UNIVERSITÉ LAVAL

Analyse et traitement de données massives

GLO-7027

Rapport 2 : Traitement des données

Buis Anh My

Bouchery Loïc

Remis le 26/04/2023

Introduction

La polypharmacie, souvent définie comme une prescription de plus de cinq médicaments, est utilisée pour des maladies multiples ou complexes nécessitant la prise de plusieurs substances afin de lutter contre un ensemble de symptômes. Ces substances pouvant potentiellement interagir entre elles, cela comporte des risques supplémentaires pour les patients et peuvent mener à des hospitalisations suite à l'apparition d'effets secondaires indésirables.

Dans le cadre de ce projet, on vise à établir une méthode permettant d'estimer si les combinaisons de médicaments étudiées sont potentiellement nocives, ce qui nécessiterai par la suite une étude médicale sur les interactions entre ces médicaments.

Il est donc impératif d'avoir une méthode claire capable d'identifier avec précision si une combinaison est à risque ou non.

I – Rappel sur le prétraitement des données

Notre étude se base sur un jeu de données de la RAMQ contenant 30 220 351 entrées correspondant à un évènement de santé, sur un total de 3 000 000 patients. Chacun de cet évènement correspond à la prise de médicaments d'un patient entre les années 2000 à 2023, et si une hospitalisation a été nécessaire pendant la durée du traitement.

On sépare le jeu de données sur la base du nombre de médicaments pris simultanément, et on obtient un second jeu de données dit « polypharmacies » contenant uniquement les observations pour lesquelles les patients prennent au moins 5 médicaments simultanément.

Le jeu de données total est déséquilibré en faveur des patients non hospitalisé (5 pour 1), et cela reste vrai dans le sous jeu contenant uniquement le polypharmacies (3.8 pour 1). Comme on a affaire à un jeu de données non équilibré, les classifieurs qu'on va employer peuvent se retrouver biaisés au profit de la classe du dataset la plus fréquente.

Au sein du jeu contenant les polypharmacies, aucun médicament n'est sous représenté, il n'y a donc pas d'ajustement à faire à ce niveau sur les données.

On veillera cependant à retirer d'eventuels doublons et données semblant aberrantes, par exemple les observations pour un même patient le même jour, contenant des données sur les médicaments completement differentes.

Une fois le jeu de données adapté et nettoyé, on va employer deux approches afin de le traiter et pouvoir identifier les polypharmacies potentiellement dangereuses : l'emploi de règles d'association puis un des classifieurs plus traditionnels, Random Forest.

II – Approche intuitive : Règles d'association

L'approche intuitive vis-à-vis de notre problème est l'emploi de règles d'association : on veut pouvoir mettre en évidence les liens complexes entre les variables médicaments et hospitalisation pour mettre à jour les combinaisons néfastes.

On utilise le jeu de données contenant uniquement les polypharmacies, et on utilise un algorithme de règles d'association pour nous permettre d'identifier les éléments les plus fréquents dans ce sous-ensemble, que l'on va comparer avec l'ensemble global.

Pour ce faire, on utilise un algorithme de référence, FPGrowth. Nous avons dans un premier temps tenté d'utiliser Apriori, mais son efficacité spatiale étant bien moindre, nous avons vite été limités lorsqu'il a été question d'aller chercher des combinaisons récurrentes contenant plus de trois médicaments.

En effet, contrairement à Apriori, cet algorithme crée un arbre de fréquence sans générer d'ensembles d'éléments candidats, ce qui permet un gain en mémoire (et de fait, en temps de calcul) non négligeable dans notre cas au vu du nombre de données à traiter.

Cet algorithme permet de plus de s'intéresser aux combinaisons à l'intérieur d'autres combinaisons, permettant par exemple de considérer des polypharmacies de 5 médicaments au sein d'une autre de 7. Cela permet d'identifier d'éventuels exemples dans lequel une combinaison néfaste est contenue mais pourrait etre « cachée » par la présence d'autres médicaments avec un effet bénin.

On identifie les itemsets les plus fréquents lorsque l'on est en situation de polypharmacie et d'hospitalisation et on ne tient compte ici que des combinaisons, puisque les médicaments individuellement ont tous un faible lien avec l'hospitalisation (il représentent tous environ 28% des cas, sachant qu'il y a 30% d'hospitalisations).

