Kooperáció és intelligencia

Tanulás többágenses szervezetekben/3

MARL – Multi Agent Reinforcement Learning
Többágenses megerősítéses tanulás: áttekintés

Kezdjük 1 db ágenssel. Legyenek a környezet állapotai x-ek, ágens cselekvései u-k, ágens cselekvéseit meghatározó eljárásmód h, ill. az ágens cselekvés-érték függvénye Q(x,u). Az állapotok és a cselekvések közötti kapcsolatot az un. **Markov döntési folyamat** írja le (átmenet-valószínűségek).
Egy ágenses megerősítéses tanulásnál egy ágens

$$R_k = E\left\{\sum_{j=0}^{\infty} \gamma^j r_{k+j+1}\right\}$$

diszkont hátralévő jutalmat maximálja, ahol γ a diszkont faktor és r a megerősítés.
Adott eljárásmód mellett az ágens

$$Q^h(x,u) = E\{\textstyle\sum_{j=0}^{\infty} \gamma^j r_{k+j+1} \mid x_k = x, u_k = u, h\}$$

cselekvés-érték függvényt tanul.

A lehető legjobb eredmény az optimális cselekvés-érték függvény:

$$Q^*(x,u) = \max_h Q^h(x,u)$$

ami teljesíti az un. **Bellman egyenletet**:

$$Q^*(x,u) = \sum_{x' \in X} f(x,u,x')\left[\rho(x,u,x') + \gamma \max_{u'} Q^*(x',u')\right]$$

Az ágens **mohó** eljárásmódja: $\bar{h}(x) = \arg\max_u Q(x,u)$

ami optimális, ha Q is optimális. A Bellman-egyenlet un. **Q-tanulással** oldható meg (jelen formában **időkülönbség Q-tanulás**):

$$Q_{k+1}(x_k,u_k) = Q_k(x_k,u_k) + \alpha_k\left[r_{k+1} + \gamma \max_{u'} Q_k(x_{k+1},u') - Q_k(x_k,u_k)\right]$$

Q-tanulás bizonyos feltételek mellett optimális Q-hoz konvergál.
A feltételek közül a legfontosabb, hogy a **tanuló ágensnek véges nem nulla valószínűséggel ki kell próbálni minden létező cselekvését**.
Nem tud tehát csak mohó lenni, a mohóságát **felfedezési** igénnyel kell vegyítenie.

A mohóság + felfedezés keverék viselkedést biztosítani tudjuk:
- **ε-mohósággal**: az ágens ε valószínűséggel véletlen cselekvést választ, ill. 1-ε valószínűséggel mohó, vagy
- **Boltzmann-felfedezési modellel**, ahol egy u cselekvés megválasztásának valószínűsége egy x állapotban:

$$h(x,u) = \frac{e^{Q(x,u)/\tau}}{\sum_{\tilde{u}} e^{Q(x,\tilde{u})/\tau}}$$

a T „hőmérséklet” a két véglet között szabályoz.
ha $T \to \infty$, akkor a választás tisztán (egyenletesen) véletlen,
ha $T \to 0$, akkor a választás mohó.

Többágenses esetben Markov döntési folyamat helyett a modell un.

Sztochasztikus Játék (Stochastic Game, SG), ahol az állapotátmeneteket az összes ágens **együttes cselekvése** határozza meg, és ahol az egyedi ágensek eljárásmódjai mellett beszéljünk a **együttes eljárásmódról** is.

Jelölje egy-egy ágens megerősítését generáló függvényt ρ_i. Beszélhetünk akkor

- **teljesen kooperatív** ágens rendszerekről $\rho_1 = \cdots = \rho_n$

- **teljesen versengő** ágens rendszerekről, ill. $\rho_1 = -\rho_2$ (két ágens esetén) és $\rho_1 + \rho_2 + \ldots \rho_n = 0$, több ágens esetén

- **vegyes** ágens rendszerekről (ahol semmilyen feltétel nem adható).

Többágenses megerősítéses tanulás problémái:

- mi a cél?
- nem stacionárius (l. előbbi előadás)
- koordinálás igénye

Általában az SG bekényszerítése valamiféle egyensúlyi helyzetbe,
tipikus választás a Nash-egyensúly.
Mindenki tartson ehhez, akkor nem lesz probléma.

