Deterministički procesi odlučivanja

Teorijske osnove

Adaptivni sistemi – L1a

Milan R. Rapaić

Katedra za automatsko upravljanje
Departman za računarstvo i automatiku
Fakultet tehničkih nauka
Univerzitet u Novom Sadu

10. januar 2020

The European Commission's support for the production of this publication does not constitute an endorsement of the contents, which reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Co-funded by the
Erasmus+ Programme
of the European Union

Sadržaj

Opšti pojmovi

Formalna postavka problema

- Procesi odlučivanja sa potpunom opservacijom stanja
- Procesi odlučivanja sa nepotpunom opservacijom stanja

Princip optimalnosti

Zaključna razmatranja

Process odlučivanja: Osnovni pojmovi 1/5

Agent i njegovo okruženje

- AGENT interaguje sa OKRUŽENJEM sekvencijalno, u DISKRETNIM VREMENSKIM TRENUCIMA.
- U trenutku t, okruženje se nalazi u nekom STANJU $s_t \in \mathcal{S}$.
  - Stanje okruženja je rezultat svih prethodnih akcija agenta i drugih činilaca (koji nam mogu biti nepoznati).
  - Tekuće stanje okruženja predstavlja svu informaciju o prethodnim događajima (istoriji okruženja) koju je neophodno poznavati da bi se poznavalo ponašanje okruženja u budućnosti.
- U svakom trenutku, agent "prima" informaciju (OPSERVACIJU) o stanju okruženja $o_t \in \mathcal{O}$, i na osnovu te informacije bira način delovanja, odnosno AKCIJU $a_t \in \mathcal{A}$.
- U *sledećem* vremenskom trenutku, okruženje reaguje na tekuću akciju agenta, MENJA SE (prelazeći u naredno stanje) i pri tome "nagrađuje" agenta NAGRADOM $r_{t+1} \in \mathcal{R} \subseteq \mathbb{R}$.

Process odlučivanja: Osnovni pojmovi 2/5

Stanje, opservacija, akcija

- STANJE okruženja možemo razumeti kao *minimalni skup informacija o tekućoj situaciji dovoljan za donošenje (ispravne, najbolje, optimalne, ...) odluke*.
- OPSERVACIJA najčešće nosi samo delimičnu informaciju o tekućem stanju. Uopšte, razlikujemo dve vrste procesa odlučivanja:
  - sa potpunom opservacijom stanja (kada je u svakom trenutku $o_t = s_t$),
  - sa nepotpunom opservacijom stanja.
- Ponekad su agentu različite akcije na raspolaganju u zavisnosti od stanja okruženja.
  - U slučaju potpune opservacije to znači da za svako stanje s imamo skup akcija dopustivih u tom stanju $\mathcal{A}(s)$, odnosno da je u svakom trenutku vremena $a_t \in \mathcal{A}(s_t)$.

Process odlučivanja: Osnovni pojmovi 3/5

Cilj i strategija agenta

- Pravilo (algoritam, logika) na osnovu koga agent bira narednu akciju poznajući stanje okruženja nazivamo STRATEGIJOM.
- Cilj agenta je da maksimizuje DOBIT, tj. određenu kumulativnu meru nagrada koje prima tokom interakcije sa okruženjem.
- UČENJE je (u ovom kontekstu) postupak promene strategije na osnovu iskustva.

Process odlučivanja: Osnovni pojmovi 4/5

Vrednost stanja i akcije

- VREDNOST STANJA s PRI STRATEGIJI u je očekivana dobit ukoliko, polazeći iz stanja s, primenjujemo strategiju u.
- VREDNOST AKCIJE a U STANJU s PRI STRATEGIJI u je očekivana dobit ukoliko, polazeći iz stanja s, u prvim koraku delujemo akcijom a, a nakon toga primenjujemo strategiju u.
- VREDNOST STANJA s jeste maksimalna dobit koju agent može ostvariti kada je okruženje u stanju s. Drugim rečima, to je dobit koju agent ostvaruje ukoliko bira najisplativiju strategiju.
- VREDNOST AKCIJE a U STANJU s jeste maksimalna dobit koju agent može ostvariti kada deluje akcijom a ukoliko je okruženje u stanju s.

