INP ENSEEIHT

Projet Master : Observabilité partielle et POMDP

Auteur:
BOULET-GILLY Edmond

March 1, 2019

Table des Matières

1 Introduction **1**
1.1 Observabilité Partielle 1
1.2 Observabilité Partielle et Chaîne de Markov 1

2 Idées Saillantes **2**
2.1 Une première modélisation possible 2
2.2 belief-MDP 3

3 Détails Techniques **3**
3.1 1D Espace de Croyance a 2 états POMDP 3
3.2 Établissement d'une Politique 4

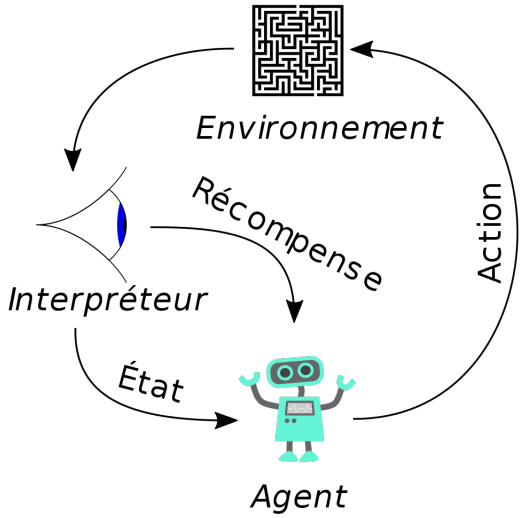
4 Conclusion **4**

5 Références **5**

1 Introduction

En apprentissage automatique, il est question qu'un système puisse apprendre à travers une première phase d'observation d'un environnement dans l'optique que, dans un second temps, ce système soit autonome et prenne lui même ses décisions pour arriver à un but donnée.

L'apprentissage par renforcement met cela en pratique en modélisant le système avec un interpréteur (une entité qui va être capable de lire l'environnement), un agent (l'entité qui va agir sur l'environnement) et un environnement lui même.



Sur cette figure on peut voir que l'agent récupère des information de l'interpréteur. Il récupère des informations sur l'état de l'environnement et des informations sur les récompenses qu'il a généré, indice de succès vis à vis du but de l'intelligence artificielle. A partir de ces informations il va pouvoir générer une politique qui va dicter ces actions de manière à maximum les récompenses.

Il y a donc plusieurs problématiques lors de l'établissement d'une intelligence artificielle à apprentissage récurent :

- La liste des actions
- La quantification des récompenses
- La lecture des états de l'environnement
- La façon dont la politique est mis à jour

1.1 Observabilité Partielle

Suivant le cas de figure d'utilisation ces paramètres sont plus ou moins faciles où quantifier, le cas à l'état qui découle de l'environnement n'est pas quantifiables de manière certaine s'appelle le cas d'observablité partielle.

On le retrouve dans les systèmes où l'environnement est trop important pour être quantifier ou les systèmes où certaines informations ne sont pas observables, comme les cartes d'un adversaire pendant une partie de poker par exemple.

1.2 Observabilité Partielle et Chaîne de Markov

Dans le cas d'un système à États finis on peut modéliser son comportement par une chaîne de markov.

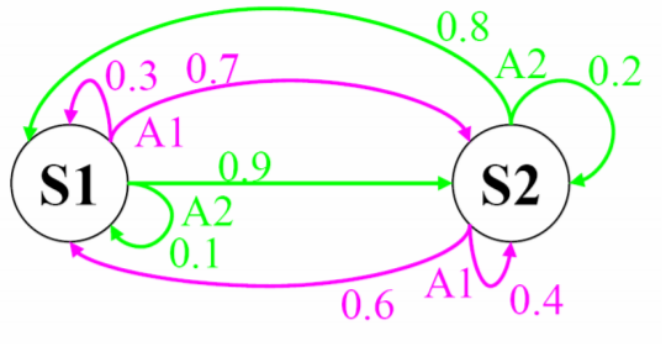


Figure 1: Système à 2 états modélisés par une chaîne de markov

Par exemple sur cette figure nous avons un système à 2 états pour lequel on sait les conséquences (probabilistes) des actions A1 et A2. Si l'on associe en plus à chaque état une récompense, on obtient une modélisation d'un système similaire à celui nécessaire pour utiliser une intelligence artificielle à apprentissage renforcé.

