Visszacsatolásos tanulási módszerek a Mario játékra alkalmazva

Schnebli Zoltán

Babeș–Bolyai Tudományegyetem, Kolozsvár

2018 május 26

A gépi tanulás

- Felügyelt tanulás

A gépi tanulás

- Felügyelt tanulás
  - regresszió, osztályozás

A gépi tanulás

- Felügyelt tanulás
  - regresszió, osztályozás
- Felügyeletlen tanulás

A gépi tanulás

- Felügyelt tanulás
  - regresszió, osztályozás
- Felügyeletlen tanulás
  - klaszterezés

A gépi tanulás

- Felügyelt tanulás
  - regresszió, osztályozás
- Felügyeletlen tanulás
  - klaszterezés
- **Félig felügyelt tanulás**

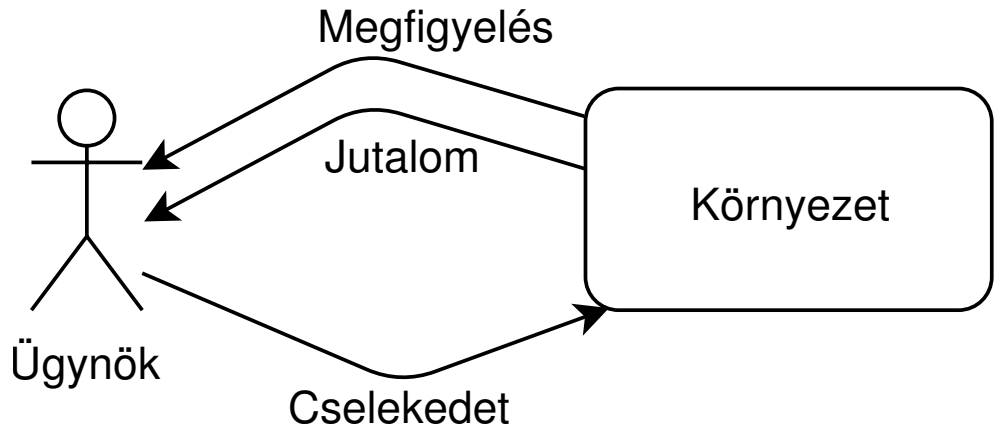
A gépi tanulás

- Felügyelt tanulás
  - regresszió, osztályozás
- Felügyeletlen tanulás
  - klaszterezés
- **Félig felügyelt tanulás**
  - robotika

A visszacsatolásos tanulás

Működési elv:

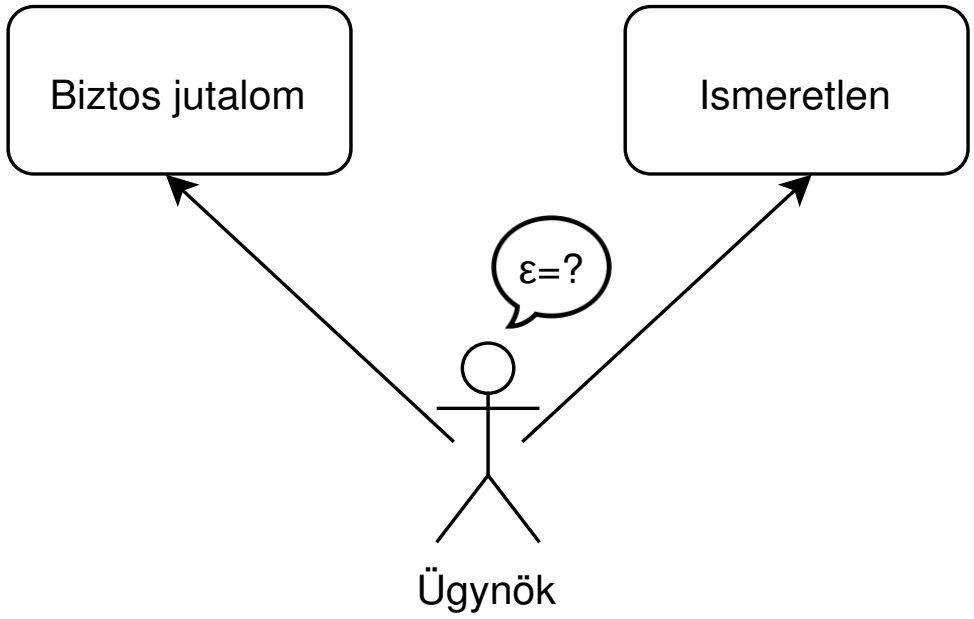
- Ügynök
- Környezet
- Állapot - Cselekedet
- Jutalom



