

Apprentissage par renforcement :

Systemes multi-agents

Frédéric Garcia



Unité de Biométrie et Intelligence Artificielle
BP 27, 31326 Castanet-Tolosan

Apprentissage dans les systèmes multi-agents

- Domaine qui a émergé il y a une dizaine d'années dans les communautés Apprentissage et SMA, très en vogue actuellement
 - special issue on Multiagent Learning, Machine Learning 33(2-3) 1998
- Il existe un fort besoin de techniques et méthodes d'apprentissage en SMA, du fait de la complexité de conception de ces systèmes.
- Passer du cadre d'agent unique à celui de société d'agents enrichit considérablement la problématique de l'apprentissage.

Interagir pour apprendre

un agent influence l'apprentissage d'un autre agent

- imiter un comportement,
- partager les expériences,
- contraindre les expériences ...

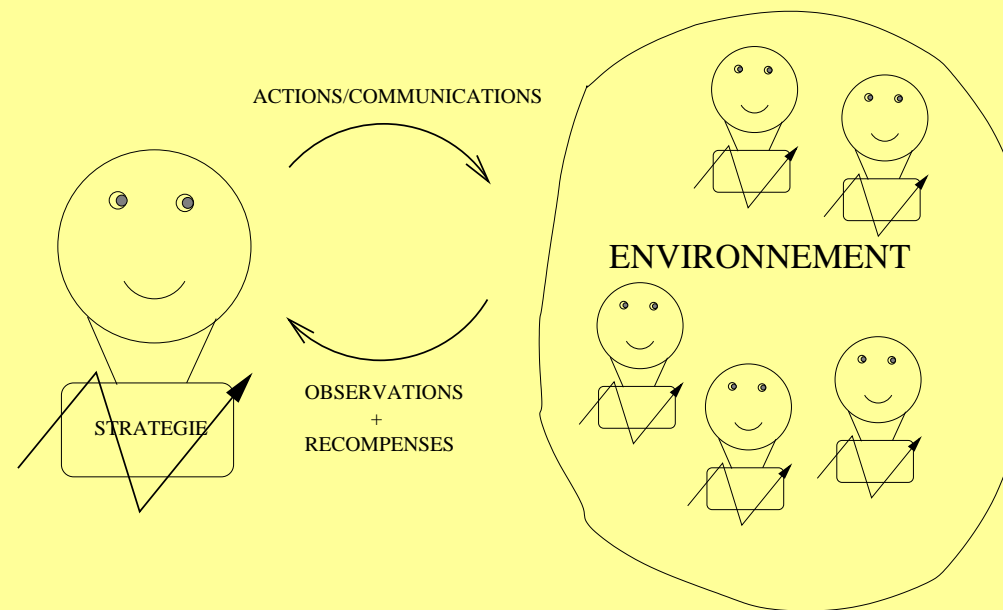
Apprendre à interagir

- quand communiquer ? avec qui ? comment ?
- comment se coordonner ?
- quelle organisation sociale ? ...

Des questions riches et complexes qui ne se posent pas dans un contexte d'agent unique

Paradigme de l'AR multi-agents

Un ensemble d'agents autonomes agissant au sein d'un environnement, qui recherchent au travers d'expériences itérées un comportement décisionnel optimal.



Jeux markoviens

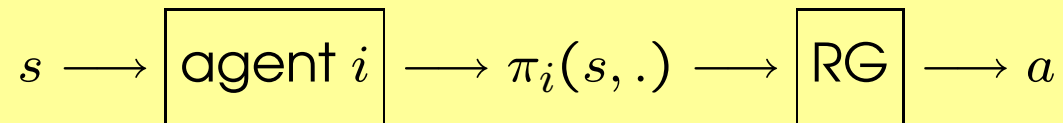
Une généralisation naturelle des PDM au cadre multi-agents

- un espace d'état S
- des espaces d'actions A_1, \dots, A_k pour chacun des k joueurs
- k fonctions de récompences $r_i : S \times A_1 \times \dots \times A_k \mapsto \mathbb{R}$
- des probas de transition $P(s' | s, a_1, \dots, a_k)$
- chaque joueur cherche à maximiser $E[\sum_{t=0}^{\infty} \gamma^t r_{i,t}]$

Politiques stochastiques

Chaque agent $i = 1, \dots, k$ suit une politique stochastique π_i

$\pi_i(s, a)$ représente la probabilité pour l'agent i de choisir l'action a dans l'état s .