Là où rentre dans le cas d'un POMDP (Processus de décision markovien partiellement observable) est lorsque les états S1 et S2 sont partiellement observable. Par exemple si S1 donne une observation O1 avec ne probabilité de 0.75 et une autre observation O2 de 0.25.

L'avantage du modèle POMDP est le fait qu'une plus grande variété de cas de figure devient modélisable mais en contre partie il devient plus compliqué d'obtenir des politiques performantes donc le système est moins simple à résoudre.

2 Idées Saillantes

La problématique des POMDP est donc d'établir une politique qui va dicter l'action suivante qui se base sur les actions et observations passés sachant que les réactions sur l'environnement sont incertaines mais aussi que l'état courant est basé sur une observation probabiliste.

2.1 Une première modélisation possible

Dans un premier temps, une modélisation possible d'une POMDP possible serait le Tuple : $\langle$S, A, T, R, Ω, O$\rangle$

- **S** : L'espace des états.
- **A** : L'espace des actions.
- **T** : La fonction de transition qui donne la probabilité d'arriver dans l'état suivant en faisant une action donnée dans l'état courant.
- **R** : La fonction de récompense qui donne une valeur de récompense en fonction de l'action faite, et de l'état de départ et d'arrivée.
- Ω : L'espace des observations possibles.

- **O** : La fonction d'observation qui donne la probabilité d'observer o $\in \Omega$ sachant l'état et l'action passé.

Il est possible de définir en plus une variable qui donne la probabilité initiale qui sert à initialiser le système et, dans la plus part des cas étudiables avec ce système, les espaces **S**, **A** et Ω sont finis.

2.2 belief-MDP

Pour compenser les inconnus liés à l'absence de pouvoir observer totalement l'état de l'environnement, il est possible que l'agent se base sur un calcul de probabilité de l'état pour prendre ses décisions. Les état de croyances (Belief state) constitue les état le plus probables du système. On obtient donc un système belief-MDP.

On peut donc utiliser une modélisation du système avec le tuple : $\langle$B, A, T, R$\rangle$, où **B** est l'espace des états dites de croyances.

Un des problèmes majeurs du modèle belief-MDP est que, même si le POMDP de base est a un nombre d'état fini, l'espace des états de croyances peut être continu et donc infini.

3 Détails Techniques

3.1 1D Espace de Croyance a 2 états POMDP

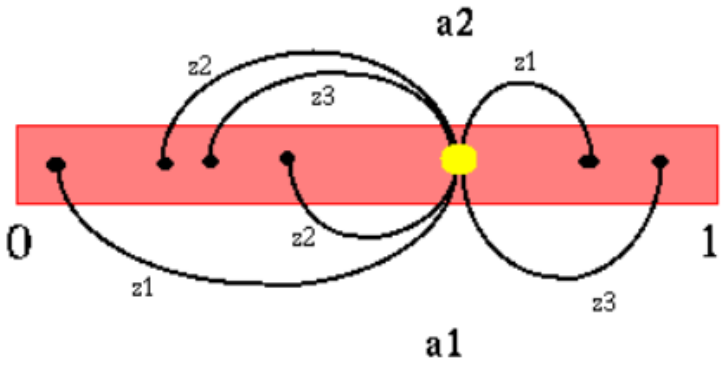


Figure 2: Système à 2 états modélisés par une chaîne de markov

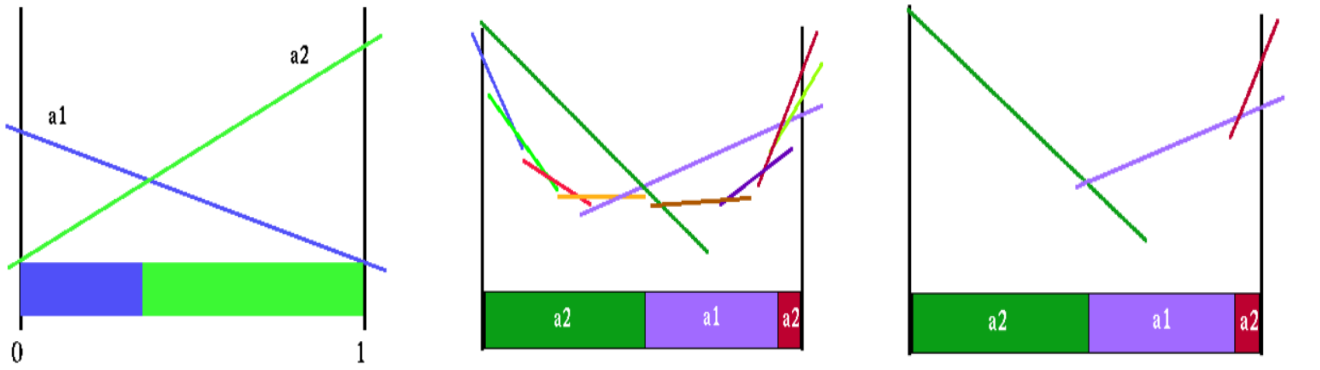
Prenons l'exemple du système à 2 états, sur la figure précédente chaque extrémité du spectre représente un de ces 2 états (on peut aussi voir comme la probabilité p $\in$ [0,1] d'être dans l'état S1 sachant l'état S2 aura par conséquent la probabilité p-1).