Cél:
- **stabilitás** (konvergencia, ha mások ua. a tanuló algoritmust használják, ha mások stacionáriusak, ha …)
- **adaptivitás** – változó másokhoz (hatékonynak maradni, ha mások megváltoznak)

Tanulás
- opponens-független
- opponens-függő (milyen mértékben „tud” róla)

Teljes együttműködés $\rho_1 = \cdots = \rho_n$

Optimális együttes Q értékek parallel tanulása (vektor az együttes cselekvés-halmazt jelenti):

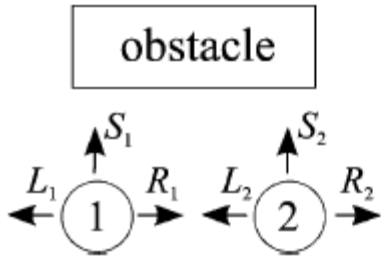
$$Q_{k+1}(x_k, \boldsymbol{u}_k) = Q_k(x_k, \boldsymbol{u}_k) + \alpha\left[r_{k+1} + \gamma \max_{\boldsymbol{u}'} Q_k(x_{k+1}, \boldsymbol{u}') - Q_k(x_k, \boldsymbol{u}_k)\right]$$

és belőle egyenkénti optimális eljárásmód származtatása

$$\bar{h}_i^*(x) = \arg\max_{u_i} \max_{u_1,\ldots,u_{i-1},u_{i+1},\ldots,u_n} Q^*(x, \boldsymbol{u})$$

A koordinálás szükségessége itt is általában megjelenik.

Példa: formáció-mozgás



Q	L_2	S_2	R_2
L_1	10	-5	0
S_1	-5	-10	-5
R_1	-10	-5	10

A két optimális helyzet ellenére, koordinálás hiányában ágensek $Q(L_1, R_2)$ szuboptimális helyzetben végezhetnek.
(ha a Q érték közös, mindkét optimális eset egy Nash egyensúly)

$$Q(L_1, L_2) = Q(R_1, R_2) = 10$$

Koordinálás kérdése

Koordinálás-mentes

pl. Team-Q: egyedi opt. együttes cselekvést tételez fel
Distributed-Q-Learning: lokális Q és h tanulás, de az egyedi Q frissítése csak akkor, ha növekszik (a közös opt.-t is el fogja kapni)

Koordinálás-alapú pl. együttes Q dekomponálása kisebb csoportullások szerint

$$Q(x, \boldsymbol{u}) = Q_1(x, u_1, u_2) + Q_2(x, u_1, u_3) + Q_3(x, u_3, u_4)$$

Indirekt koordinálás

pl. tanulva, hogy más ágensek bizonyos cselekedeteit milyen gyakorisággal használják:
JAL – Joint Action Learner: $C^i_j(u_j)$ – i-edik ágens hányszor tapasztalja, hogy j-edik ágens egy u_j cselekvéshez folyamodik (mások prob. modellje)

$$\hat{\sigma}^i_j(u_j) = \frac{C^i_j(u_j)}{\sum_{\tilde{u}_j \in U_j} C^i_j(\tilde{u}_j)}$$

Frequency Maximum Q-value heurisztika:
r_{max} – u_i mellett legjobb megerősítés,
C_{max} – ennek gyakorisága
C – a cselekvés gyakorisága

$$\tilde{Q}_i(u_i) = Q_i(u_i) + \nu \frac{C^i_{\max}(u_i)}{C^i(u_i)} r_{\max}(u_i)$$

a számított Q értéket a Boltzmann-felfedezés képletében használja

Koordinálás kérdése

Explicit koordinálás

pl. társadalmi szabályok
normatívák, törvények
kommunikáció
szerepek
...

pl. ágens 1 < ágens 2
L < R < S
döntés (L_1, L_2)

Teljes versengés

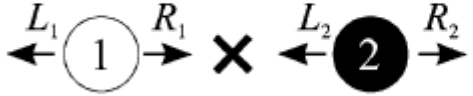
Minimax Q-tanulás
(1. ágens)

$$h_{1,k}(x_k, \cdot) = \arg \mathbf{m}_1(Q_k, x_k)$$

$$Q_{k+1}(x_k, u_{1,k}, u_{2,k}) = Q_k(x_k, u_{1,k}, u_{2,k}) + \alpha[r_{k+1} + \gamma\, \mathbf{m}_1(Q_k, x_{k+1}) - Q_k(x_k, u_{1,k}, u_{2,k})]$$

$$\mathbf{m}_1(Q, x) = \max_{h_1(x,\cdot)} \min_{u_2} \sum_{u_1} h_1(x, u_1) Q(x, u_1, u_2)$$

Teljes versengés



Q	L_2	R_2
L_1	0	1
R_1	-10	10

Minimax Q-tanulás, példa

1. ágens szeretne elfoglalni a keresztet és elmenekülni.
2. ágens szeretne elkapni az 1. ágenst.
A Q táblázat az 1. ágens perspektíváját mutatja, a 2. ágens Q függvénye ennek -1-szerese.

