Q learning - aproximácia funkcie ohodnotení neurónovou sieťou

Michal CHOVANEC
Fakulta riadenia a informatiky

Jún 2015

Úvod

Cieľom je nájsť optimálnu stratégiu - maximalizácia odmeny (účelovej funckie)

- Vopred nie je známa hodnota odmeny vykonanej akcie
- Vopred nie je známi ani stav do ktorého sa systém dostane
- Je možné určiť v akom stave sa systém nachádza
- Je presne daná množina akcií v každom stave
- Aspoň pre cieľový stav je daná výška odmeny

Úvod

Aplikácie zo sveta robotických sytémov

- Učenie sa pohybu, s ohľadom na techické prostriedky a terén
- Multirobotické plánovanie - hľadanie optimálneho rozhodnutia pre celú skupinu
  - Mapovanie
  - Hľadanie cieľa
  - Robotický futbal
  - Capture the flag
- Optimalizácia v automatických dopravných systémov
  - Dať prednosť, alebo predbehnúť
  - Kedy ísť zobrať náklad, predikcia
- Všetky problémy kde : ako niečo urobiť je zložité popísať
  - Systém si učením sám nájde postup ako niečo robiť
  - Vyžaduje sa adativita a samostatnosť

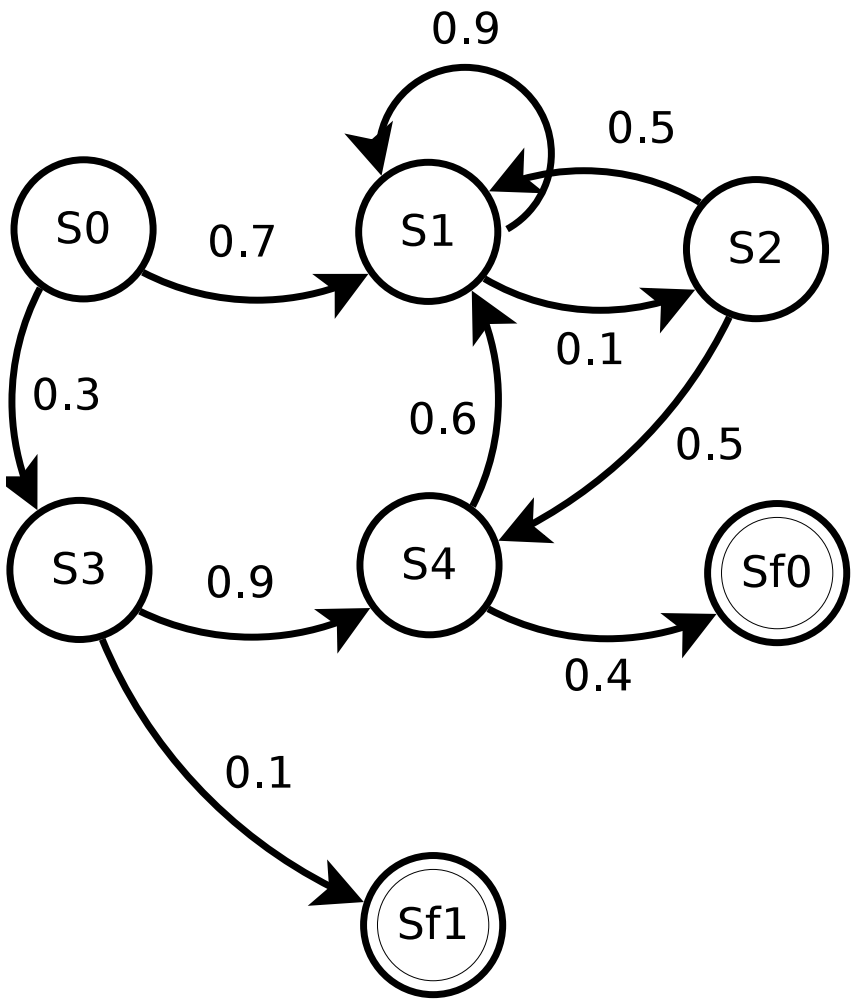
Q learning

- $Q(s, a)$ - funkcia ohodnotení
- s - stav
- a - akcia v stave s
- $R(s, a)$ - funkcia okamžitých odmien, za vykonanie a v stave s

$$s \in \mathbb{S}$$
$$a_s \in \mathbb{A}_s$$

Q learning - prechod stavovým priestorom

Markovov rozhodovací proces



Q learning - ohodnotenie Q(s, a)