Si l'on tient uniquement compte des polypharmacies, on obtient après avoir calibré le support minimum à 2.5% le tableau suivant :

support	itemsets	length
0.034070	[drug_18, drug_17, drug_10, drug_4, drug_5]	5
0.033820	[drug_18, drug_17, drug_10, drug_11, drug_4]	5
0.033775	[drug_18, drug_10, drug_11, drug_4, drug_5]	5
0.033701	[drug_18, drug_17, drug_11, drug_4, drug_5]	5
0.033495	[drug_18, drug_17, drug_10, drug_11, drug_5]	5
...	...	...
0.025869	[drug_15, drug_3, drug_14, drug_13, drug_8]	5
0.025850	[drug_3, drug_14, drug_2, drug_8, drug_6]	5
0.025784	[drug_7, drug_3, drug_14, drug_8, drug_6]	5
0.025768	[drug_3, drug_14, drug_13, drug_4, drug_8]	5
0.025625	[drug_3, drug_14, drug_13, drug_8, drug_6]	5

La variable support indique quel pourcentage des hospitalisations est observé en même temps que la prise de la polypharmacie par rapport au total des hospitalisation pour toutes les polypharmacies.

Sachant que les combinaisons de médicaments s'étalent sur 19 attributs, un faible écart de support de 1% est très pertinent dans le cadre de notre étude : le classement par ordre décroissant du support permet ici d'identifier les suspects potentiels.

En effet, on remarque que tous les 5 polypharmacies présentes en tête du classement contiennent les médicaments 18, 17, 10, 11 et 5 et 4. On s'attend donc à ce que les combinaisons nocives soient parmi celles-ci. Comme l'algorithme tiens compte des combinaisons de plus de 5 éléments, on se doute que des combinaisons plus grandes contenant ces 6 médicaments contribuent à ce support.

Pour s'assurer de la pertinence de ce choix, et pour développer un critère de décision précis, on considère l'étude de ces combinaisons dans le jeu de données total des polypharmacies, afin de s'assurer qu'il y a bien une différence significative entre les combinaisons classées en haut et en bas du tableau.

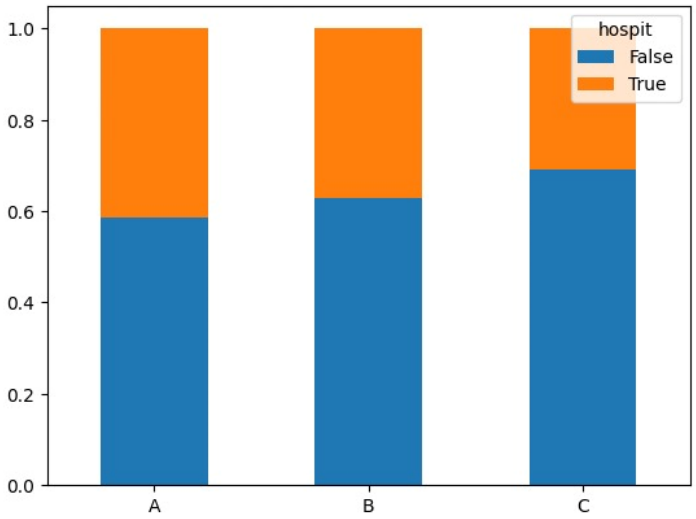
On effectue pour cela le rapport entre les observations avec et sans hospitalisation pour 3 combinaisons de médicaments :

A : 18, 17, 10, 4, 5

B : 10, 11, 12, 1, 5 (prise en milieu de tableau)

C : 3, 14, 13, 8, 6

Le résultat est présenté sous forme du graphique ci-dessous :



Sachant que la taille des échantillons est comparable pour ces 3 combinaisons, on calcule un taux d'hospitalisation différent de 10.5 % entre les combinaisons A et C (41.3 contre 30.8), ce qui n'est absolument pas négligeable.

On compare cela au taux d'hospitalisation moyen pour l'ensemble des polypharmacies, qui est de 30.2 %. On remarque donc que la plupart des combinaisons sont plus proches de la valeur basse que de la valeur haute, et sont sous le support utilisé précédemment.

On décide donc d'utiliser un critère simple sur ces combinaisons identifiées afin de déterminer lesquelles sont réellement nocive. Tout d'abord, on ajuste donc le support de manière à avoir un nombre raisonnable de combinaisons.

Sachant que certaines combinaisons sont plus susceptibles de provoquer des hospitalisations, on cumule le critère du support de l'algorithme à un critère d'écart à la moyenne afin d'affiner les résultats obtenus.

Tout d'abord, on limite le support minimum a 3.2 : on obtient alors un nombre de combinaisons potentielles de 220. Parmi ces combinaisons, toutes ont un taux d'hospitalisation bien supérieur à la moyenne, d'au moins 38%.

Pour affiner nos règles d'association, on peut se restreindre à celles dont le taux est d'au moins 9% supérieure à la moyenne, et on en obtient alors 114.