Process odlučivanja: Osnovni pojmovi 5/5

Problem optimalnog odlučivanja

- OPTIMALNA STRATEGIJA je ona koja u svakom stanju deluje tako da se maksimizira dobit.
- ... drugim rečima, u^* je optimalna strategija ako (i samo ako) je vrednost proizvoljnog stanja pri toj strategiji veća ili jednaka od vrednosti istog stanja pri proizvoljnoj drugoj strategiji!

Okruženje

- Za svako tekuće stanje okruženja s i svaku tekuću akciju agenta a jednoznačno je određeno naredno stanje s^+.

$$s^+ = f(s, a)$$

- Za svako tekuće stanje okruženja s i svaku tekuću akciju agenta a jednoznačno je određena nagrada r koju agent dobija.

$$r = h(s, a)$$

Dobit

- Dobit agenta u nekom trenutku definišemo kao (otežanu) sumu svih budućih nagrada.

$$g_t := \sum_{\tau=0}^{\infty} \gamma^\tau r_{t+1+\tau}$$

  $\gamma > 0$ nazivamo *faktorom obezvređivanja*.
- Dobit se može računati rekurzivno

$$g_t = r_{t+1} + \gamma g_{t+1}$$

Strategija

- Strategija agenta jeste preslikavanje $u \in \mathcal{U}$ koje svakom stanju okruženja dodeljuje akciju kojom će agent delovati u tom stanju

$$a = u(s)$$

- Vrednost stanja s pri strategiji u

$$v_u(s) = g_t \text{ kada je } s_t = s \text{ i } (\forall k \geq t)\ a_k = u(s_k)$$

- Vrednost akcije a u stanju s pri strategiji u

$$q_u(s, a) = g_t \text{ kada je } s_t = s\ , a_t = a \text{ i } (\forall k \geq t + 1)\ a_k = u(s_k)$$

Optimalno odlučivanje

- u^* je optimalna strategija ako i samo ako je

$$v_{u^*}(s) \geq v_u(s) \quad (\forall s \in \mathcal{S})(\forall u \in \mathcal{U})$$

- Optimalna akcija u stanju s jeste ona koju propisuje optimalna strategija $a^* = u^*(s)$
- Vrednost stanja s jeste vrednost tog stanja pri optimalnoj strategiji

$$v(s) := v_{u^*}(s)$$

Princip optimalnosti

An optimal policy has the property that whatever the initial state and initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision.

Bellman, 1957, Chap. III.3.

Isaacs vs Bellman

Dinamičko programiranje

Dynamic Programming (DP) is used heavily in optimization problems (finding the maximum and the minimum of something). Applications range from financial models and operation research to biology and basic algorithm research. So the good news is that understanding DP is profitable. However, the bad news is that DP is not an algorithm or a data structure that you can memorize. It is a powerful algorithmic design technique

http://courses.csail.mit.edu/6.006/fall11/rec/rec19.pdf

Belmanova jednačina 1/2

Vrednost proizvoljnog stanja u ma kojoj datoj strategiji se može računati rekurzivno!

$$
\begin{aligned}
v_u(s) &= g_t \text{ kada je } s_t = s \text{ i } (\forall k \geq t)\ a_k = u(s_k) \\
&= r_{t+1} + \gamma g_{t+1} \text{ ako } s_t = s \text{ i } (\forall k \geq t)\ a_k = u(s_k) \\
&= h(s, u(s)) + \gamma\,(g_{t+1} \text{ ako } s_t = s \text{ i } (\forall k \geq t)\ a_k = u(s_k)) \\
&= h(s, u(s)) + \gamma\,(g_{t+1} \text{ ako } s_{t+1} = f(s, u(s)) \text{ i } (\forall k \geq t+1)\ a_k = u(s_k)) \\
&= h(s, u(s)) + \gamma v_u(f(s, u(s)))
\end{aligned}
$$

$$v_u(s) = h(s, u(s)) + \gamma v_u(f(s, u(s)))$$

Belmanova jednačina 2/2

$$v(s) = \max_{a \in \mathcal{A}} \{h(s,a) + \gamma v(f(s,a))\}$$

Optimalno dejstvo (akcija) a^* jeste ona za koje se postiže maksimum u Belmanovoj jednačini.

Pojmovnik

srpski	oznaka	engleski
stanje	x	state
akcija / dejstvo	a	action
nagrada	r	reward
dobit	g	gain
vrednost		value
faktor obezvređivanja	γ	discount factor