Z Bellmanového princípu optimality

$$Q_{n+1}(s,a) = R_{n+1}(s,a) + \gamma \max_{a'} Q_n(s'_{n+1}, a') \qquad (1)$$

Kde
$R_{n+1}(s,a)$ je získaná odmena (reward) za vykonanie akcie a v stave s v čase $n+1$

$\max_{a'} Q_n(s'_{n+1}, a')$ je výber akcie v stave s_{n+1} ktorá má najväčšia odmenu

γ je podiel z maximálnej odmeny v stave s'_{n+1} pri vykonaní najlepšej možnej akcie v tomto stave

Q learning - ohodnotenie Q(s, a)

Varianty algoritmu
Filtrovanie v stochastickom prostredí

$$Q_{n+1}(s,a) = \alpha Q_n(s,a) + (1-\alpha)(R_{n+1}(s,a) + \gamma \max_{a'} Q_n(s'_{n+1}, a'))$$

SARSA algoritmus

$$Q_{n+1}(s,a) = \alpha Q_n(s,a) + (1-\alpha)(R_{n+1}(s,a) + \gamma Q_n(s'_{n+1}, a'))$$

kde $\alpha \in \langle 0, 1)$

Q learning - výber akcie

Boltzmanové rozdelanie

$$P(s|a_i) = \frac{k^{Q(s,a_i)}}{\sum_{j \in \mathbb{A}} k^{Q(s,a_j)}}$$

Kde $k \in \langle 0, \infty)$ a určuje správanie sa agenta, pre $Q(s,a) \in \langle -1, 1 \rangle$ možno pozorovať tieto druhy správania

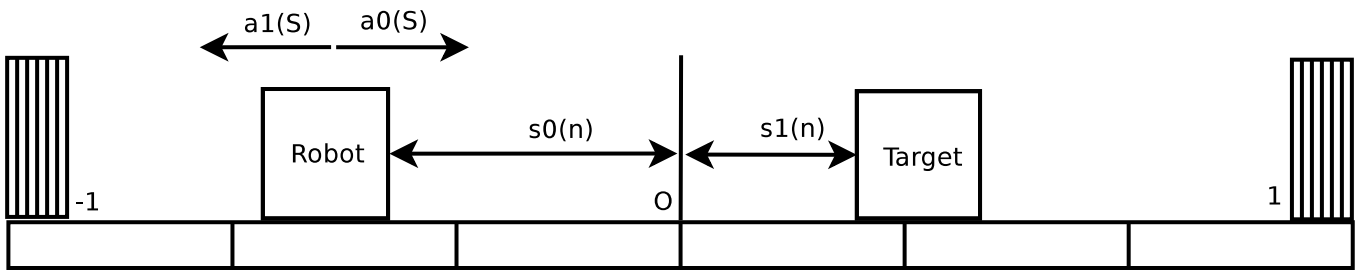
- $k = 1$ prieskumník
- $k = \langle 2, 10 \rangle$ zlatý stred
- $k >> 10$ pažravý (greedy) agent

Q learning - problémy

- Rosiahly stavový priestor, $Q(s, a)$ je možné nájsť len približne
  - Interpolácia
  - Transformácia pomocou features a ich lineárna kombinácia
  - Aproximácia neurónovou sieťou
- Výber akcie
  - Tak aby neboli prehľadávané akcie ktoré nemajú cenu
- Zdieľanie a syntéza $Q(s, a)$ medzi viacerými agentami

Experiment

Cieľom je overiť aproximáciu $Q(s, a)$ dvoma rôznymi neurónovými sieťami pri rôznych veľkostiach k v malom stavovom priestore - Q(s, a) vieme spočítať presne.



Experiment

Boli porovnávané dva modely neurónov s optimálnym riešením

$$y(n) = \varphi(\sum_{i=0}^{N-1} x_i(n) w_i(n)) \tag{2}$$

$$y(n) = \sum_{j=1}^{\infty} \sum_{i=0}^{N-1} x_i^j(n) w_{ji}(n) + \sum_{j=0}^{N-1} \sum_{i=j+1}^{N-1} x_i(n) x_j(n) v_{ji}(n) \tag{3}$$

Prečo by 3 mal byt lepší? Kolmogorov teorém 3 skryté vrstvy !

