Représentations numériques de codes médicaux dans un espace sémantique par factorisation de matrice d'information mutuelle

Corentin Faujour*[1,2], Stéphane Bouee[2], Corinne Emery[2], and Anne-Sophie Jannot[1,3]

[1]HeKA, Inserm, Université Paris Cité – Inria de Paris, Centre de Recherche des Cordeliers – France
[2]CEMKA, Bourg La Reine – Cemka – France
[3]Banque Nationale de Données Maladies Rares (BNDMR) – Assistance publique - Hôpitaux de Paris (AP-HP) – France

Résumé

Introduction - A partir des bases de données médico-administratives telles que le Système National des Données de Santé, les parcours de soins sont fréquemment représentés par des séquences de codes médicaux issus de nomenclatures différentes (diagnostics, traitements...). Les méthodes proposées pour étudier ces séquences reposent en grande partie sur les notions de distance entre séquences (alignement optimal, distance d'édition...) et d'extraction de variables latentes (pattern mining, topic models...).
De manière classique, ces méthodes utilisent une représentation binaire (one-hot) des codes médicaux possédant plusieurs limites. D'une part, la représentation binaire des codes ne permet pas de prendre en compte la similarité potentielle entre deux codes différents. D'autre part, l'encodage one-hot donne lieu à des vecteurs de grande dimension qui rendent nécessaire la sélection et le regroupement manuel des codes lors de chaque application.

Pour ces raisons, il semble pertinent de remplacer les vecteurs one-hot par une représentation numérique des codes médicaux de faible dimension, dense et dont la structure capture le lien sémantique qui les unit.

Méthodes - La méthode employée s'appuie sur l'approche distributionnelle issue du Traitement Automatique du Langage : un calcul de cooccurrences entre codes est réalisé à partir des séquences individuelles. La matrice d'information mutuelle (PMI) correspondante est factorisée par décomposition en valeurs singulières de manière à obtenir des vecteurs denses. Nous représentons la similarité entre deux codes comme étant l'angle formé par leurs vecteurs respectifs (similarité cosinus).

Résultats - A partir d'un dataset de 100.000 individus représentant 17 millions de codes-évènements ponctuels, nous obtenons des représentations numériques pour 16.000 codes médicaux issus de 5 nomenclatures nationales et internationales (CIM-10, Classification ATC, Classification Commune des Actes Médicaux, Nomenclature des Actes de Biologie Médicale, Liste des Produits et Prestations Remboursables).

L'étude du voisinage d'un set de codes nous permet de nous assurer que les représentations

*Intervenant

obtenues sont sémantiquement cohérentes. Les codes médicaux liés sémantiquement sont regroupés au sein du même voisinage (au sens de la similarité cosinus). Cette propriété permet d'identifier des clusters homogènes d'évènements médicaux issus de nomenclatures différentes de manière automatique. Par exemple, le voisinage du code diagnostic "Affections de l'iris" est constitué d'un ensemble d'autres codes diagnostics liés aux affections de l'œil (cataracte, kyste rétinien, malformations du cristallin...) ainsi que d'un ensemble d'actes ophtalmologiques (sclérotomie, vitrectomie, plastie de l'iris...).

Conclusion - La factorisation de la matrice PMI permet d'obtenir des représentations numériques sémantiquement cohérentes pour les codes médicaux. Ces représentations permettent de constituer des clusters de codes issus de nomenclatures différentes de manière automatique. Ces vecteurs doivent pouvoir être utilisés à la place de l'encodage one-hot dans une variété de tâches de Machine Learning appliquées aux séquences de soins.