Exploration et système de recommandation pour l'aide au raisonnement médical

Joris Falip, Frédéric Blanchard, Michel Herbin

CRESTIC, Université de Reims Champagne-Ardenne, Reims, France
{joris.falip,frederic.blanchard,michel.herbin}@univ-reims.fr

Résumé :
Avec la multiplication des systèmes d'information liés à la santé, l'analyse *a posteriori* des données médicales est un enjeu important. Les informations et les connaissances potentiellement contenues dans ces gigantesques bases de données sont nombreuses mais nécessitent un travail considérable pour émerger. L'approche classique consiste à définir une problématique médicale, filtrer les données et utiliser un modèle informatique et statistique pour répondre à la question posée. En l'absence d'une problématique clairement posée, il est nécessaire d'explorer, visualiser, observer et comprendre les données pour faire émerger des hypothèses médicales. Nous proposons une solution algorithmique et logicielle qui permet d'effectuer une exploration préliminaire de ces masses de données. Cet outil constitue une sorte de guide pour la « fouille » puis suggère, à la manière d'un système de recommandation, des associations entre patients, en utilisant des indicateurs de pertinence. Notre solution, basée-instances, est centrée sur le patient et permet d'accroître le raisonnement par analogie de l'utilisateur.

Mots-clés : Aide au raisonnement, système de recommandation, exploration, visualisation, données massives, données médicales

1 Introduction

Au cours des dernières décennies, l'informatisation progressive du suivi des patients a conduit les professionnels de santé à alimenter en données des bases dédiées initialement à la mémorisation, à la communication et à la mutualisation des informations. De nombreux médecins ont aujourd'hui conscience du potentiel que représentent ces bases de données en termes de connaissances cliniques. La médiatisation de l'intelligence artificielle et de l'apprentissage automatique notamment, ont fait naître beaucoup d'espoirs quant à l'exploitation a posteriori de ces données.

Les objectifs sont variés et les attentes sont nombreuses. Les experts médicaux qui alimentent les systèmes d'informations aimeraient faire progresser les connaissances sur les pathologies qui les concernent. Chaque spécialiste a ses propres problématiques et ses propres attentes. Et chacune de ces questions devrait faire l'objet d'une analyse *ad hoc* avec un modèle et un choix d'algorithme de résolution spécifiques. Ainsi, il n'existe pas « une » analyse secondaire, mais autant que de questions formulées par les spécialistes. Toutes ont en commun la difficulté de préparer ces données, complexes et brutes.

Au delà du travail de prétraitement difficile et laborieux de ces données massives complexes, il est nécessaire de structurer et d'organiser les données pour permettre eux experts médicaux de mieux les appréhender. Nous proposons, dans ce travail, une méthodologie d'exploration et d'enrichissement des données de santé dont l'objectif est de les rendre exploitables. En corollaire de cette première approche, nous proposons un outil de recommandation permettant d'explorer ces données enrichies, sans *a priori* sur le sujet d'expertise ni sur la finalité. Le principe de cette utilisation est d'étendre le raisonnement par « analogie ». De nombreux systèmes experts utilisant ce mode de raisonnement ont déjà été imaginés mais la disponibilité et le volume des données offrent de nouvelles perspectives. Ils posent aussi de nouveaux problèmes. Notre proposition apporte des éléments de réponse à certains d'entre eux.

Le contexte médical induit des spécificités qui sont le leitmotiv des choix que nous proposons. La médecine étant souvent appelée une « science des cas particuliers », nous avons opté pour une approche basée-instances afin d'éviter tout biais lié à la généralisation, sans devoir utiliser des connaissances *a priori*. Nous associons ensuite les patients en fonction de

leurs spécificités et présentons ces rapprochements sous forme de suggestions et de recommandations. Nous avons donc conçu notre outil comme un moyen d'augmenter, grâce aux données, le raisonnement par analogie. Pour y parvenir nous retrouvons et associons, dans la masse des données, les patients dont les situations sont les plus similaires. Pour calculer cette pertinence, nous proposons des indicateurs de généricité, de spécificité et d'atypicité. En élargissant la contextualisation et en diminuant les risques de biais cognitifs, notre approche permet d'étendre la réflexion des experts médicaux sans prétendre s'y substituer.

Après avoir décrit le contexte et les données qui ont servi de point de départ à ce travail, nous présentons notre méthodologie de structuration et d'association des patients, puis le calcul d'indicateurs de pertinence, pour finir par le mécanisme de recommandation.