En partant d'un état de croyance donné, ici représenté en jaune, on peut effectuer les actions a1 ou a2 et chacune de ces actions a ses observations de l'état d'après associés z1, z2 et z3, représentées en noir. Autrement dit, tant que nous n'avons pas pris d'actions et reçus une observation nous ne pouvoir pas savoir quel sera notre prochain état de croyance. Nous savons justes que qu'il y a une probabilité d'observer z1, z2 ou bien z3 et que la somme de ses probabilités est égal à 1.

3.2 Établissement d'une Politique

La question de l'établissement de la politique se pose donc, quel action faire en conséquence de l'état de croyance actuel avec une connaissance approximative de l'état de croyance suivant ?

Pour cela on établie la fonction de valeur qui donne la valeur d'une action dans un état donné, on obtient une valeur qui quantifie l'impact positif (si valeur est élevée) ou proche de 0 si l'action n'est pas pertinente dans l'état actuel. Ainsi si on reprend l'exemple précédente, on pose que l'action a1 a un impact de 1 dans l'état S1 (côté gauche du spectre) et 0 dans l'état S2 (côté droit). Tant dis que a2 a un impact de respectivement 0 et 1.5 dans S1 et S2. Naturellement plus on est dans un état proche de S1, plus il est intéressant de faire l'action a1 et pareil vis à vis de S2 et a2. C'est exactement ce qu'illustre la figure 3(a) qui montre le recoupement de l'espace de croyance où il est plus intéressant de faire a1 (partie en bleu) ou a2 (partie en vert).



(a) Modèle à 2 actions sans futur (b) Modèle après Entraînement (c) On garde les décisions majorantes

Figure 3: Établissement d'une politique linéaire par morceaux et convexe

Par contre, ce cas de figure peut être bon pour le premier état de croyance rencontré sans être adéquate pour tous les autres. C'est pour cela qu'il est pertinent de réitérer sur plusieurs cas, de manière à entraîner le système, en recalculant à chaque fois l'impact de chaque action en prenant en compte les états et actions précédentes. Cette méthode donne comme résultat la figure 3(b).

Il est normal de retirer toutes les tracés qui ne prennent jamais le dessus sur les tracés supérieurs, qui indiquent un meilleur impact de l'action attitrée. Ce qui donne la figure 3(c) et on obtient une politique qui est pertinente à long terme.

4 Conclusion

Les POMDP permettent d'approcher une variétés importantes de systèmes car, dans les applications concrètes, il est difficile de capturer un état global de l'environnement. De plus les POMDP conservent une partie de l'information lorsqu'une décision est prise quand il y a une incertitude sur l'état courant, cet ajout d'information contribue à posteriori à la price de décision futur.

Un autre aspect des POMPD sont les MOMPD, les procédés à chaînes markoviennes à observabilité mixtes, c'est à dire les cas ou une partie des variables d'environnement sont connues de manières certaines et une autre partie sont partiellement observables.

5 Références

- Partially Observable Markov Decision Processes (POMDPs) par Geoff Hollinger, automne 2007
- Des POMDPs avec des variables d'état visible par Mauricio Araya-López, Vincent Thomas, Olivier Buffet and François Charpillet, 2009
- Les POMDP: une solution pour modéliser des problèmes de gestion adaptative en biologie de la conservation par Iadine Chadès, Josie Carwardine, Tara G. Martin, Samuel Nicol and Olivier Buffet, 2010
- Apprentissage par Renforcement par Wikipedia
- The POMDP Page