Sujet de thèse :
Utilisation de méthodes d'apprentissage par renforcement dans le cadre de la recherche médicale

Candidate : Sophia Yazzourh

Ecole Doctorale : Ecole Doctorale 475 - Mathématiques, Informatique, Télécommunications de Toulouse

Laboratoire d'accueil : Institut de Mathématiques de Toulouse

Direction de thèse :

- Nicolas SAVY – Maître de conférences HDR - Institut de Mathématiques de Toulouse
- Philippe SAINT-PIERRE - Maître de conférences - Institut de Mathématiques de Toulouse

---

Description du projet de thèse : La mise en œuvre de techniques d'apprentissage statistique pour traiter les questions de santé a maintenant une longue histoire. Cette histoire montre l'énorme puissance et la polyvalence de ces techniques, mais a également montré plusieurs faiblesses. La question notamment de l'optimisation des séquences de traitement posent de nombreux problèmes. Ce projet se concentre sur l'apprentissage par renforcement typiquement utilisé pour optimiser des stratégies de traitement (DTR – Dynamic Treatment Regimes) et vise à proposer des améliorations à ces méthodes d'apprentissage par renforcement. Le projet se décline en trois axes d'amélioration :

- étudier la pertinence de l'utilisation du deep Q-learning pour appréhender des situations plus complexes que le Q-learning usuellement utilisé pour l'analyse des DTR,
- aborder la question de l'analyse des trajectoires en développant une méthode d'apprentissage hybride intégrant des données et des connaissances d'experts,
- explorer la pertinence d'introduire des méthodes GAN (Generative Adversarial Networks) pour utiliser des techniques DTR dans le contexte des petites bases de données.

Eléments de contexte : L'apprentissage par renforcement est une méthode d'apprentissage qui consiste à apprendre, à partir des données, les actions à prendre afin de maximiser une récompense acquise au cours du temps. Pour ce faire, on considère un agent placé dans un environnement et devant prendre des décisions selon son état actuel. En retour, l'environnement procure à l'agent une récompense (positive ou négative). L'agent cherche alors à optimiser son comportement décisionnel au travers de ses expériences, plus précisément à maximiser la somme des récompenses obtenues.
Dans le cadre de la santé, l'apprentissage par renforcement est moyen efficace de traiter plusieurs aspects de l'analyse des parcours de soins notamment pour les maladies chroniques. En effet, une maladie chronique est caractérisée par une séquence d'observations cliniques ou une séquence de soins administrés. L'utilisation de méthodes d'apprentissage par renforcement permet notamment de

- définir des trajectoires « typiques » et d'en étudier les déterminants,
- définir des trajectoires menant à un état de santé donné et déterminer les facteurs de risque d'être dans un état donné à un moment donné,
- décrire et optimiser les modalités de gestion des traitements adaptés à chacun des patients,
- comprendre et optimiser les soins des patients,
- mesurer l'impact de l'intervention sur le parcours de traitement par simulation dans des conditions données.

L'intérêt de l'apprentissage par renforcement a déjà été démontré, et plus particulièrement la méthode Q-learning a déjà été utilisée dans le contexte médical de traitement à longue durée [1, 2] mais a aussi montré de nombreuses faiblesses [3] dont certains - détaillées ci-dessous - font l'objet de ce projet de thèse. L'objectifs de la thèse est d'apporter des solutions et des pistes d'amélioration de ces techniques

dans le domaine de la santé en s'appuyant sur les innovations récentes des techniques d'apprentissage par renforcement.

Etape 1 du projet de thèse : du Q-learning au Deep-Q-learning. L'apprentissage par renforcement de base est modélisé par un processus de décision de Markov (MDP) impliquant

- **S** un ensemble d'états de l'environnement sur lequel agit un « agent »,
- **A** un ensemble d'actions,
- $\boldsymbol{P_a}(s, s') = P(s_{t+1} = s' \mid s_t = s\,, a_t = a)$ la probabilité de transition (à l'instant t) de l'état s à l'état s' sous l'action a,
- $\boldsymbol{R_a}(s, s')$, la récompense immédiate après le passage de l'état s à l'état s' sous l'action a.