Fonctions de valeur

Soit $\Pi = \{\pi_1, \dots, \pi_k\}$ les politiques stochastiques des agents

La valeur espérée de Π pour l'agent i (en s) est

$$V_i^\Pi(s) = \sum_{a_1, \dots, a_k} \pi_1(s, a_1) \dots \pi_k(s, a_k) Q_i^\Pi(s, a_1, \dots, a_k)$$

avec

$$Q_i^\Pi(s, a_1, \dots, a_k) = r_i(s, a_1, \dots, a_k) + \gamma \sum_{s' \in S} P(s' | s, a_1, \dots, a_k) V_i^\Pi(s')$$

Différents types de problèmes

- De quelle connaissance les agents disposent-ils ?
- Qu'observent-ils ?
- quelles types de politiques peuvent-ils suivre ?

L'AR s'intéresse au cas où les agents ne connaissent ni P ni R

L'agent i peut observer s, a_i, r_i , et parfois les $a_{j \neq i}$ et $r_{j \neq i}$

Changements vis à vis du modèle MDP classique

Apprentissage par renforcement non-stationnaire pour un agent donné

Les observations d'un agent peuvent être partielles

On recherche des équilibres entre politiques

L'exploration influence le résultat de l'apprentissage

Jeux à 2 joueurs à somme nulle

Actions simultanées : les politiques optimales sont stochastiques

Optimisation du "cas pire"

$$V^*(s) = \max_{\pi \in \Pi(A)} \min_{o \in O} \sum_{a \in A} \pi_a Q^*(s, a, o)$$

avec

$$Q^*(s, a, o) = r(s, a, o) + \gamma \sum_{s'} P(s' | s, a, o) V^*(s')$$

Algorithme *minimax* Q-learning (Littman 94)

Observation de $\langle s_n, a_n, o_n, s_{n+1}, r_n \rangle$

$$Q_{n+1}(s_n, a_n, o_n) \leftarrow Q_n(s_n, a_n, o_n) + \alpha_n(r_n + \gamma V_n(s_{n+1}) - Q_n(s_n, a_n, o_n))$$

avec V_n obtenue par PL à partir de Q_n

$$V_n(s) = \max_{\pi \in \Pi(A)} \min_{o \in O} \sum_{a \in A} \pi_a Q_n(s, a, o)$$

Convergence vers Q optimale prouvée.

Meilleurs résultats que Q-learning.

Jeux successifs

Un joueur après l'autre : un cas classique en AR (backgammon, dames, échecs, etc.).

Approche MDP classique, apprentissage d'une fonction de valeur de l'état $V(s)$, puis choix de l'action $\pi(s)$ par recherche minimax

$$V^*(s) = \max_{a \in A} \min_{o \in O} Q^*(s, a, o)$$

(politiques déterministes)

Jeux à somme quelconque

Généralisation du cas précédent.

On recherche des politiques optimales “en équilibre”

Equilibre de Nash

$\Pi = \{\pi_i\}$ est un équilibre de Nash ssi

$$V_i^\Pi(s) = \max_{\pi} V_i^{\Pi_{-i} \cup \pi}(s) \quad \forall i \quad \forall s$$

le comportement de chaque agent est optimal étant donné le comportement des autres agents.

Il peut exister plusieurs équilibres de Nash

Exemple d'équilibre de Nash

$$\begin{array}{c} u \quad v \quad w \\ a \left(\begin{array}{ccc} 1 & -2 & 4 \\ 0 & 1 & 1 \end{array} \right) \quad a \left(\begin{array}{ccc} 2 & 1 & 0 \\ 0 & -3 & 2 \end{array} \right) \\ b \end{array}$$

(a, u) est un équilibre déterministe pour ce jeu

Nash Q-learning (Hu & Wellman 98)

Chaque agent i observe $\langle s_n, a_n^1, \dots, a_n^k, s_{n+1}, r_n \rangle$

$$Q_{n+1}^i(s_n, a_n^1, \dots, a_n^k) \leftarrow (1 - \alpha_n)Q_n^i(s_n, a_n^1, \dots, a_n^k) + \alpha_n(r_n^i + \gamma V_i^\Pi)$$

où V_i^Π est la valeur pour l'agent i de l'équilibre de nash Π pour le jeu

$$(Q_n^1(s_{n+1}), \dots, Q_n^k(s_{n+1}))$$

Certaines hypothèses permettent de prouver la convergence de cet algorithme.

Problème de sélection en cas de multiplicité des équilibres.

Adversarial equilibria

chaque agent préfère ne pas changer

$$V_i^\pi = \max_{\pi} V_i^{\pi_{-i} \cup \pi} \quad \forall i$$

chaque agent aimerait que les autres changent ...

$$V_i^\pi \leq V_i^{\pi'_{-i} \cup \pi_i} \quad \forall \pi', \forall i$$

Si chaque équilibre utilisé dans le Nash Q-learning est de ce type, l'algorithme converge vers un équilibre optimal.

