectées sur un échantillon de DME et déterminé 14 concepts SKOS désignant des sujets médicaux pertinents pour la prédiction d'hospitalisation, car relatifs à de lourdes pathologies : Anatomo-pathologie des tumeurs, Cancérologie, Radio-oncologie, Maladie cardio-vasculaire, Trouble du rythme cardiaque, Maladie neuro-vasculaire, Hémopathie maligne, Maladie auto-immune, Etat médical lié à l'obésité, Maladie génétique, Ablation chirurgicale, Défaillance d'organe, Urgence médicale et Urgence en cardiologie. Pour chaque DME à représenter, à partir de la liste des ressources identifiées par DBpedia Spotlight, nous interrogeons le point d'accès francophone de DBpedia afin de déterminer si ces ressources ont pour sujet (propriété `dcterms:subject`) un ou plusieurs des 14 concepts retenus. Afin d'améliorer l'annotation de DBpedia Spotlight, les mots ou expressions abrégées au sein des DMEs sont ajoutés aux champs textuels à l'aide d'une approche symbolique, par l'utilisation de règles et dictionnaire.

Extraction de connaissances de Wikidata Wikidata est une base de connaissances ouverte centralisant les données des divers projets de la fondation Wikimedia. Sa couverture est sur

certains domaines de connaissances différente de celle de DBpedia. Nous avons extrait des connaissances relatives aux médicaments en requêtant le point d'accès de Wikidata. Plus précisément, nous avons identifié trois propriétés de médicaments pertinentes pour la prédiction d'une hospitalisation : 'agit en tant que tel' (propriété `wdt:P2868`), 'maladie traitée' (propriété `wdt:P2175`), et 'médicament interagit avec' (propriété `wdt:P769`). A partir de l'URI d'un médicament, nous extrayons des couples propriété-concept liés aux médicaments pour les trois propriétés retenues (e.g. la péthidine est un narcotique, le méproprobamate soigne la céphalée, l'atazanavir interagit avec le rabéprazole).

Extraction de connaissances issues d'ontologies spécifiques au domaine Nous nous sommes intéressés à l'impact de connaissances issues d'ontologies spécifiques au domaine notamment sur les champs textuels comportant des codes internationaux de médicaments issus de la classification Anatomique, Thérapeutique et Chimique (ATC) et des codes relatifs aux raisons de la consultation auprès d'un médecin généraliste avec la Classification Internationale des Soins Primaires (CISP-2). Le choix de CISP2 et ATC dans notre étude vient du fait que la base de données PRIMEGE adopte ces nomenclatures. Nous avons extrait du vocabulaire ATC [1] représenté à l'aide de primitives OWL et SKOS les labels des super classes des classes relatives aux médicaments répertoriés dans la base de données PRIMEGE, grâce aux propriétés `rdfs:subClassOf` et `atc:member_of` sur différents niveaux de profondeur à l'aide de requêtes SPARQL avec des chemins de propriétés (e.g. le 'meprednisone' (code H02AB15) a pour super classe 'Glucocorticoids, Systemic' (code H02AB) qui a elle-même pour super classe 'CORTICOSTEROIDS FOR SYSTEMIC USE, PLAIN' (code H02)). De manière analogue, nous avons extrait de la représentation OWL-SKOS de CISP2 [2] les labels des super classes avec la propriété `rdfs:subClassOf`, cependant étant donné la faible profondeur de cette représentation, il n'est possible d'extraire qu'une super classe par problème de santé diagnostiqué ou procédure de soins identifiée (e.g. Symptôme et plaintes (code H05) a pour super classe Oreille (code H)).

2.2 Intégrer des connaissances ontologiques au vecteur de représentation

Il est crucial avec un corpus spécifique à un domaine de générer sa propre représentation, car de nombreux termes peuvent se trouver en dehors d'une représentation généraliste ou une notion ambiguë peut être associée à un terme alors qu'il possède un sens bien précis dans le domaine considéré. Nous avons opté pour un modèle exploitant la représentation par sac de mots (BOW) pour différentes raisons : (i) les principales informations de documents textuels sont extraites sans nécessiter un large corpus ; (ii) les attributs ne sont pas transformés ce qui permet d'identifier quels termes participent à la distinction de patients devant être hospitalisés ou non ; (iii) l'intégration de données hétérogènes est facilitée, car il suffit de concaténer d'autres attributs à ce modèle. À l'image de la structure employée dans PRIMEGE, certaines données textuelles doivent être distinguables les unes des autres lors du passage à la représentation vectorielle des DMEs comme, par exemple, les antécédents d'un patient et ses antécédents familiaux. Pour ce faire, un préfixe a été introduit à la création du BOW en fonction du champ textuel source. Soit $C^i = \{c_1^i, c_2^i, ..., c_n^i\}$ le sac de concepts résultant de l'extraction de

