Offre de thèse dans le cadre d'un contrat CIFRE

Entreprise/Organisme : Partenariat entre l'équipe HeKA et la société CEMKA

Niveau d'études : Master / Ecole d'Ingénieur

Sujet : Analyse des données du Système National de Données de Santé pour identifier les séquences de soins des patients en errance diagnostique

Date de début : Septembre 2023

Durée du contrat : 3 ans

Rémunération : 33 000 € brut / an

Secteur d'activité : Data Science

Description :

Une des problématiques importantes pour les patients est l'errance diagnostique définie comme le délai entre les premiers signes de la maladie et le diagnostic et qui peut parfois prendre plusieurs années pour les maladies rares. De nombreuses actions sont menées pour diminuer ce délai, mais ces actions sont limitées par le fait que les patients en errance diagnostique sont difficilement identifiables. Nous partons de l'hypothèse pour ce projet que certaines séquences de traitement permettent d'identifier des patients en errance diagnostique de façon suffisamment spécifique et sensible pour pouvoir être utile en pratique pour diminuer le délai d'errance diagnostique. Pour identifier ces séquences, des enjeux méthodologiques se posent car les données du SNDS sont hétérogènes et de très grande dimension : l'extraction d'information de ce grand ensemble de données jusqu'alors peu exploité constitue un défi.

Les questions posées ;

- Quels classifieurs utiliser pour identifier les séquences de traitement les plus discriminantes entre les patients atteints d'une maladie donnée et les autres patients lors de leur parcours de soin avant la prise en charge appropriée?
- Quelles maladies ont des séquences de traitement suffisamment spécifique au cours de la phase d'errance pour être identifiable plus rapidement ?

Un des cas d'usage pourra être le projet Dromos qui vise à étudier le parcours de soin pour 850 maladies rares regroupant plus de 500000 patients en chainant les données de la BNDMR, qui permettront de disposer de la date des premiers signes et la première date d'activité dans un centre de référence, avec les données du système national de données de santé (SNDS) qui permet d'avoir l'historique des consommations de soin de ces patients au cours des 10 dernières années.

1. les méthodes ;

Les classifieurs utilisés doivent se fonder sur des mesures de distance qui intègrent la complexité des données du SNDS à savoir :

- le morcellement des informations : La difficulté dans l'utilisation des données du SNDS pour identifier des parcours de soin est le morcellement des informations. Par exemple, la prise en charge d'une même pathologie peut être identifiée de multiples façons en fonction du fait qu'elle est traitée en ville ou non, par un traitement médicamenteux ou un autre ayant la même indication. Il existe cependant une relation sémantique entre les différentes informations : ainsi un médicament anti-diabétique appartient à la classe

des anti-diabétique et donc son code est relié sémantiquement au code diagnostic du diabète. La prise en compte des relations sémantiques entre les différents types d'information constitue un enjeu majeur pour augmenter l'informativité des données du SNDS. Il est donc nécessaire d'intégrer cette dimension dans un classifieur.

- Le caractère hétérogène des données : Les données présentes dans le SNDS peuvent être des évènements ponctuels ou bien des variables évoluant dans le temps (exemple : dose mensuelle d'un traitement médicamenteux). Il s'agit donc de prendre en compte dans une même distance des données de nature très différente
- La grande dimension des données : Il existe des dizaines de milliers d'actes, diagnostics, médicaments qui peuvent être codés et les cohortes maladies rares sont des cohortes de taille inférieure au nombre de variables disponibles. Il est donc nécessaire de construire le classifieur à partir de méthodes d'apprentissage permettant de réduire la dimension des données (méthodes pénalisées, arbres, …).

Les classifieurs développés seront donc fondés sur des mesures de distance entre les deux groupes permettant de discriminer le plus fortement les deux groupes en explorant différentes pistes pour intégrer la complexité des données issues du SNDS. Les critères de performance évalués seront les critères habituels (rappel, précision, F1-mesure)., Plusieurs groupes témoins seront étudiés : témoins issus de la population générale et témoins ayant d'autres maladies rares du même spectre.

Ces méthodes seront appliquées à plusieurs maladies rares (cas d'usage) afin d'identifier le classifieur le plus généralisable possible.

Nous avons déjà étudié dans des travaux précédents différentes distances pour prendre en compte des données longitudinales issues du SNDS (1) et les liens sémantiques entre médicaments (2).

(1) Lambert, J., Leutenegger, A. L., Jannot, A. S., & Baudot, A. (2023). Tracking clusters of patients over time enables extracting information from medico-administrative databases. Journal of Biomedical Informatics, 139, 104309.

(2) Lambert, J., Leutenegger, A. L., Baudot, A., & Jannot, A. S. (2023). Improving patient clustering by incorporating structured label relationships in similarity measures. medRxiv, 2023-06.

Contact : stephane.bouee@cemka.fr