2 Contexte, problématique et données

Les données qui nous ont servi de sujet d'expérimentation sont issues du système d'information d'un réseau de santé. Ce système a été initialement conçu pour regrouper les informations recueillies par les différents spécialistes sur des patients diabétiques. Les données anonymisées dont nous disposons concernent près de 1900 patients [1]. Elles contiennent plusieurs centaines de variables (600 environ), de tous types, et sont recueillies en flux d'enregistrements. Hétérogènes et inégalement renseignées, elles relèvent des *big data* de par les difficultés d'analyse qu'elles suscitent.

Les experts médicaux qui aimeraient exploiter ces données ont des objectifs différents et des problématiques médicales variées : comprendre et anticiper la survenue de complications liées au diabète, améliorer et individualiser la démarche thérapeutique, en sont des exemples.

La préparation des données est une étape cruciale et commune à toutes ces problématiques. Ces données arrivent en flux sous forme d'enregistrements liés aux consultations. Elles nécessitent d'être prétraitées et consolidées.

Après ce prétraitement nécessaire, nous proposons de structurer les données dans un graphe orienté. Cette structure de graphe nous permet alors de calculer des indicateurs de pertinence des instances. Cet enrichissement des données permet de guider leur exploration. Enfin, cette exploration est mise en pratique sous forme d'un système de recommandation.

3 Méthodologie proposée

La méthodologie que nous avons choisie repose sur la construction d'une sorte de topologie discrète sur l'ensemble des patients (les instances). L'idée est dans un premier temps d'associer ces instances par paires dans un graphe orienté, en fonction de leurs caractéristiques et de leurs ressemblances. Dans un second temps nous calculons sur cette structure des indicateurs permettant de quantifier la pertinence des instances en terme d'information.

Le choix de l'approche basée-instances a plusieurs buts :

— éviter le paradigme de généralisation et rester centré sur l'individu (le patient),
— éviter l'exclusion des patients les plus atypiques, qui restent porteurs d'informations
— pallier la carence de métrique pertinente en grande dimension, en l'absence de connaissances ou d'expertises *a priori*.

L'objectif des indicateurs de pertinence est d'exhiber les instances les plus informatives. En nous basant sur le raisonnement médical par analogie et en ayant conscience des biais cognitifs inhérents (le biais de disponibilité plus particulièrement), nous proposons de quantifier la pertinence des instances en fonction de leur atypicité et de leur représentativité. L'idée est d'associer à un patient donné, les patients de la base qui lui sont les plus ressemblants et dont les profils sont les plus atypiques et les plus génériques.

1. pour illustrer notre méthodologie, nous nous sommes plus particulièrement intéressés à un sous-échantillon de 196 patients, dont le recueil des informations (notamment sur les mesures d'HbA1C) a fait l'objet d'une attention particulière dans le cadre d'une étude clinique.

3.1 Structuration des données

Après l'étape de préparation, les données sont présentées comme un ensemble d'instances, les patients, décrits à l'aide de plusieurs variables. Notre premier objectif est donc de pouvoir comparer ces patients à l'aide des variables disponibles. Cette étape de comparaison, essentielle, n'est pas simple. Habituellement, une analyse de données requiert une mesure de dissimilarité (le plus souvent une distance) et un espace de description (le plus souvent euclidien). On procède pour cela généralement à une étape préliminaire de *features engineering* qui permet de construire cet espace, et d'y choisir une métrique.

Mais notre système d'exploration se doit d'être aussi générique que possible (tout en restant interprétable). Il ne s'agit pas de cibler une problématique particulière avec une ou plusieurs variables d'intérêt. On ne peut donc pas injecter de connaissance *a priori* en transformant les variables disponibles initialement, en procédant à une réduction de dimension de l'espace de description, ni en émettant d'hypothèses sur la distribution des données. Nous avons donc choisi de garder l'ensemble des variables disponibles au départ et notre approche consiste :

— à traiter indépendamment chacune des variables comme un descripteur particulier et agréger ces informations le plus tard possible dans le processus,
— à définir une topologie discrète sur les instances, et utiliser cet espace relativement « pauvre » en propriétés géométriques pour caractériser et organiser ces instances.

Dans Falip *et al.* (2017), nous avons proposé une stratégie de construction de ce graphe de représentativité, qui est une extension de celle proposée dans Blanchard *et al.* (2014, 2015). C'est la méthode par défaut de l'outil que nous avons conçu.

