On-line learning in real-time environments

http://op.matfyz.cz/sarn
Ondřej Pacovský
op@matfyz.cz

Cíle

- učení přímo v průběhu života agenta
- minimalizace zásahů člověka
- dynamické prostředí
- prototyp ve virtuálním 3D světě

Možné využití

- Samotný SARN algoritmus
- Techniky v něm použité (ESN, hebb, reinforcement)
- Interface k UT2004

Zpětnovazebné učení (Reinforcement learning)

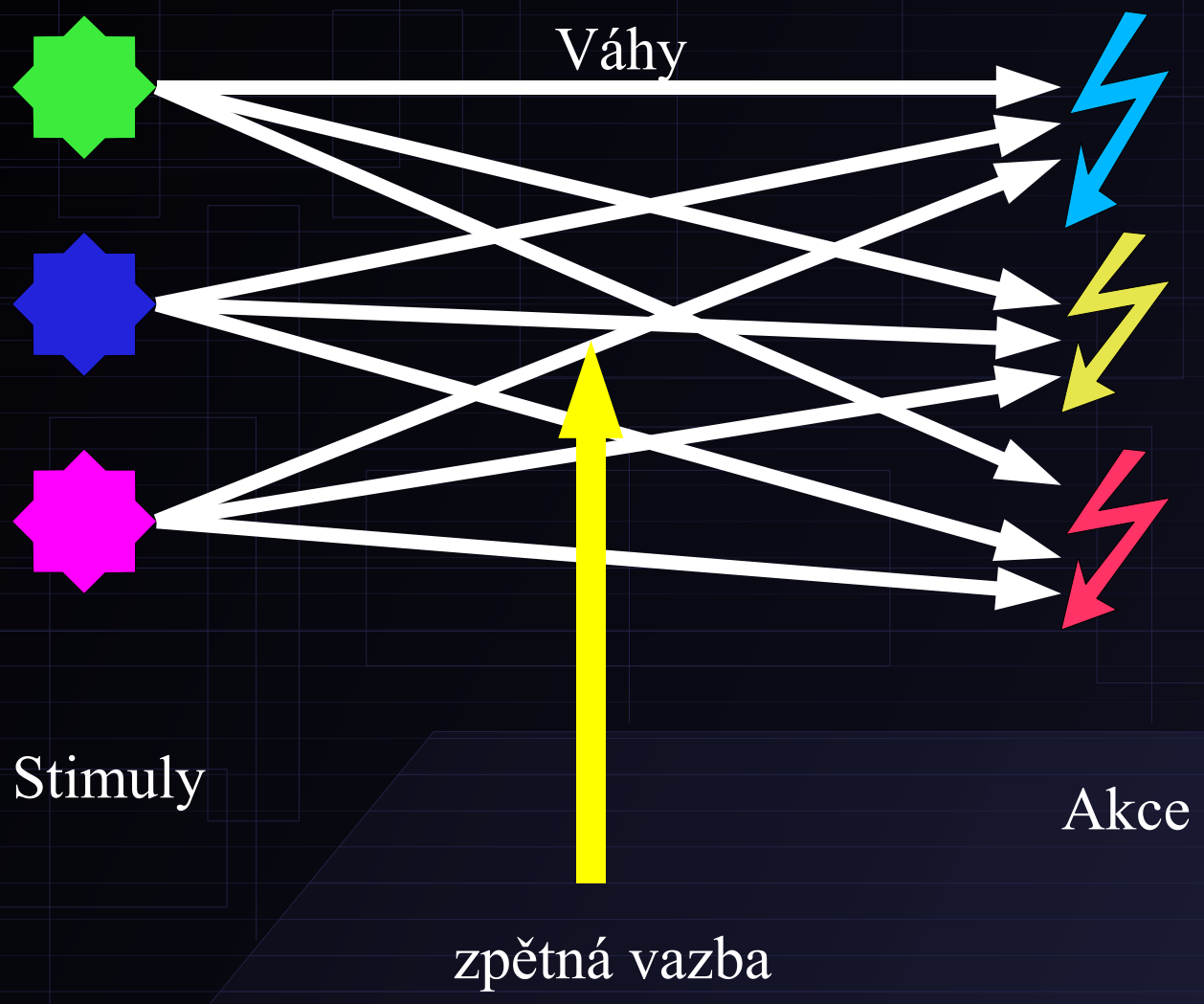
- učitel neposkytuje vzorová řešení
- systém se učí jen pomocí “cukru a biče”
- agent v každém kroku vybírá akci a dostává odezvu

- Standardní řešení: algoritmus Q-learning
  - diskrétní stavy
  - postupná aproximace užitkové funkce

Stimulus Action Reward Network (1)

- konekcionistický přístup - prvky hebbovského učení
- diskrétní stavový prostor nahrazuje množina stimulů
- pomocí zpětné vazby posiluje vazby stimulů relevantních k dané akci a naopak

Stimulus Action Reward Network (2)



Krok adaptace - detail

a index vzbrané akce

f zpětná vazba

S vektor stimulů $S \in \{-1;1\}^M$

$\mathbf{w}_{ab}$ váha vazby stimulu **a** na akci **b**

pro i = {1,..,M}

$$\Delta w_{ia^t} = S_i^t f^t$$

Přiřazování zásluh (credit assigment)

- Akce v čase **t** ovlivňuje odměny v dalších časových okamžicích, nejen **f**^t+1

proto SARN používá metodu sledování historie - eligibility traces

ukládáme historii **h** stavů **S** a vybraných akcí

- v každém kroku přiřadíme odměnu i akcím v historii s klesající silou

$$\Delta w_{ia^t} = \gamma^h S_i^{t-h} f^{t-h}$$

Kdo určí stimuly?

- manuálně
  - pro každý problém zvlášť (podobné problémy jako u diskretizace)
  - předem omezuje možná řešení
  - chování systému je lépe analyzovatelné
- automaticky – Echo State Network
  - náhodná, předem zkonstruovaná rekurentní neuronová síť
  - není závislá na daném problému
  - zajímavé dynamické vlastnosti

Prototyp

- virtuální 3D prostředí Unreal 2004
- agent (“bot”) má za úkol uhybat se projektilům
- akce: vlevo, vpravo, žádná
- zpětná vazba: penalizace za zásah a také za jakýkoliv pohyb agenta

- ukázka

Výsledky

+

- v dané časové škále efektivnější než Q-learning
- malá účast člověka
- prototyp slibný pro reálné použití

-

- adaptačního pravidlo trpí nedostatky při změně podmínek prostředí
- je třeba otestovat na širší škále problémů

A .. dál ?

- cart pole bench
- GA
- roboti
- integrace do reálně použitelného bota