A minimax megoldás 1. ágensre:
Ha L1-et lép, akkor a 2. minimalizálva L2-et lép, eredményben 0.
Ha R1-et lép, akkor a 2. minimalizálva szintén L2-et lép, eredményben -10.
Az 1. ágensnek tehát L1-et kell lépnie, mert így legfeljebb 0-val megússza

Vegyes feladatok

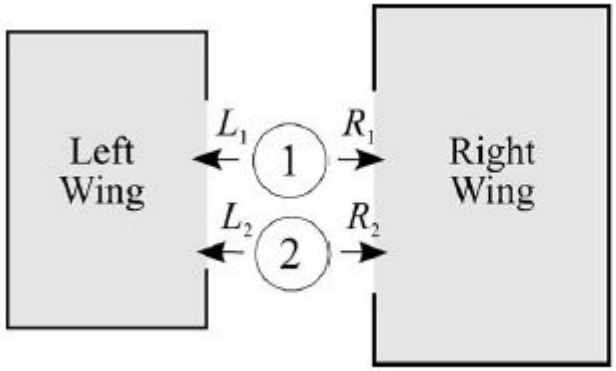
Nincsenek feltételek megerősítésekre. Valamilyen egyensúly felé kell húzni.
Lehet pl. a Nash-egyensúly, de mi van, ha több van?

Vegyes feladatok

- egyedi ágens Q-tanulás (a többi implicite a környezeti információban)
- ágens-független módszerek (egymástól független, de egy feltehetően közös egyensúly felé … (Nash-Q-learning, correlated equilibrium Q-learning, asymetric Q-learning, …)

Problémák egyensúlyi helyzetekkel, pl.:

Két porszívóágens feladata a két szobából álló lakás kitakarítása. Mindegyik jobban szeretne a bal szobát megkapni, mert ez kisebb.



Q_1	L_2	R_2
L_1	0	3
R_1	2	0

Q_2	L_2	R_2
L_1	0	2
R_1	3	0

Két Nash-egyensúly van: (L_1, R_2) (R_1, L_2)
de ha a két ágens között nincs koordináció, akkor mindketten ugyanabban a szobában végeznek, kisebb hasznossággal.

$$Q_1(L_1, L_2) = Q_1(R_1, R_2) = Q_2(L_1, L_2) = Q_2(R_1, R_2) = 0.$$

Vegyes feladatok

Ágens-követő, ágens-tudatos módszerek (más ágensek modellezése, a modell használata tanulásban: - érzékelés + stratégia-váltás)

AWESOME (Adapt When Everyone is Stationary, Otherwise Move to Equilibrium)

IGA (Infinitensimal Gradient Ascent) – ágensek cselekvéseinek valószínűsége az, amit az ágens tanul (2 ágens, 2 cselekvés,

»

α 1. ágens 1. cselekvése és
β 2. ágens 1. cselekvése:

$$\begin{cases} \alpha_{k+1} = \alpha_k + \delta_{1,k} \dfrac{\partial E\{r_1 \mid \alpha, \beta\}}{\partial \alpha} \\ \beta_{k+1} = \beta_k + \delta_{2,k} \dfrac{\partial E\{r_2 \mid \alpha, \beta\}}{\partial \beta} \end{cases}$$

$$\delta_{1,k} = \delta_{2,k} = \delta$$

WoLF-IGA (Win-or-Learn-Fast)

- győztes helyzetben ágens óvatos, kis δ-val lassan tanul, nehogy az előnyös pozícióját elveszítse
- vesztes esetben viszont nagyobb δ-val gyorsan kikerül a jelen helyzetből.

Lucian Busoniu, Robert Babuska, and Bart De Schutter, A Comprehensive Survey of Multiagent Reinforcement Learning, IEEE Trans. on Systems, Man, and Cybernetics—Part C: Applications and Reviews, Vol. 38, No. 2, March 2008