Il y a au total 11 628 combinaisons possibles de médicaments, et parmi les 11 millions de données sur les polypharmacies, toutes les combinaisons possibles sont prescrites entre 280 000 et 290 000 fois (plusieurs combinaisons de 5 peuvent se retrouver au sein d'une même combinaison de 6 ou plus).

On ne semble donc pas avoir besoin de pondérer les critères par la popularité des combinaisons. Cette popularité est des plus bruitée, puisque toutes les combinaisons linéaires des polypharmacies établies peuvent elles aussi être considérées comme nocives.

Avec les règles de d'association, on trouve qu'environ 0.1% des polypharmacies étudiées sont nocives, ce qui est raisonnable puisque ces cas sont censés être rares et difficile à détecter (puisque dans le cas inverse, les médecins ne prescrivent pas ces combinaisons).

En effet, comme les données sont anonymisées, nous ne pouvons pas savoir quel est l'état de santé global du patient lors de la prise de médicaments. Si une polypharmacie a été identifiée comme nocive par nos règles d'association mais que cette combinaison n'est prescrite qu'a des patients dont l'état de santé est déjà extrêmement fragile, alors les professionnels de santé chargé d'étudier la liste donnée sauront que de telles combinaisons sont à ignorer.

II – Approche via classifieurs

Nous avons d'abord pensé à employer la classification naïve de Bayes. Ce classifieur se repose sur le calcul des probabilités a priori et conditionnelles des caractéristiques pour chaque classe.

Il est souvent adapté aux ensembles de données de grande dimension et à grande échelle, mais ses prédictions suggèrent une indépendance entre les dimensions, qui n'est pas respectée ici puisque certains médicaments sont souvent prescrits ensembles, et d'autres combinaisons sont évitées. On évitera donc l'emploi d'un tel classifieur.

Comme les arbres de décision sont plus adaptés, on va alors utiliser un classifieur simple mais puissant : Random Forest. Il s'agit d'une méthode par ensemble utilisant plusieurs arbres de décision indépendants construits à partir d'échantillons de données. On utilise comme facteur de décision des arbres le coefficient de Gini.

En termes de métriques, on s'intéresse surtout à la capacité de nos classifieurs à identifier les combinaisons néfastes. Pour cela, nous considérons une polypharmacie comme dangereuse si le classifieur prédit que prendre cette combinaison de médicaments a pour conséquence l'hospitalisation du patient, et nous l'évaluons à l'aide de l'exactitude (accuracy), la précision, le rappel et le F1-Score. La précision nous donne une indication sur le taux de faux positifs, et le rappel sur les faux négatifs.

Les médicaments sont prescrits afin de combattre plusieurs symptômes, en retirer un par crainte d'une polypharmacie nocive alors qu'elle ne l'est pas réellement est dangereux, et que comme le stipule l'énoncé, lancer une étude pour déterminer si les combinaisons sont dangereuses est très coûteux. C'est pourquoi il est important de garder à l'esprit qu'un faux positif est moins souhaitable qu'un faux négatif.

On donne à chacun des classifieurs ci-dessous un jeu d'entrainement contenant 80% des données, et un jeu de test permettant de réaliser la mesure des métriques contenant les 20% restants, et ces données sont initialisées au hasard dans les jeu de données indiqués.

Classifieur	Nombre d'arbres	Jeu de données	Exactitude	Précision	Rappel	F1 Score
RF1	50	Polypharmacies	66	38	18.8	25.1
RF2	75	Polypharmacies	67	40.1	16.4	23.2
RF3	100	Polypharmacies	68	42.5	13.2	20.1
RF4	150	Polypharmacies	67	40	16	23.0
RF ALL	100	Entier	71	46	16.9	23.7

Nous avons établi plusieurs classifieurs en fonction du nombre d'arbres et également du jeu de données utilisées. Pour les quatre premiers classifieurs, le jeu de données se base que sur des cas de polypharmacies, et le dernier sur l'ensemble des données.

L'entrainement sur l'ensemble du jeu de données, qui contient environ 3 fois plus de données, peut sembler plus intéressant car ses métriques sont légèrement supérieures, mais comme les données liées à des combinaisons très simples de médicaments (1 ou 2) donnent des résultats peu significatifs du point de vue des polypharmacies, alors que son temps d'entrainement est bien supérieur. On se restreindra donc au jeu de données des polypharmacies pour la suite de l'étude.

Cette étude permet d'identifier un nombre d'arbre 'optimal' de 100. Au-delà, les performances ne s'améliorent pas vraiment, malgré une hausse exponentielle du temps de calcul.