Experiment

Hypotéza :

Problém aproximácie $q = Q(s, a)$ je pridelenie každému s, a práve jedno q. To vedie na formuláciu :
AK je systém v stave s A bola vykonaná akcia a, POTOM výstup je q.

Pre jeden stav a dve akcie (ohodnotené ako a_0, a_1) možno napísať

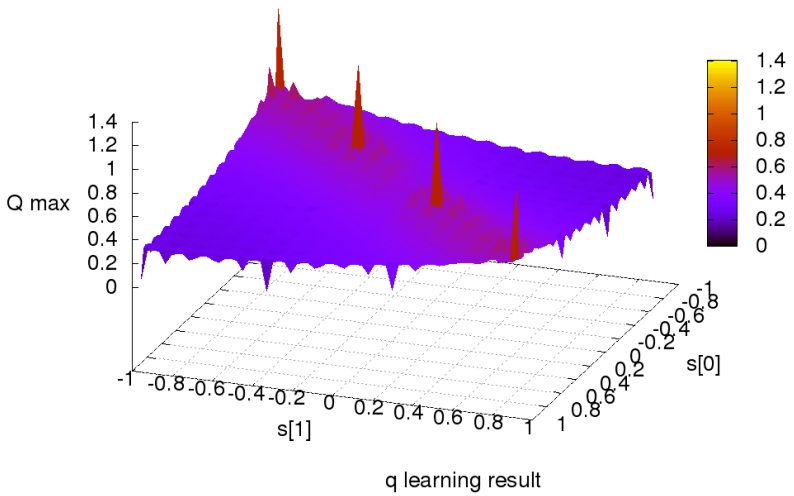
$$q = va_0 + (1 - v)a_1 \tag{4}$$

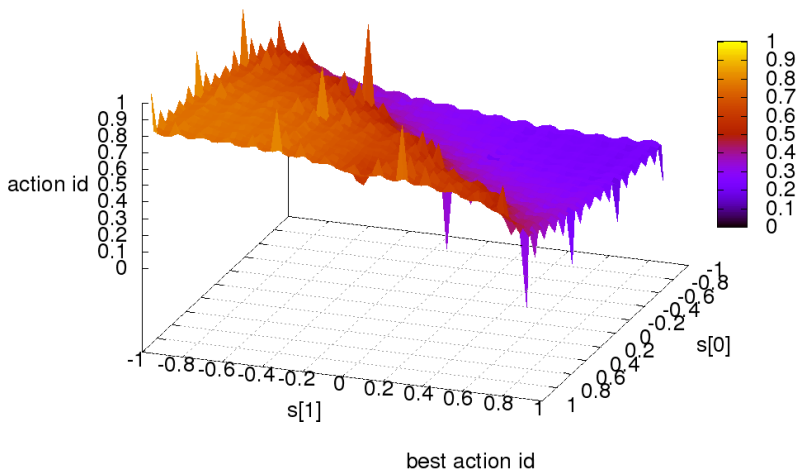
Kde v je výber akcie a platí $v \in \langle 0, 1 \rangle$. Pozn. všetky veličiny sú premenné. To je ale potom v súlade s 3.

Experiment - parametre

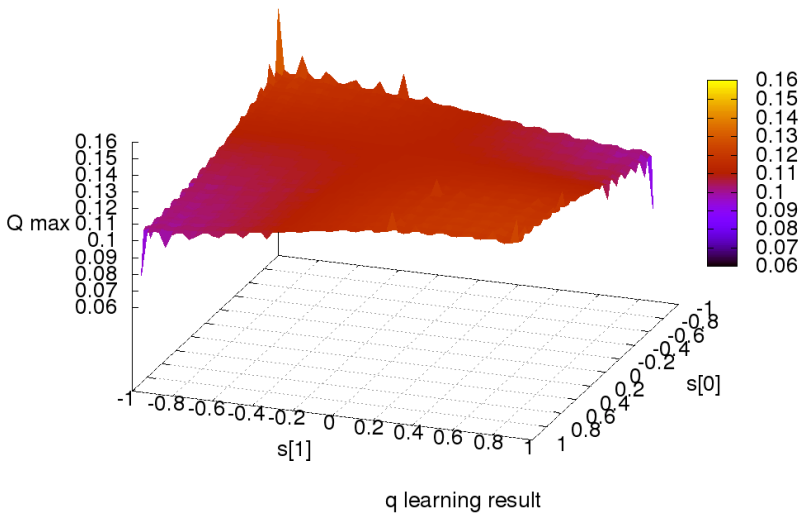
```
iterations = 10000000

agent :

state_density = 1/8.0
alpha = 0.98
gamma = 0.7

neural network :

hidden layers = 2
neurons in hidden layers = 10
weight range = 4.0
neuron order = 7
init weight range = 0.1
eta = 0.001
```

