Intégration de recommandations simples dans un MDP

Florian Benavent, Bruno Zanuttini

Normandie Université, GREYC, CNRS UMR 6072, F-14032 Caen, France
prénom.nom@unicaen.fr

Nous nous intéressons aux recommandations d'un utilisateur à un système autonome. En pratique, on souhaite que l'utilisateur puisse spécifier au système des modifications ponctuelles du monde, ou ses propres préférences. Par exemple, dans le cas d'un G.P.S, l'utilisateur peut souhaiter lui indiquer des embouteillages, des travaux, ou une préférence sur les routes à prendre. Nous souhaitons avant tout que l'utilisateur puisse fournir de telles recommandations sans avoir besoin de représenter en détail ses connaissances sur le monde.

Notre but est de permettre à l'agent d'estimer le modèle sous-jacent de l'utilisateur, à partir de recommandations simples de ce dernier, et d'en déduire une politique d'actions répondant à l'objectif du système tout en satisfaisant les recommandations.

Ce problème peut être vu comme un problème d'apprentissage par renforcement inverse, où nous cherchons à déterminer la fonction de récompense de l'utilisateur au travers d'informations qu'il nous donne [Abbeel & Ng, ICML 2004, Ramachandran & Amir, IJCAI 2007]. Il existe également de nombreuses études cherchant à simplifier le transfert d'informations utilisateur-agent, en particulier à partir de démonstrations sous-optimales [Brys *et al.*, IJCAI 2015] et de renforcements locaux [Knox & Stone, *Artif. Intell.* 2015]. Plus proches de nos travaux, on trouve dans la littérature récente de nombreuses approches pour prendre des décisions à partir d'informations partielles sur la fonction de récompense de l'utilisateur [Regan & Boutilier, UAI 2009, Fürnkranz *et al.*, *Machine Learning* 2012].

Les premiers de ces travaux nécessitent que l'utilisateur aient une idée assez précise d'une politique d'actions, et les seconds demandent un compromis entre la qualité de la politique calculée par l'agent et la quantité d'intéractions utilisateur-agent. Notre approche se distingue par le fait que nous étudions des recommandations extrêmement simples en terme de quantité d'informations, au prix d'hypothèses plus fortes sur la similitude entre le modèle que l'agent a de la tâche, et le modèle (sous-jacent) de l'utilisateur.

Recommandation d'évitement

De façon générale, nous supposons une modélisation d'une tâche sous la forme d'un processus de décision markovien $M = (S, A, T, R)$, avec un état de départ s_0 et un but g, et notons π la politique optimale pour ce processus et V_π^M la fonction de valeur associée dans M. Nous distinguons alors deux MDP, celui de l'agent (M^A) et celui de l'utilisateur (M^U), sur les mêmes ensembles d'états S et d'actions A. L'idée est que l'agent a un modèle de la tâche, et qu'un modèle différent sous-tend les souhaits et/ou connaissances de l'utilisateur concernant cette tâche.

Par souci de simplicité, nous présentons ici un type particulier de recommandation, la recommandation d'*évitement*. Cette recommandation indique au système que l'utilisateur souhaite une politique d'actions évitant à un certain degré un ou plusieurs états, dits *répulsifs*. Nous supposons alors que le modèle M^A de l'agent est correct, c'est-à-dire égal à M^U, sauf en la récompense affectée à ces états. Sur l'exemple du G.P.S., nous formalisons ainsi des cas d'utilisations tels que celui d'un utilisateur souhaitant aller de Lille à Marseille en évitant Paris et Lyon.

Nous considérons ici un seul état répulsif, noté s_m. Une recommandation évidente consisterait, pour l'humain, à donner à l'agent la valeur (négative) de s_m dans son modèle M^U, mais nous ne supposons pas qu'il connaît cette valeur, et souhaitons lui permettre de communiquer sa recommandation avec un minimum d'interaction. Pour cela, nous permettons qu'il donne une distance à garder avec l'état répulsif. Il s'agirait d'indiquer à son G.P.S., par exemple, la ville de Paris (comme état répulsif s_m) et la ville de Melun (comme *point de passage* s_p). L'interprétation que nous plaçons sur une telle recommandation est que l'action à effectuer à Melun devra l'éloigner de Paris, ou encore que Melun est la ville la plus proche (localement) de Paris par laquelle il accepte de passer.

Toutefois, dans un environnement stochastique, on ne peut pas nécessairement s'éloigner d'un état avec

probabilité 1, c'est pourquoi nous interprétons la *recommandation d'évitement* (s_m, s_p) par :

$$d(s^*, s_m) \geq d(s_p, s_m)$$

où d est la distance utilisée sur S (euclidienne par exemple, dans le cas du G.P.S.), et s^* le successeur *préféré* de s_p, défini par :

$$s^* = \arg\max_{s'} T(s_p, a, s')V(s')$$

où a (resp. V) est l'action (resp. la fonction de valeur) optimale pour l'humain (dans M^U). Intuitivement, le successeur préféré de s via a est le successeur sur lequel on « parie » lorsqu'on choisit l'action a dans s.

Pour résumer, une recommandation d'évitement est la donnée d'un état à éviter s_m et d'un point de passage s_p « localement maximalement proche » de s_m, en ce sens que l'action optimale, pour l'humain, en s_p, est motivée par le désir de s'éloigner de s_m.

Travaux en cours

Notre objectif est que l'agent, ayant pour données son propre modèle M^A et une recommandation (s_m, s_p), calcule une politique qui soit aussi proche de l'optimal que possible dans le modèle de l'humain M^U. Par nos hypothèses sur la différence entre M^A et M^U, il suffit pour cela que l'agent soit capable de retrouver la récompense immédiate $R^U(s_m)$. La recommandation n'est bien entendu pas assez précise dans le cas général, mais nous pouvons montrer, avec des calculs simples, qu'elle permet d'en calculer un minorant (un majorant étant donné par la récompense que l'*agent* attribue à s_m, par hypothèse). Le calcul de ce minorant (et donc d'un encadrement) requiert essentiellement un calcul de la fonction de valeur optimale dans le modèle de l'*agent*, M^A, pour les successeurs possibles de s_p.

Il s'agit donc de déterminer si l'encadrement ainsi obtenu est précis, en ce sens qu'il permet de retrouver une politique de très bonne qualité pour M^U en pratique. Nous menons actuellement un ensemble d'expérimentations visant à évaluer cette précision. Pour calculer une politique à partir d'une récompense ainsi encadrée, nous considérons le minorant lui-même ainsi que la politique qui minimise le regret maximum dans l'encadrement [Regan & Boutilier, UAI 2009]. Parmi les indicateurs, nous examinons en particulier si la politique ainsi calculée à partir de l'encadrement a la même suite de successeurs préférés de s_0 à g : il s'agit donc de vérifier que la recommandation suffit à retrouver une politique qui « parie » sur la même trajectoire que la politique optimale (dans M^U), en ignorant les divergences sur les états peu probables ou de faible valeur. Les premiers résultats expérimentaux sont prometteurs.

Par ailleurs, la notion de recommandation « simple » peut se décliner de nombreuses manières. Nous considérons en particulier des recommandations plus générales, des recommandations d'une action donnée dans un état donné, des recommandations d'attirance, etc. De façon plus générale, on peut voir les recommandations comme des *contraintes* sur la politique optimale dans le modèle de l'agent, ce qui permet d'envisager un algorithme générique pour l'intégration.