Prise de décision séquentielle en environnements incertains et non stationnaires

Emmanuel Hadoux*[1], Aurélie Beynier[1], and Paul Weng[1]

[1]Laboratoire d'Informatique de Paris 6 (LIP6) – Université Pierre et Marie Curie (UPMC) - Paris VI, CNRS : UMR7606 – 4 Place JUSSIEU 75252 PARIS CEDEX 05, France

Résumé

Le modèle des Processus Décisionnels de Markov (MDP) (Puterman94) permet de représenter et de résoudre des problèmes de décision séquentielle dans l'incertain.
Il suppose que l'environnement dans lequel les décisions sont prises est stationnaire.
Cependant, dans la pratique, cela peut ne pas être le cas.
Choi et al. ont proposé les Hidden-Mode MDP (HM-MDP) (Choi99) pour répondre à cette limitation.
Dans leur nouveau modèle, l'environnement est supposé évoluer selon une chaîne de Markov.
Chaque mode m d'un HM-MDP est un MDP défini par le tuple avec S l'ensemble des états, A l'ensemble des actions, T_m la fonction de transition entre les états et R_m la fonction de récompense. L'ensemble des états et l'ensemble des actions sont donc communs pour tous les modes. Un HM-MDP est alors défini par le tuple avec M l'ensemble des modes et C la fonction de transition entre les modes.
Nous proposons les Hidden-Semi-Markov-Mode MDP (HS3MDP) comme extension des HM-MDP pour les cas où l'environnement évolue selon une chaîne semi-markovienne.
Cette hypothèse est d'après nous plus réaliste car l'environnement n'évolue pas forcément à tous les pas de temps.

Un HS3MDP est défini par un tuple avec M et C définis comme précédemment et la fonction H(m, m', h) indiquant la probabilité, après avoir changé de mode de m à m', de rester h pas de temps dans le nouveau mode m'.

L'un des problèmes de la littérature est celui de la gestion d'ascenseurs.

Dans ce problème, l'ensemble des états représente toutes les combinaisons possibles de positions des ascenseurs ainsi que de l'état des boutons d'appel (à l'intérieur) et de sélection d'étages (à l'extérieur).

Les actions sont monter, descendre et ouvrir les portes.

La fonction de transition entre les états est dépendante des probabilités d'arrivée de personnes à chacun des étages. Cette probabilité est modifiée en fonction des différentes heures de pointes, des réunions inattendues, etc. représentées par les modes.

Les fonctions de récompenses engendrent un coup pour chaque utilisateur dont la requête

*Intervenant

n'est pas satisfaite.
C et H sont définies suivant les dynamiques du problème.

Les HM-MDP et les HS3MDP sont des sous-classes des MDP partiellement observables (Puterman94) et peuvent donc être résolus en utilisant les méthodes déjà existantes.
Cependant, ils souffrent de la même malédiction de la dimension que les POMDP.
Nous nous sommes donc intéressés à la résolution approchée des HS3MDP en utilisant POMCP (Silver10) l'un des meilleurs algorithmes de résolution approchée des POMDP à ce jour.
Nous l'avons adapté en exploitant la strucutre particulière des HS3MDP afin d'en améliorer les performances et nous avons expérimenté POMCP original et adapté sur différents problèmes non stationnaires de la littérature.

Mots-Clés: décision séquentielle dans l'incertain, environnements non stationnaires