À cette étape, les données sont associées deux à deux dans un graphe orienté. La construction du graphe dépend donc du voisinage de chaque instance. Ainsi la taille de ce voisinage (le nombre de voisins) influence la structure du graphe. Nous obtenons donc un graphe paramétrique. Nous étudions ensuite l'évolution de la structure en fonction de la taille choisie pour le voisinage, et ce afin de caractériser les instances.

Sur la cohorte réduite à 196 patients, la figure 2 représente ce graphe pour différentes valeur du paramètre de voisinage. Ces graphes illustrent l'influence de ce paramètre sur la granularité des regroupements en composantes connexes. La figure 2 nous montre l'évolution du nombre de composantes connexes en fonction de la taille du voisinage.

3.2 Caractérisation des instances

Nous définissons des indices de singularité et de généricité à partir de notre méthode de structuration précédente. Ces indices, dont le calcul est présenté dans Falip *et al.* (2017); Aït Younes *et al.* (2014), permettent de caractériser la manière dont certaines instances se distinguent des autres, et de quantifier leur capacité à représenter d'autres instances. Le calcul des indices caractérisant les patients se fait en étudiant l'évolution du graphe de similarité, pour toutes les tailles de voisinage possibles. En faisant varier l'unique paramètre de l'algorithme, nous étudions à chaque itération le nombre de patients représentés par cet individu, ainsi que le nombre d'itérations où un individu se choisi lui-même comme représentant.

L'indice de généricité quantifie la capacité à résumer un grand nombre d'instances. Ainsi une instance à la généricité élevée sera représentative de nombreuses instances, tandis qu'une généricité nulle indique une instance n'en subsumant aucune autre.

La singularité peut être interprétée comme l'incapacité d'une instance à être subsumée par une autre. Les instances dotée d'une forte singularité sont entourées d'instances peu représentatives, et n'ont donc d'autre choix qu'être leur propre représentant. C'est le cas pour les instances très génériques, mais aussi à l'inverse pour les instances les plus atypiques qui ne peuvent trouver de représentant du fait de leurs particularités.

3.3 Visualisation et recommandation

Le premier intérêt de cette méthode « d'enrichissement » des données est d'aider à la compréhension des données en exhibant les instances d'intérêt et en visualisant leurs associations.

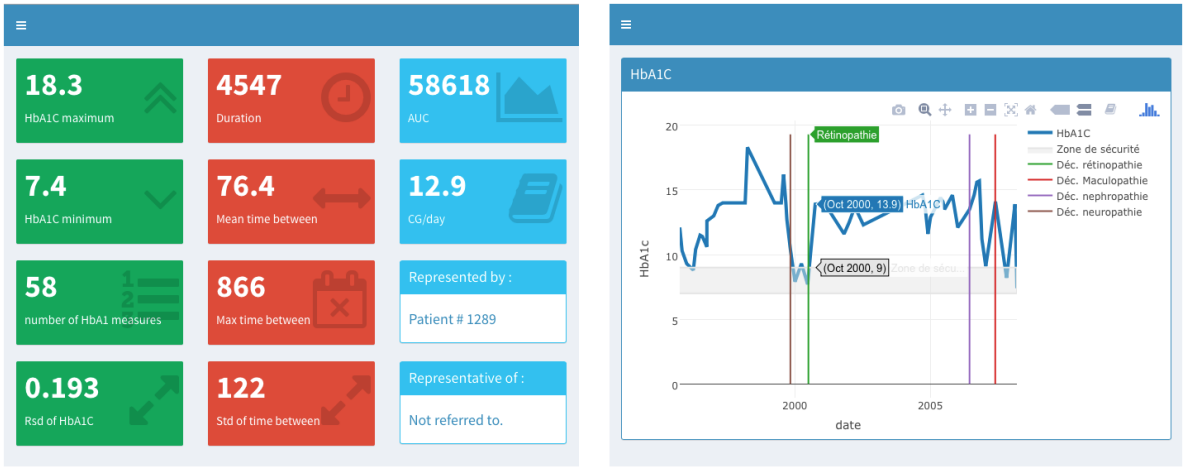

FIGURE *1 – Fiche d'un patient, comportant (A) ses informations de suivi et (B) son historique d'hémoglobine glyquée ainsi que les survenues de complications liées au diabète.*