Experiment, k = 1.1, optimálne riešenie

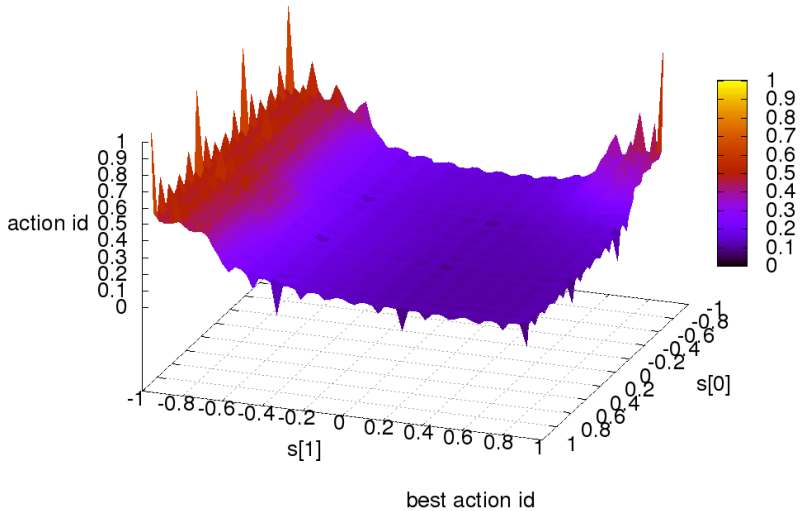




Experiment, k = 1.1, mcculloch pitts neurón

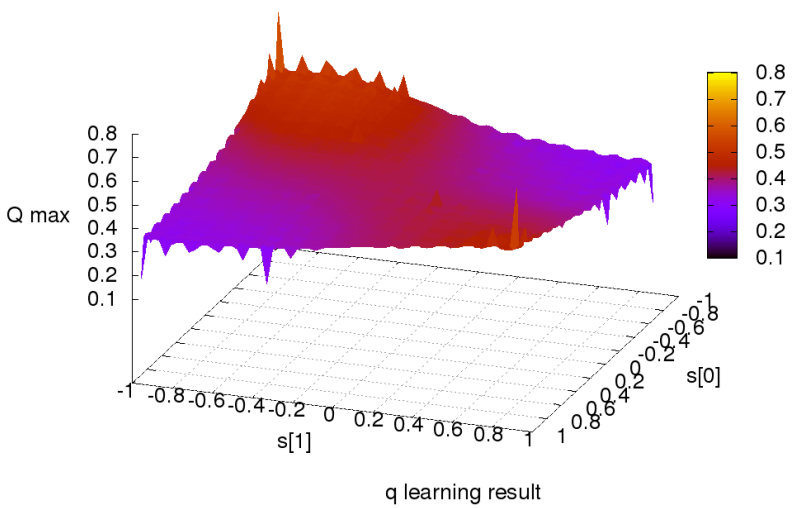


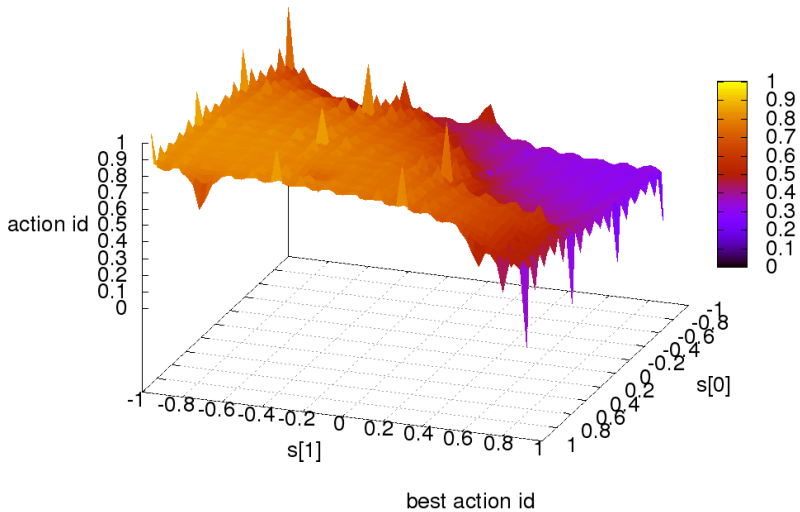