A visszacsatolásos tanulás

Felfedezés és kizsákmányolás

- ϵ - mohó stratégia
- csökkenő ϵ - paraméter



A visszacsatolásos tanulás

Elemei:

- Irányelv
- Jutalomfüggvény
- Értékfüggvény
- Környezet modellje

A visszacsatolásos tanulás

Markov döntési folyamat

- $(S, A, \mathcal{P}_{\cdot}(\cdot,\cdot), \mathcal{R}_{\cdot}(\cdot,\cdot), \gamma)$

A visszacsatolásos tanulás

Markov döntési folyamat

- $(S, A, \mathcal{P}_{\cdot}(\cdot,\cdot), \mathcal{R}_{\cdot}(\cdot,\cdot), \gamma)$
  - S - állapottér

A visszacsatolásos tanulás

Markov döntési folyamat

- $(S, A, \mathcal{P}_{\cdot}(\cdot,\cdot), \mathcal{R}_{\cdot}(\cdot,\cdot), \gamma)$
  - S - állapottér
  - A - cselekvéstér

A visszacsatolásos tanulás

Markov döntési folyamat

- $(S, A, \mathcal{P}_\cdot(\cdot,\cdot), \mathcal{R}_\cdot(\cdot,\cdot), \gamma)$
  - S - állapottér
  - A - cselekvéstér
  - $\mathcal{P}_a(s, s')$ - s' valószínűségét határozza meg

A visszacsatolásos tanulás

Markov döntési folyamat

- $(S, A, \mathcal{P}_\cdot(\cdot,\cdot), \mathcal{R}_\cdot(\cdot,\cdot), \gamma)$
  - S - állapottér
  - A - cselekvéstér
  - $\mathcal{P}_a(s, s')$ - s' valószínűségét határozza meg
  - $\mathcal{R}_a(s, s')$ - s'-be vezető cselekedet jutalmát hazározza meg

A visszacsatolásos tanulás

Markov döntési folyamat

- $(S, A, \mathcal{P}_\cdot(\cdot,\cdot), \mathcal{R}_\cdot(\cdot,\cdot), \gamma)$
  - S - állapottér
  - A - cselekvéstér
  - $\mathcal{P}_a(s, s')$ - s' valószínűségét határozza meg
  - $\mathcal{R}_a(s, s')$ - s'-be vezető cselekedet jutalmát hazározza meg
  - γ - engedményfaktor

A visszacsatolásos tanulás

Megoldási módszerek:

- Dinamikus programozás
  - kipróbál minden lehetőséget mielőtt lép
  - sok memória
- Monte Carlo módszerek
  - tapasztalat
  - pontos környezeti modell
- **Időbeli-differencia tanulás** (ID tanulás)
  - előző két elv egyesítése

Időbeli-differencia tanulás

A legegyszerűbb ID tanulás

- TD(0)
- $V(s_t) \leftarrow V(s_t) + \alpha[r_t + \gamma V(s_{t+1}) - V(s_t)]$
- α - tanulási ráta

Időbeli-differencia tanulás

Q - tanulás

- $Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \left[r_{t+1} + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t) \right]$
- nem ígényel környezeti modellt
- képes optimális irányelvet találni