1. http://bioportal.bioontology.org/ontologies/ATC
2. http://bioportal.lirmm.fr/ontologies/CISP-2

concepts issus d'ontologies sur le i^{ème} patient après analyse des données textuelles structurées et des textes libres tels que les observations. Soit $V^i = \{w_1^i, w_2^i, ..., w_n^i\}$ le BOW obtenu à partir des données textuelles. Les différents algorithmes d'apprentissage exploitent l'agrégation de ces deux vecteurs : $x^i = V^i \oplus C^i$. Les concepts issus d'ontologies sont ainsi considérés comme un token dans un texte, lorsqu'un concept est identifié, il est ajouté à un vecteur de concepts et son attribut aura pour valeur le nombre d'occurrences de ce concept au sein du DME du patient.

3 Expérimentations et Résultats

3.1 Protocole

Nous avons expérimenté et évalué notre approche sur un ensemble de données équilibré DS_B contenant 714 patients hospitalisés et 732 non hospitalisés. Comme nous nous servons d'algorithmes d'apprentissage non-séquentiels pour évaluer l'enrichissement apporté par les connaissances ontologiques, nous avons dû agréger toutes les consultations d'un patient afin de s'affranchir de la dimension temporelle inhérente aux épisodes médicaux dans la vie d'un patient. Ainsi, toutes les consultations survenant avant une hospitalisation sont agrégées en une représentation vectorielle du dossier du patient. Pour les patients n'ayant pas été hospitalisés, l'ensemble de leurs consultations est agrégé. Nous avons évalué les représentations vectorielles ainsi construites par validation croisée imbriquée (Cawley et Talbot (2010)), avec une boucle externe ayant un K fixé à 10 et pour la boucle interne un K fixé à 3 avec exploration des hyperparamètres par recherche aléatoire sur 150 itérations. Les différentes expériences ont été menées sur un HP EliteBook 840 G2, 2,6 hHz, 16 go de RAM sous Python 3.6.3 ainsi qu'un Precision Tower 5810, 3.7GHz, 64GB RAM sous Python 3.5.4. La création des représentations vectorielles a été effectuée sur le HP EliteBook et sur cette même machine ont été déployés DBpedia Spotlight ainsi que les ontologies spécifiques au domaine.

3.2 Algorithmes d'apprentissage automatique

Nous avons effectué la tâche de prédiction d'hospitalisation avec différents algorithmes de l'état de l'art disponibles dans la bibliothèque Scikit-Learn avec des hyperparamètres déterminés par validation croisée imbriquée :

- SVC : Machine à vecteurs de support dont l'implémentation se base sur celle de libsvm (Chang et Lin (2011)). Le coefficient de régularisation C, le noyau utilisé par l'algorithme ainsi que le coefficient gamma du noyau ont été optimisés.
- RF : L'algorithme des forêts aléatoire (Breiman (2001)). Le nombre d'arbres dans la forêt, la profondeur maximale des arbres, le nombre minimal d'échantillons requis afin de diviser un nœud interne, le nombre minimal d'échantillons à prélever au niveau d'un nœud foliaire et le nombre maximal de nœuds foliaires ont été optimisés.
- Log : L'algorithme de la régression logistique (McCullagh et Nelder (1989)). Le coefficient de régularisation C et la norme utilisée dans la pénalisation ont été optimisés.

Nous avons opté pour ces algorithmes comme il est possible de fournir une interprétation native de leur décision, permettant ainsi de préciser au médecin les raisons d'hospitaliser un patient avec les facteurs sur lesquels il peut intervenir afin d'éviter que cet événement ait lieu.

3.3 Résultats et discussion

Afin d'évaluer l'intérêt de prendre en compte des connaissances ontologiques, nous nous sommes servis de la mesure $F_{tp,fp}$ (Forman et Scholz (2010)) pour évaluer les performances des algorithmes de classification. Soit TN, le nombre d'instances négatives correctement classées, FP le nombre d'instances négatives incorrectement classées, FN le nombre d'instances positives incorrectement classées et TP le nombre d'instances positives correctement classées.

$$TP_f = \sum_{i=1}^{K} TP^{(i)} \qquad FP_f = \sum_{i=1}^{K} FP^{(i)} \qquad FN_f = \sum_{i=1}^{K} FN^{(i)}$$

$$F_{tp,fp} = \frac{2.TP_f}{2.TP_f + FP_f + FN_f}$$

La Table 1 synthétise les résultats pour chaque méthode que nous avons testée sur DS_B :