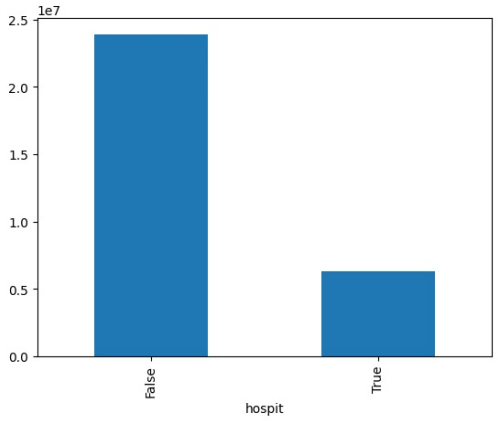
Malgré les améliorations successives lorsque l'on ajoute des agents de classification (nombre d'arbres), le F1 score du modèle reste faible comparé aux attentes, plus particulièrement le score de rappel.

Pour mieux comprendre comment le modèle obtient de tels scores, on s'intéresse à l'importance des features pour le classifieur RF3 résumés dans la table suivante :

	feature	importance
0	drug_8	0.072113
1	drug_9	0.070247
2	drug_2	0.068286
3	drug_6	0.067512
4	drug_0	0.065137
5	drug_3	0.062829
6	drug_13	0.061323
7	drug_7	0.060800
8	drug_1	0.059943
9	drug_15	0.058003
10	drug_14	0.057875
11	drug_10	0.052035
12	drug_16	0.042052
13	drug_5	0.041413
14	drug_17	0.037990
15	drug_11	0.033802
16	drug_18	0.033694
17	drug_4	0.029571
18	drug_12	0.025375

On observe un déséquilibre dans l'importance des features. Au vu des résultats établis sur les règles de décision, on s'attend à ce que les médicaments 18, 17 et 4 fassent parti des features les plus importantes pour la prise de décision positive, puisqu'ils se retrouvent dans les polypharmacies ayant un taux d'hospitalisation supérieur à la moyenne.

Si l'on rappelle la répartition de données sur l'ensemble du jeu :



Cette distribution est très inégale (ratio de 2.3 entre les 2 classes), il y a plus de chance qu'une instance soit classée négative (non hospitalisé) mais soit en réalité positive (hospitalisé) ce qui résulte en un nombre important de faux négatif, donnant donc un score faible au niveau du rappel : le modèle tend à prédire la classe la plus populaire.

C'est aussi la raison du déséquilibre dans les features du modèle : il privilégie des features capables de lui donner la meilleure précision sur l'ensemble, soit la décision la plus populaire.

Afin de gérer le déséquilibre de classe qui peut affecter la performance du modèle en termes de rappel, nous allons attribuer un poids plus élevé à la classe minoritaire "hospitalisé" pour la construction des arbres.

Cela permet de donner plus d'importance à la prédiction correcte des cas d'hospitalisation et d'éviter que le classifieur ne soit biaisé en faveur de la classe majoritaire « non hospitalisé ». Ainsi, on s'assure que le modèle apprend mieux à partir des exemples de la classe « hospitalisé », ce qui peut améliorer la performance du modèle sur cette classe spécifique. Cependant, on s'attend à perdre en exactitude et dans une moindre mesure, en précision.

Concrètement, cela revient à ajuster le critère de séparation à chaque nœud de l'arbre afin de mieux discriminer la classe minoritaire, ici une hospitalisation, cela permet de mettre l'emphase sur les attributs les plus pertinents dans notre classification.

On conservera un nombre d'arbres de 100 pour les classifieurs suivants, et on applique des poids plus forts sur la classe positive, indiqué comme « Ajustement X : 1 » ou X est le poids de la classe positive par rapport à celui de la classe négative, gardé à 1.

Classifieur	Exactitude	Précision	Rappel	F1 Score
RF3	68	42.5	13.2	20.1
Ajustement 1.1 : 1	65	39.3	20.9	27.3
Ajustement 1.5 : 1	65.0	38.3	28.0	32.3
Ajustement 2.3 : 1	58.8	36.5	49.7	42.1
Ajustement 3 : 1	53.3	34.1	58.2	43

Comme prévu, à mesure que le facteur de poids augmente, l'exactitude diminue. Cette manipulation a tout de même permis d'augmenter significativement les scores de rappel, même si la précision a légèrement diminué. Cela explique pourquoi le score F1 est considérablement plus élevé.

En effet, avec l'ajustement des poids de classe, le score F1 est bien meilleur, et l'on observe une hausse de plus de 20%. Cependant, dans notre cas, la baisse de la précision n'est pas souhaitable, puisque l'on veut en particulier éviter les faux positifs.