L'« agent » (le patient dans le contexte médical) interagit avec son environnement à des instants discrets. A l'instant t, son état actuel s_t et sa récompense r_t sont collectés. L'agent choisit une action a_t et l'envoie à l'environnement qui passe à l'état s_{t+1} et récompense r_{t+1} selon la transition $\boldsymbol{P_a}$. Un objectif majeur est d'apprendre une politique dite optimale définit par la fonction $\pi *$: $\boldsymbol{A} \times \boldsymbol{S} \rightarrow [0,1]$ qui maximise la récompense cumulée attendue mesurée par la Q-fonction $Q^{\pi}(a, s) = E_{\pi}[R_t \mid a_t = a, s_t = s]$:

$$\pi *(a, s) = \underset{\pi}{\operatorname{argmax}}\, Q^{\pi}(a, s).$$

La méthode la plus courante pour répondre à cette question est le Q-learning basée sur les équations de Bellman. Dans les cas les plus simples (espace d'état et espace d'action discrets), le problème se résume à une recherche d'optimum dans une table mais pour répondre à des situations plus complexes, il est nécessaire d'avoir recours à des modélisations de la Q-fonction. Un des objectifs principaux de la thèse est d'explorer les performances et la faisabilité du Deep Q-learning [4] où la Q-fonction est modélisée par des algorithmes de deep learning. Si ces premières investigations donnent des résultats convaincants, un effort sera porté sur la méthodologie et sur la vulgarisation de la technique.

Etape 2 du projet de thèse : d'une approche « data-driven » à une approche hybride « data / expert - driven ». La méthode Q-learning est une méthode purement « data-driven » dans le sens où elle ne demande pas de modélisation du MDP. C'est une propriété très confortable car elle repose sur aucune hypothèse mais elle est aussi très sensible aux données et à leur représentativité. Ce problème est étroitement lié au dilemme exploration-exploitation bien connu en renforcement learning. En effet, lors du choix entre les différentes options, l'agent est fréquemment confronté à choisir entre :

- quelque chose de familier afin de maximiser les chances d'obtenir ce que vous vouliez,
- quelque chose qui n'a pas été essayé et peut-être en apprendre davantage, ce qui peut (ou non) aboutir à de meilleures décisions à l'avenir.

Ce compromis affectera le fait que l'agent gagne sa récompense plus tôt ou apprenne d'abord l'environnement, puis gagne ses récompenses plus tard. Dans ce contexte des approches « hybrides » incorporant de l'a priori d'expert dans ces méthodes data-driven sont envisagées. Le second objectif de la thèse consiste donc à décrire et à étudier les propriétés statistiques d'approches hybrides mêlant techniques « data-driven » et « expert-driven » dans le contexte des approches DTR par Q-learning.

Etape 3 du projet de thèse : les méthodes de « data augmentation », d'une petite base de données à une grande base de données. Une des limitations de ces méthodes est directement reliée au volume des données. Dans le contexte du parcours médical, il est difficile de disposer d'échantillons impliquant de nombreux patients, les données étant difficiles et couteuses à recueillir (du moins en assurant une qualité suffisante de ces données). Le troisième objectif de cette thèse consiste donc à considérer des techniques dites de « data augmentation » qui consiste à artificiellement augmenter l'effectifs de la base de données en intégrant des patients virtuels obtenus par simulation numérique. Plusieurs techniques ont été explorés par N. Savy et P. Saint-Pierre. Un troisième axe de ce projet de thèse est d'étudier les propriétés des GAN (Generative Adversarial Network [5]) pour la génération de patients virtuels et d'appréhender les propriétés statistiques et computationnelles des approches de reinforcement learning dans ce contexte de données augmentées.

Retombées attendues : Les résultats des travaux présentés ci-dessus qui pourront être valorisés par des publication dans des revues de statistiques, machine learning et biostatistiques. Le souci méthodologique sera au centre de nos préoccupations pour délimiter le contour des applications et pour recenser les hypothèses de ces méthodes. Un effort de vulgarisation sera fait pour rendre accessible ces méthodes complexes à la communauté médicale par le biais de package R ou Python et de communications orales et écrites. Enfin cette thèse est une opportunité pour renforcer l'interaction entre l'IMT et les partenaires de santé en leurs permettant de mettre en œuvre des méthodes innovantes.

Références :

[1] M.R. Kosorok and E.E.M. Moodie (Eds). *Adaptive Treatment Strategies in Practice*. ASA-SIAM Series on Statistics and Applied Mathematics, 2015.

[2] B. Chakraborty and E.E.M. Moodie. *Statistical Methods for Dynamic Treatment Regimes: Reinforcement Learning, Causal Inference, and Personalized Medicine*. Springer. 2013.

[3] Y. Chao, L. Jiming, and N. Shamim. *Reinforcement learning in healthcare: A survey*. arXiv preprint arXiv:1908.08796, 2019.

[4] V. François-Lavet, P. Henderson, R. Islam, M. G. Bellemare and J. Pineau, *An Introduction to Deep Reinforcement Learning*. Foundations and Trends in Machine Learning: Vol. 11, No. 3-4, 2018.

[5] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, *Generative adversarial nets*. Advances in neural information processing systems, pp. 2672–2680, 2014.

Signatures :

La candidate Sophia Yazzourh	Le directeur de Thèse Nicolas Savy	Le co-directeur de thèse Philippe Saint-Pierre	Le directeur du laboratoire Franck Barthe