General Game Learning using Transfer Knowledge

Par Bikramjit Banerjee et Peter Stone

I) Problème général

Dans le cadre de GGP (General Game Playing), les auteurs de l'article présentent un agent joueur utilisant un apprentissage par renforcement dont la particularité est de réutiliser des connaissances apprises contre un adversaire. Le but est de montrer que ces connaissances permettent de favoriser la victoire contre le même adversaire dans un autre jeu.

II) Méthode présentée

L'apprentissage par renforcement se déroule dans un MDP (Markov Decision Process). Le but est grossièrement de réussir à apprendre la fonction action-valeur $Q(s,a) \leftarrow Q(s,a) + \alpha[r_{sa} + \max(b)\ \gamma\ Q(s',b) - Q(s,a)]$ pour avoir la meilleure récompense possible pour chaque état en fonction d'une action.

Dans GGP, l'environnement est représenté par le maître du jeu. L'interaction se fait comme dans un MDP, la majeure différence étant que l'on ne reçoit une récompense que lors de la fin de la partie. Il est à noter que les auteurs travaillent non pas en terme de couples état/action, mais post-états, qui sont plus expressifs dans le cadre des jeux opposant deux adversaires. Les auteurs proposent donc d'appliquer la méthode TD(λ) pour mettre à jour les valeurs $Q(\sigma)$.

La contribution majeure des auteurs est l'extraction des connaissances d'une partie pour les réutiliser par la suite. A chaque post-état, on explore l'arbre des coups possibles (on se fixe une profondeur max de 2). Chaque nœud est classé selon son état (victoire, défaite …) et on regroupe les nœuds identiques dans une même branche. Un arbre sera considéré comme étant une connaissance à apprendre à partir du moment où l'un de ses nœuds est terminal.

On compare ensuite chaque connaissance avec tous les post-états σ. Si on trouve une équivalence, la valeur de la connaissance $val(Fi) = avg_w\{Q(\sigma) \mid \sigma \text{ matches } Fi\}$. Toutes ces valeurs apprises seront ensuite réutilisées dans d'autres jeux contre le même adversaire. Il suffit d'initialiser chaque $Q(\sigma)$ dont σ est équivalent à une connaissance Fi avec le val(Fi).

III) Résultats

L'agent joueur construit ses connaissances à partir du jeu Tic-Tac-Toe, puis il est confronté et évalué sur trois jeux à chaque fois contre trois joueurs aux comportements différents : un aléatoire, un faible (ne choisira jamais le meilleur choix) et un ε-greedy (choisit toujours le meilleur chemin dans un cas critique).

Deux autres agents joueurs sont évalués. L'un identique à l'agent testé sauf qu'il n'implémente pas le transfert de connaissance, et l'autre dont l'apprentissage se fait par une recherche en profondeur en se servant de minmax pour estimer la valeur d'un post-état.

Les résultats des expérimentations sont partagés. Si on compare les résultats obtenus entre les différentes méthodes contre le joueur ε-greedy (sur n'importe quel jeu), celle avec la recherche en profondeur sera toujours plus efficace. Cependant, contre les deux autres joueurs la méthode présentée par les auteurs donne de meilleurs résultats. Ce phénomène est principalement du à la stratégie appliquée par l'agent minmax, où il fait l'hypothèse que le joueur adverse prendra toujours la meilleure solution.

Les auteurs concluent sur le fait que leur méthode est finalement plus efficace étant donné qu'il suffit que l'heuristique implémentée soit fausse pour que l'agent minmax donne de moins bons résultats. La méthode utilisant le transfert des connaissances est beaucoup plus générale.

IV) Etat de l’art

L’extraction de connaissances a déjà été abordée par Fawcet. Asgharbeygi avait aussi développé une technique basée sur la méthode TD dans le cadre du transfert de connaissance, cette fois en manipulant des prédicats de la logique du premier ordre. La différence est que la méthode des auteurs reste plus générale car ils n’ont pas besoin de définir des caractéristiques spécifiques au jeu.

V) Perspectives

Les perspectives évoquées par les auteurs correspondent surtout à tester leur méthode sur des jeux plus complexes et à améliorer la modélisation du transfert en intégrant des connaissances plus profondes.

VI) Commentaires personnels

L’article pose un problème intéressant : pouvoir apprendre les stratégies d’un adversaire, son comportement, au cours de plusieurs parties, puis de se servir de ses connaissances et les réutiliser dans des jeux différents. Cependant, j’ai trouvé les résultats peu concluants. Un premier problème est que l’agent joueur n’est pas évalué contre des adversaires aux comportements plus complexes. Mais le véritable grief est la comparaison avec une méthode plus classique d’apprentissage utilisant une recherche en profondeur et minmax. Il a été montré qu’avec la bonne heuristique, la méthode proposée par les auteurs était inférieure. Le fait est qu’elle est applicable dans n’importe quel cas car on n’a pas besoin de spécifier des caractéristiques liées au jeu. Je pense cependant qu’une méthode réellement efficace serait de trouver un juste milieu entre la généralisation proposée par les auteurs et une spécification à base d’heuristiques.