q learning result



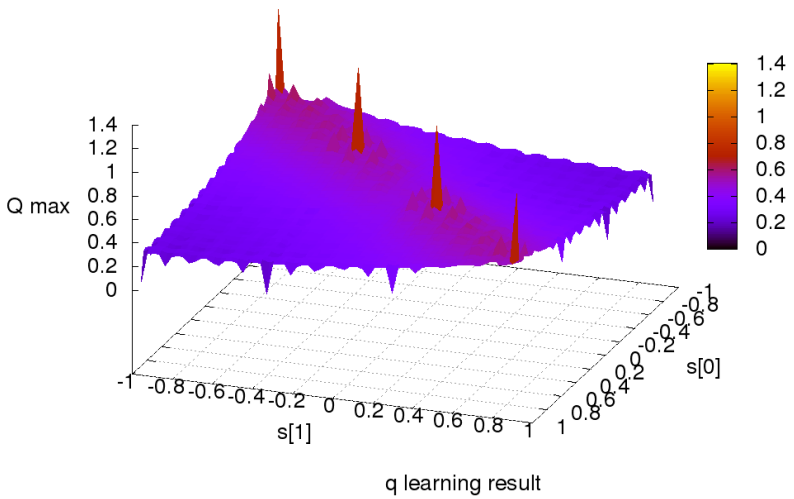
best action id

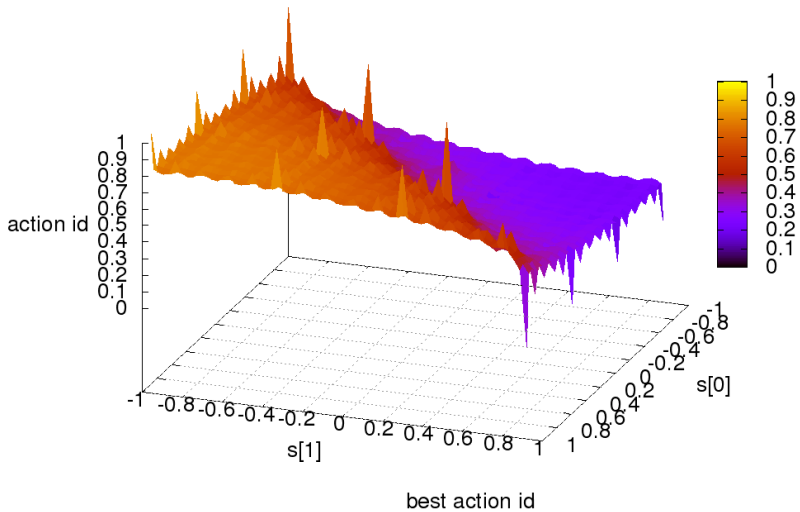
Experiment, k = 1.1, testovaný neurón



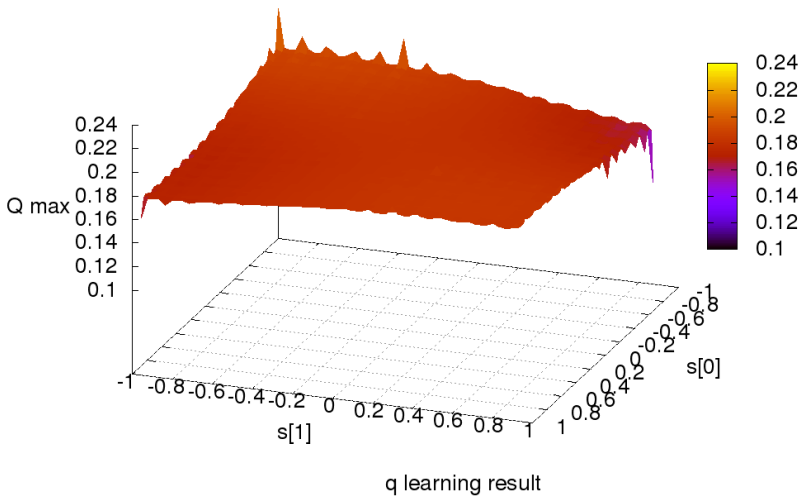