Exemple : jeux à 2 joueurs à somme nulle.

Coordination equilibria

chaque agent préfère ne pas changer

$$V_i^\Pi = \max_{\pi} V_i^{\Pi_{-i} \cup \pi} \quad \forall i$$

ce qui est bon pour un est bon pour tous ...

$$V_i^\Pi(s) = \max_{a_1, \dots, a_k} Q_i^\Pi(s, a_1, \dots, a_k) \quad \forall i \quad \forall s$$

Si chaque équilibre utilisé dans le Nash Q-learning est de ce type, l'algorithme converge vers un équilibre optimal.

Exemple : jeux d'équipe markoviens.

Jeux d'équipe markoviens

Une unique fonction $r(s, a_1, \dots, a_k)$

$$Q^1 = Q^2 = \dots Q^k$$

$$a_1^*, \dots, a_k^* = \operatorname{argmax}_{a_1, \dots, a_k} Q(s, a_1, \dots, a_k)$$

est un équilibre de coordination.

On retrouve un MDP (*Multi-agents decision process*)

Team Q-learning (Littman 2001)

Chaque agent i observe $\langle s_n, a_n^1, \dots, a_n^k, s_{n+1}, r_n \rangle$

$$Q_{n+1}^i(s_n, a_n^1, \dots, a_n^k) \leftarrow (1 - \alpha_n) Q_n^i(s_n, a_n^1, \dots, a_n^k) + \alpha_n (r_n + \gamma \max_{a_1, \dots, a_k} Q_n^i(s_{n+1}, a_1, \dots, a_k))$$

Converge vers Q optimal.

Independent Learner (Claus and Boutillier 98)

Etudes menées dans le cas $|S| = 1$

l'agent i observe $\langle a_n^i, r_n \rangle$

$$Q_{n+1}^i(a_n^i) \leftarrow (1 - \alpha_n)Q_n^i(a_n^i) + \alpha_n r_n$$

(Q-learning)

Exploration probabiliste des a_n^i selon Q_n^i .

Joint Action Learner (Claus and Boutillier 98)

l'agent i observe $\langle a_n^1, \dots, a_n^k, r_n \rangle$

$$Q_{n+1}^i(a_n^1, \dots, a_n^k) \leftarrow (1 - \alpha_n)Q_n^i(a_n^1, \dots, a_n^k) + \alpha_n r_n$$

Maintien d'un modèle des autres agents $P^i(a^j)$ pour $j \neq i$

Exploration probabiliste des a_n^i selon

$$\sum_{a^1, \dots, a^{i-1}, a^{i+1}, \dots, a^k} P^i(a^1) \dots P^i(a^k) Q_n^i(a_n^1, \dots, a_n^k)$$

Propriétés de IL et JAL

Sous certaines hypothèses d'exploration, la probabilité de jouer un équilibre tend vers 1 avec n .

Expérimentalement, JAL est un peu meilleur que IL.

La décroissance de l'exploration rend de plus en plus difficile les sauts d'un équilibre à l'autre

Convergence non assurée vers un équilibre optimal

Amélioration possible de JAL en modifiant sa fonction d'exploration

Entre jeux d'équipe et jeux à somme nulle

Exemple du dilemme du prisonnier

$$|S| = 1$$

Q-learning classique (Crites 96)

Observation de $\langle a_n, o_n, r_n \rangle$

- apprentissage machine contre stratégie fixée ou machine
- différents états et représentations de Q
- différents modes d'exploration

Difficulté générale à coopérer

Théorie économique de l'AR

objectif d'explication des comportements observés (Roth and Erev 95)

Choix probabiliste des actions, renforcement des probabilités des actions qui réussissent

Pas de maintien de modèles

Les équilibres de Nash décrivent mal les comportement observées.

AR and the El Farol Model (Franke 99)

N personnes décident indépendamment chaque jeudi soir d'aller ou non au bar "El Farol".

Les personnes présentes sont satisfaites de la soirée si le bar n'est pas plein, c'est à dire s'il y a moins de B ($0 < B < N$) clients. Sinon elles préfèrent rester à la maison.

On modélise le comportement d'un agent par une probabilité de sortir.

L'apprentissage par renforcement permet de modéliser la manière dont ces probabilités évoluent au cours des expériences.

Conclusion

Apprentissage par renforcement multi-agents :

- Problème de plus en plus étudié dans la communauté AR
- Quelques résultats, mais pas encore de cadre formel définitif
- Liens à creuser avec les théories économiques