— $référence$: représente notre base de comparaison où aucun enrichissement ontologique n'est fait sur les DMEs i.e. seulement les données textuelles sous forme de BOW.
— $+s$: enrichissement apporté avec des concepts de la base de connaissances DBpedia.
— $+s*$: indique un enrichissement apporté avec des concepts de DBpedia, contrairement à $+s$, la totalité des champs textuels n'est pas exploitée, ainsi sont extraits les concepts des champs relatifs aux antécédents du patient, ses allergies, les facteurs environnementaux, ses problèmes de santé actuels, ses motifs de consultations, ses diagnostics, ses médicaments, ses procédures de soin suivies, ses motifs de prescription de médicaments et les observations du médecin.
— $+t$: enrichissement avec des concepts de la représentation OWL-SKOS de CISP-2.
— $+c$: enrichissement avec des concepts de la représentation OWL-SKOS de ATC, le nombre ou la fenêtre de nombres accolés indique les niveaux hiérarchiques utilisés.
— $+wa$: enrichissement avec la propriété 'agit en tant que tel' de Wikidata.
— $+wi$: enrichissement avec la propriété 'médicament interagit avec' de Wikidata.
— $+wm$: enrichissement avec la propriété 'maladie traitée' de Wikidata.

Représentation	SVC	RF	Log	Moyenne
$référence$	0.8270	**0.8533**	0.8491	0.8431
$+t$	0.8239	0.8522	**0.8545**	0.8435
$+s$	0.8221	0.8522	0.8485	0.8409
$+s*$	0.8339	0.8449	0.8514	0.8434
$+c_1$	0.8235	0.8433	0.8453	0.8245
$+c_{1-2}$	0.8254	0.8480	0.8510	0.8415
$+c_2$	0.8348	0.8522	0.8505	**0.8458**
$+wa$	0.8223	0.8468	**0.8545**	0.8412
$+wi$	0.8149	0.8484	0.8501	0.8378
$+wm$	0.8221	0.8453	0.8458	0.8377

TAB. 1 – *$F_{tp,fp}$ pour les différents ensembles vectoriels envisagés avec différentes méthodes d'apprentissage sur le jeu de données équilibré DS_B.*

Malgré la faible profondeur de la représentation OWL-SKOS de CISP2 la configuration $+t$, est suffisante pour améliorer la prédiction d'hospitalisation d'un patient. Un deuxième niveau de hiérarchie de super classes, $+c_2$ de la représentation OWL-SKOS de ATC fournit de

meilleurs résultats qu'un seul niveau avec $+c_1$. Cependant, les résultats montrent que l'expansion par DBpedia à des champs indirectement liés à l'état du patient, tels que les antécédents familiaux, peut conduire les algorithmes de classification à tirer de mauvaises conclusions même si un préfixe a été ajouté pour distinguer l'origine des champs textuels. Le champ de texte relatif aux symptômes a été mal complété (souvent renseigné comme le champ d'observation) et la majorité des remarques ainsi détectées par DBpedia Spotlight sont principalement des fausses alertes. De plus, l'analyse qualitative des résultats a révélé des cas de négation (e.g. 'pas de SC d'insuffisance cardiaque') et de mauvaise prise en compte de plusieurs termes (e.g. "brûlures mictionnelles" qui est associé par DBpedia Spotlight à une "Brûlure" rapportant ainsi ce terme à une 'Urgence médicale').

4 Conclusion

Dans cet article, nous avons présenté une méthode pour coupler connaissances ontologiques, spécialisées ou généralistes, et données textuelles pour prédire l'hospitalisation de patients. Ainsi, nous avons généré diverses représentations couplant vecteurs de concepts et BOWs puis évalué leurs efficacités pour la prédiction avec divers algorithmes de classification. Nous prévoyons à court terme d'évaluer l'impact de nouvelles ontologies spécifiques au domaine et d'autres propriétés participant à la prédiction de l'hospitalisation de patients. Nous projetons aussi de travailler sur une représentation alternative couplant relations sémantiques et données textuelles, ainsi que la détection de la négation et des expressions complexes.

Références

Breiman (2001). Random forests. *Machine learning 45*(1), 5–32.

Cawley et Talbot (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research 11*(Jul), 2079–2107.

Chang et Lin (2011). Libsvm : a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST) 2*(3), 27.

Forman et Scholz (2010). Apples-to-apples in cross-validation studies : pitfalls in classifier performance measurement. *ACM SIGKDD Explorations Newsletter 12*(1), 49–57.

Lacroix-Hugues et al. (2017). Creation of the first french database in primary care using the icpc2 : Feasibility study. *Studies in health technology and informatics 245*, 462–466.

McCullagh et Nelder (1989). *Generalized linear models*, Volume 37. CRC press.

Summary

The knowledge available through electronic medical records (EMR) themselves remains limited by the fact the features used by a machine learning algorithms from a text alone do not contain all the implicit information known by a domain expert. We propose and evaluate the ontological augmentations of features extracted from textual information from EMRs on several machine learning algorithms to predict hospitalization.