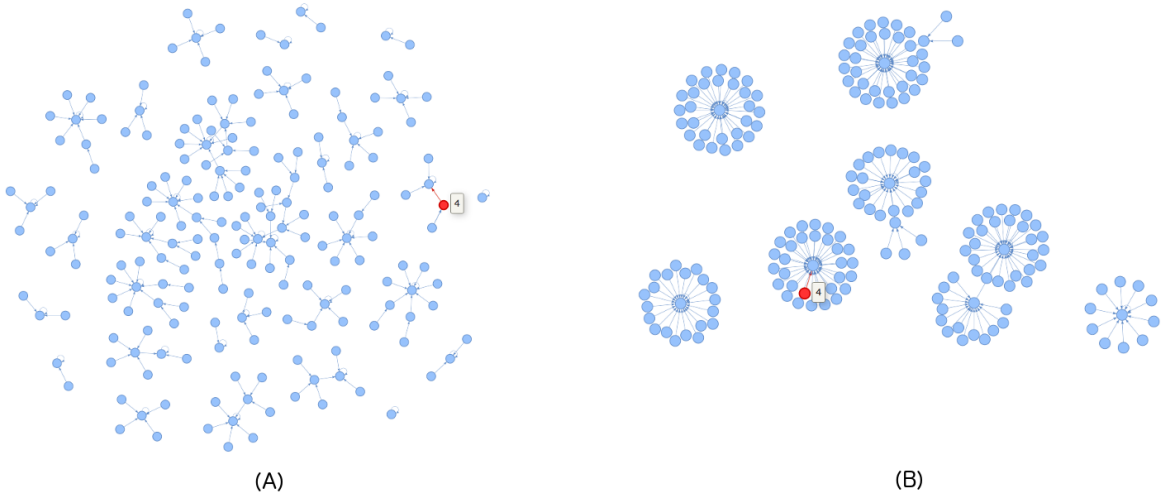

FIGURE *2 – Graphe associant chaque patient à son représentant. Comparaison de l'utilisation d'un voisinage restreint (k = 5) (A) ou étendu (k = 50) (B) pour la découverte de représentants.*

Il permet de guider l'exploration de l'utilisateur.

L'interface d'exploration que nous avons développée propose ainsi de visualiser les patients sur ce graphe et permet de suggérer l'observation des patients « proches » (figure 2). Les suggestions sont formalisées par l'affichage, sur chaque fiche-patient, de liens vers d'autres fiches-patients (figure 1). Notre algorithme « estime » que ces fiches sont suffisamment ressemblantes et présentent suffisamment d'originalité ou de généricité pour être recommandées au médecin. Cette façon de procéder, à la manière d'un système de recommandation, exploite et étend le raisonnement par analogie.

3.4 Perspectives : *feedback* et personnalisation

En filant la métaphore du « système de recommandation », nous proposons de mettre en place un système de *feedback* qui permettra d'adapter l'exploration et la recommandation à la problématique de l'utilisateur. L'idée est de construire, de manière incrémentale, une

sélection d'attributs qui modifiera le processus décrit avant. Plusieurs pistes sont envisagées parmi lesquelles : l'apprentissage d'une métrique primaire sur l'ensemble des instances et définition de voisinages adaptatifs lors de la création du graphe.

4 Conclusion

Nous avons construit un outil d'exploration basé-instances qui guide l'utilisateur dans son parcours et lui suggère des associations pertinentes. L'idée est de solliciter le raisonnement par analogie des experts pour faire émerger de nouvelles connaissances sur les pathologies étudiées. Les caractéristiques de notre solution sont sa plurivalence (aucune hypothèse préalable n'est faite sur les variables de description des instances), et sa capacité à mettre en avant les cas particuliers ou atypiques. Dans la suite de nos travaux nous allons compléter le système en personnalisant les recommandations afin d'adapter les associations aux questions propres à chaque utilisateur.

Références

AÏT YOUNES A., BLANCHARD F., DELEMER B. & HERBIN M. (2014). Singular profile of diabetics. In *International Conference on Innovations for Community Services*, Reims.

BLANCHARD F., AÏT YOUNES A. & HERBIN M. (2014). Structuring complex data using representativeness graphs. In *International Conference on Innovations for Community Services*, Reims.

BLANCHARD F., AÏT YOUNES A. & HERBIN M. (2015). Linking data according to their degree of representativeness (dor). *EAI Endorsed Transactions on Industrial Networks and Intelligent Systems*, **15**(4).

FALIP J., AÏT YOUNES A., BLANCHARD F., DELEMER B., DIALLO A. & HERBIN M. (2017). Visual instance-based recommendation system for medical data mining. In *Knowledge-Based and Intelligent Information Engineering Systems*, p. 1747–1754, Marseille : Elsevier.