De plus, ce classifieur reste médiocre dans son ensemble : il est capable d'identifier seulement la moitié des cas, et la confiance sur les résultats obtenus est assez faible comparée à nos attentes, ce qui rend ce modèle peu utilisable en pratique pour la détection des polypharmacies novices, même suite à l'ajustement des poids de classe.

Si l'on l'utilise tout de même sur l'ensemble des combinaisons possibles de 5 médicaments, on obtient 126 combinaisons potentielles.

III- Etude de cas

Au vu des performances du Random Forest, on utilisera pour l'étude de cas les règles d'association obtenues avec l'algorithme FPGrowth. Il utilise les combinaisons de médicaments pris par le patient, et juge si cette combinaison résulte en un taux d'hospitalisation bien supérieur à celui des autres, à partir d'un arbre de fréquence construit sur le jeu de donnée initial.

Nous avons choisi une combinaison de médicaments potentiellement dangereuses en utilisant notre règle d'association : la combinaison [4, 11, 16, 17, 18], ainsi que la combinaison [0, 2, 6, 9, 10] ne faisant pas partie des combinaisons jugées dangereuses. Pour examiner ces combinaisons plus en détail, nous avons examiné leurs répartitions entre les patients hospitalisés et non hospitalisés, ainsi que les antécédents médicaux des patients 9437 et 2045783 qui ont pris ces combinaisons.

Pour simplifier notre étude et de nous concentrer sur les effets spécifiques de chaque combinaison, nous avons choisi des cas où chaque combinaison n'était pas mélangée à d'autres médicaments. De cette façon, nous avons pu isoler les effets potentiels de chaque combinaison et les examiner en détail.

Combinaisons	Hospitalisés	Non Hospitalisés
[4, 11, 16, 17, 18]	78	121
[0,2,6,9,10]	43	130

Patient	Combinaisons	timestamp	hospit
9437	[4, 11, 16, 17, 18]	2001-11-29	Oui
9437	[4, 11, 16, 17, 18]	2002-01-25	Non
9437	[4, 11, 16, 17, 18]	2002-02-04	Oui
9437	[4, 11, 16, 17, 18]	2002-05-03	Oui
2045783	[0, 2, 6, 9, 10]	2016-12-01	Oui
2045783	[0, 2, 6, 9, 10]	2017-01-13	Oui
2045783	[0, 2, 6, 9, 10]	2017-02-27	Oui

Dans le cas spécifique du patient 9437, la combinaison de médicaments [4, 11, 16, 17, 18] peut potentiellement avoir contribué aux hospitalisations du patient au vu d'un nombre plus important d'hospitalisations sur 6 mois avec son utilisation.

Cependant, pour le patient 2045783, la combinaison [0, 2, 6, 9, 10] qui était considérée non nocive selon notre règle d'association a été potentiellement responsable des 3 hospitalisations du patient.

Cette étude de cas illustre la capacité de notre règle d'association à détecter des risques de polypharmacie nocive dans le cas du patient 9437, en identifiant des combinaisons fréquentes de médicaments qui peuvent potentiellement entrainer l'hospitalisation.

Mais le cas du patient 2045783 illustre aussi bien notre règle d'association, où même si la combinaison [0, 2, 6, 9, 10] n'a que 43 cas d'hospitalisations contre 78 pour la combinaison [4, 11, 16, 17, 18], elle peut se montrer nocive pour certains patients.

Conclusion et retrospective

Bien que le Random Forest soit un outil puissant pour la classification dans de nombreux domaines, dans ce cas précis, il n'est pas l'approche la plus appropriée pour la détection de polypharmacie nocive, même si plus d'efforts ont été fournis par rapport aux autres méthodes employées. S'il était question de refaire le projet, on pourrait utiliser les données sur 1 année pour l'entrainement de ce type de modèle afin d'accélérer le traitement et faire une recherche en grille d'hyperparamètres optimaux.

Il peut être également intéressant de prendre en compte la temporalité, étant donné que beaucoup de patients prennent une même combinaison régulièrement dans un laps de temps, par exemple en ajoutant une colonne contenant des informations sur le nombre de fois que la combinaison de médicaments de l'instance a été prise par un patient, puis d'entrainer le classifieur avec ce genre de données. Cependant, de part le manque d'information sur l'état du patient lors de l'observation, il est possible que cette information n'ajoute pas grand-chose d'utile.

Les règles d'association déterminées avec FPGrowth nous donnent cependant une liste suffisamment précise pour être utilisée en pratique, et c'est ces résultats contenant 114 combinaisons que nous garderons.