Projet Machine Learning :
Analyse de base de données médicamenteuse

Kouyou Essohanam, Abdou Chafin Dean, Aujogue Jean-baptiste

February 13, 2018

Introduction

Ce travail se base sur un jeu de donnée de la plateforme open data de l'assurance maladie Française [4]. Le fichier considéré est la table NB_2016_cip13_age_sexe_reg_spe.CSV téléchargeable sur le lien [3], qui recence l'ensemble des médicaments vendus en officines de villes durant l'année 2016 selon un code unique par médicament, et suivant le sexe, la tranche d'âge et la région de résidence des consommateurs.

Le problème auquel nous nous somme intéressé est la question simple suivante : Etant donné un nouveau médicament, par qui sera-t-il consommé, et en quelles quantités ?

Dans ce travail on considérera une version simplifiée de ce problème, et on cherchera à prédire le mieux possible, pour un médicament donné, simplement le sexe et la tranche d'âge des personnes les plus susceptibles de consommer ce médicament. Pour cela, il est nécessaire de rassembler des attributs pouvant décrire nos médicaments, sur lesquels ont basera notre prédiction. Nous considérerons ici deux ensembles d'attributs : l'ensemble des codes de classification ATC d'une part, et l'ensemble des principes actifs d'autre part. L'idée naive étant ici que des médicaments ayant le plus de codes en commun, ou partageant le plus de principes actifs, sont en principes similaires et donc destinés à un même groupe de personnes. La principale difficulté dans ce travail vient de la bonne sélection des attributs explicatifs, et du fait que tous les attributs utilisés ici sont qualitatifs à énormément de modalités.

I. Table descriptive des médicaments

Dans cette partie on se sert uniquement du code CIP13 servant à identifier chaque médicament dans notre base de données.

I.1 Table de médicaments par les 5 codes ATC

On recoupe la liste de ces codes avec la table de données disponible sur [1], permettant de joindre à chaque code CIP13 cinq codes différents dans la classification ATC de 1 à 5 de nos médicaments. La classification ATC1 est la plus grossière avec 15 classes seulement, correspondant aux groupes anatomiques comme le système nerveux, le système respiratoire, etc. Chaque classification raffine la classification précédente. On génère alors une table, notée **T_1_bin** dans notre notebook, contenant les 5 codes ATC sous une forme binarisée par un one hot encodage. La table ainsi obtenue

possède 1874 attributs binaires. On réduit alors la dimension de cette table par décomposition en valeurs singulières tronquée, et on considère dans la suite une description des médicaments par les 20 ou 50 premiers vecteurs fournis par la décomposition en valeurs singulière, à travers deux tables T_1_svd_20 et T_1_svd_50. On a également essayé la réduction de dimension avec une normalisation suivie d'une ACP mais les résultat d'apprentissages ne sont pas convaincants dans ce cas, on ne présente donc pas cette solution plus en détail.

I.2 Table des médicaments par les principes actifs

On joint tout d'abord un code CIS correspondant à chaque médicament, que l'on exploite ensuite pour extraire l'ensemble des principes actifs de chaque médicament de la table téléchargeable en [2].

Remarque : Deux médicaments avec le même code CIS sont vraissemblablement les mêmes médicaments mais avec des dosages différents, ce qui a un impact sur la tranche d'âge des consommateurs (et probablement le sexe aussi). Mais cet impact est 'inversé' au sens où deux médicaments avec même CIS donnent probablement des résultats opposés de prédiction d'âge et de sexe. On n'est pour l'instant pas sûr de la méthode a suivre pour incorporer ce code pour que l'apprentissage se fasse correctement, ie de façon à ce que l'algorithme observe que "même CIS => âge et sexe des consommants opposés".

On génère alors une table binaire **T_3_bin** décrivant chaque médicament par un vecteur de 0 et de 1 signalant la présence/l'abscence de chaque principe actif. On aurait également souhaité intégrer le dosage de chaque substance, en prennant garde au fait que la quantification est relative à une gélule, un comprimé, une 'dose', un volume, un poids... une référence qui n'est pas la même pour tous les médicaments Une petite vérification montre d'ailleurs que notre fichier de médicaments contient 72 désignation différentes ! Des pilules, gels, champooings et autres.. Il est donc très difficile de donner un 'dosage' pour des désignations si différentes. On se concentre donc uniquement sur la présence /l'abscence de chaque principe actif. Un calcul permet de dénombrer 1645 substances différentes, soit autant d'attributs binaires pour décrire chaque médicament. Une fois générée cette table binaire, onapplique une réduction de dimension par décomposition en valeurs singulières comme précédemment.

I.3 Rassemblement des tables

On fusionne ensuite les tables construites à partir des codes ATC et des principes actifs. On a deux possibilités : soit on prend les tables des 20 ou 50 premières composantes singulières que l'on fusionne, soit on fusionne les tables binaires complètes T_1_bin et T_3_bin pour ensuite effectuer une réduction de dimension. On applique chacune des possibilités ici, donnant les tables T_4_svd_20, T_4_svd_50 et T_5 respectivement.