Időbeli-differencia tanulás

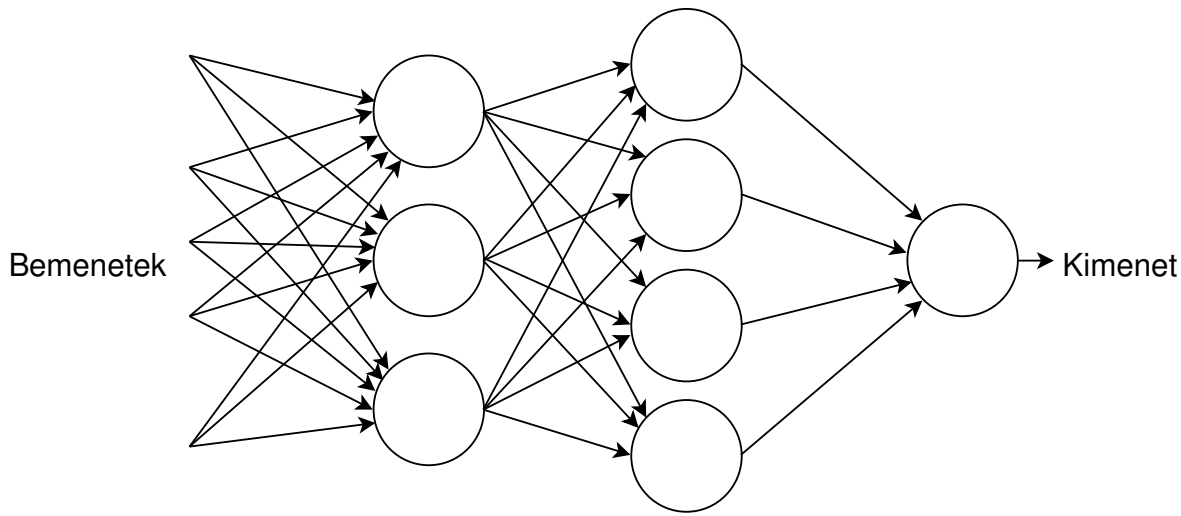
SARSA - tanulás

- Állapot-Cselekvés-Jutalom-Állapot-Cselekvés
- Q - tanulásból származik
- $Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \ [\ r_{t+1} + \gamma \ Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)\ ]$

Időbeli-differencia tanulás

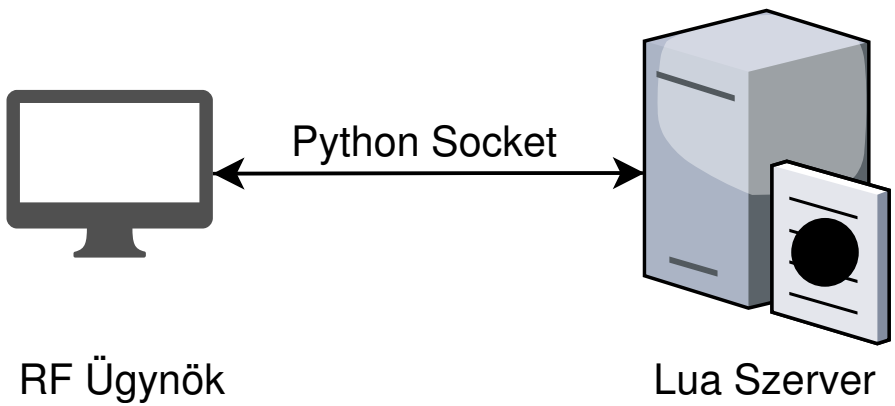
Deep Q - tanulás

- Neurális háló
- Tapasztalat
- 3 réteg



Alkalmazás

- Visszacsatolásos ügynök
- Lua szerver
- Környezeti modell



Alkalmazás

- JSON üzenetek
- a játék reprezentációja

$$
\begin{matrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 3 & 0 & 0 & 2 & 2 & 0 & 0 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{matrix}
$$

Eredmények

- 55+ százalékos sikeresség
- maximálisan elért távolság