II. Table des valeurs à prédire

On se propose de prédire deux sorties, toute deux qualitatives : le sexe le plus probable d'un consommateur aléatoire d'un médicament donné, qui possède 2 modalités, et la tranche d'âge, qui en possède 3. Ce problème est bien sûr différent de prédire le *couple* (sexe, âge) le plus probable puisque ces deux variables sont vraissemblablement dépendantes l'une de l'autre.

II.1 Sortie 'sexe'

On génère une table donnant, pour chaque médicament, le nombre de boites vendues aux consommateurs de chaque sexe (plus le sexe inconnu), à partir de laquelle on génère un vecteur de sortie binaire, qui prend les valeurs 0 ou 1 suivant si un médicament est plus consommé par les hommes ou par les femmes. En plus du vecteur de sortie binaire créé, on veut créer une sortie **probabiliste** donnant pour chaque sexe sa probabilité empirique pour un consommateur aléatoire d'un médicament donné, ou de manière équivalente, la probabilité que ce sexe soit masculin. Un classifieur déterministe est alors naturellement obtenu en classifiant par 0 ou 1 selon le seuil 0.5 (ou un seuil possiblement calibré).

II.2 Sortie 'âge'

On génère une table qui à chaque médicament donne un quadruplet, formé des nombres de boites vendues à chaque tranche d'âge (plus l'âge inconnu), à partir de laquelle on génère la table qui à un médicament associe un triplet binaire, ie un triplet de la forme (0,1,0) par exemple, où la position du 1 doit correspondre à la tranche d'âge la plus consommatrice de ce médicament.

III. Algorithmes d'apprentissage

III.1 Prédiction du sexe

On effectue une prédiction du sexe prédominant parmi les consommateurs d'un médicament donné, sur la base des codes ATC de 1 à 5 de chaque médicament : Que l'on considère la table T_1_bin à 1874 attributs ou bien l'une des tables T_1_svd_20 et T_1_svd_50 à 20 et 50 attributs respectivement, on peut prédire le sexe prédominant parmi les consommateurs d'un médicament donné avec un taux de biens classés de 79 %, avec une AUC de l'ordre de 0.85. Ces résultats de prédiction sont donc relativement satisfaisants.

On effectue ensuite une prédiction du sexe prédominant parmi les consommateurs d'un médicament sur la base de ses principes actifs, qui donne des résultats légerement moindres que l'analyse précédante. Puis on teste enfin avec les tables T_4 et T_5 : On observe que, malgré la considération simultanée des description de médicaments par codes ATC et principes actifs, la base des 80 % des biens classés estimé par validation croisée n'est pas atteinte. Le résultat de cette démarche est donc décevant. Un point qui pourrait être important serait de trouver une méthode afin d'intégrer le mieux possible les proportions de chaque principe actif dans les médicaments, afin d'améliorer le taux de biens classés. Un second point à développer est d'effectuer l'apprentissage d'un classifieur probabiliste, pour ensuite construire un classifieur déterministe avec un seuil de 0.5 (ou laissé pour calibration).

III.2 Prédiction de la tranche d'âge

On estime le taux de biens classés pour un ensemble d'algorithmes construits sur la classification ATC binarisée, ainsi que sur le 20 et 50 premiers vecteurs singuliers issus de cette table : Les classifieurs obtenus ont de très bons résultats, avec des taux de biens classés estimé par validation croisée à 81 % pour la plupart. La réduction de dimension n'altère presque pas les résultats, et offre un gain notable en temps de calcul. Des résultats très légèrement plus faibles sont obtenus à partir de la table des principes actifs, traitée après décomposition en valeurs singulières tronquée, avec

des taux de biens classés estimé par validation croisée à 79 % dans le meilleur des cas. On souhaite enfin savoir si la fusion des deux tables améliore cette prédiction : Il s'avère malheureusement que non, c'est encore les attributs issus des codes ATC qui semblent porter toute la qualité de prédiction.

Conclusion

Dans ce travail on a tenté de fournir un modèle de prédiction du sexe et de la tranche d'âge prédominants parmis les consommateurs d'un médicament donné. Une première remarque est que cette analyse ne tire pas tout le potentiel offert par les principes actifs de médicaments, ce qui devrait pouvoir être fait via l'introduction de dosage ou par une méthode de réduction de dimension alternative à la svd tronquée. La méthode suivie pour construire les algorithmes de prédiction devrait être revue sous un angle plus probabiliste. Enfin, il serait important de prédire les couples (sexe , âge) le plus probable pour un médicament donné, plutôt que chaque variable séparément.

References

[1] Base de données open medic. *http://open-data-assurance-maladie.ameli.fr/medicaments*.

[2] Fichier des compositions. *http://agence-prd.ansm.sante.fr/php/ecodex/telecharger/telecharger.php*.

[3] Jeu de données. *http://open-data-assurance-maladie.ameli.fr/medicaments/download2.php?Dir_Rep=2016_CIP13*.

[4] Site de l'assurance maladie. *http://open-data-assurance-maladie.ameli.fr*.