Experiment, k = 2.0, optimálne riešenie

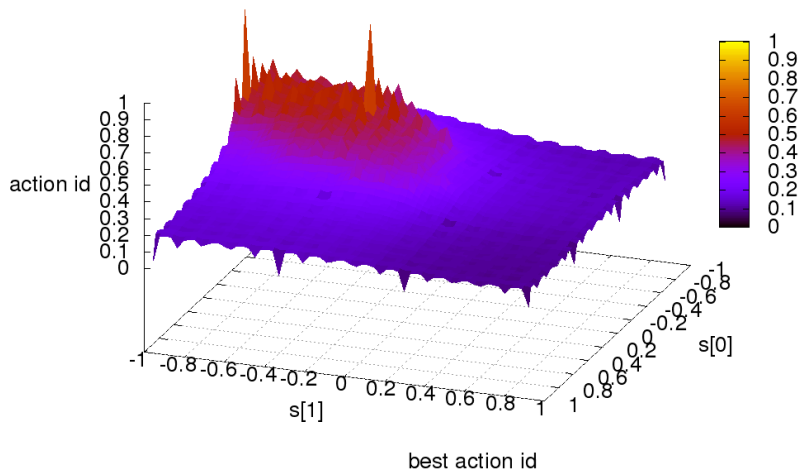




Experiment, k = 2.0, mcculloch pitts neurón

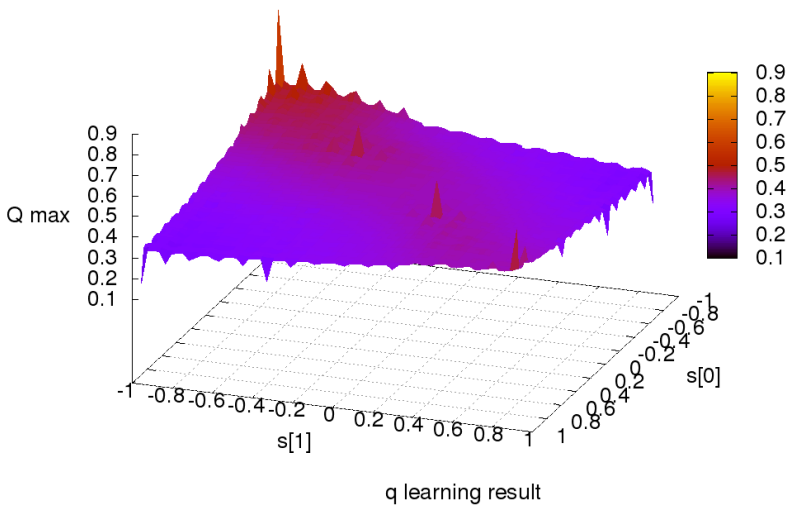


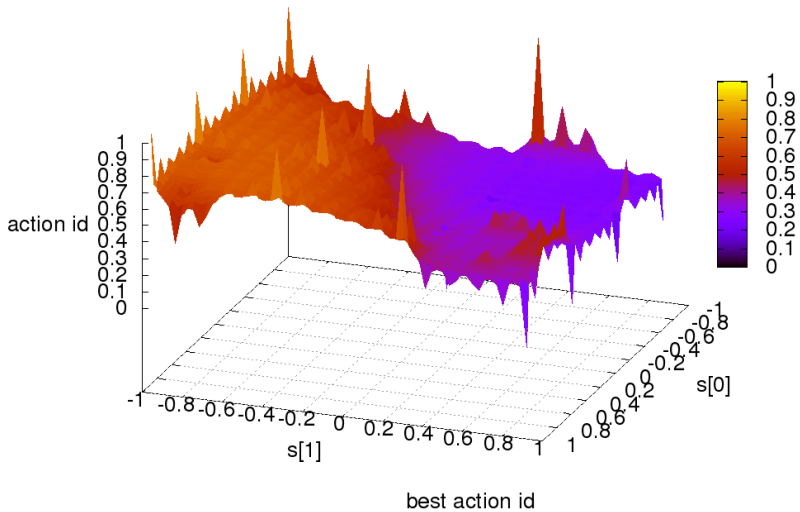
q learning result



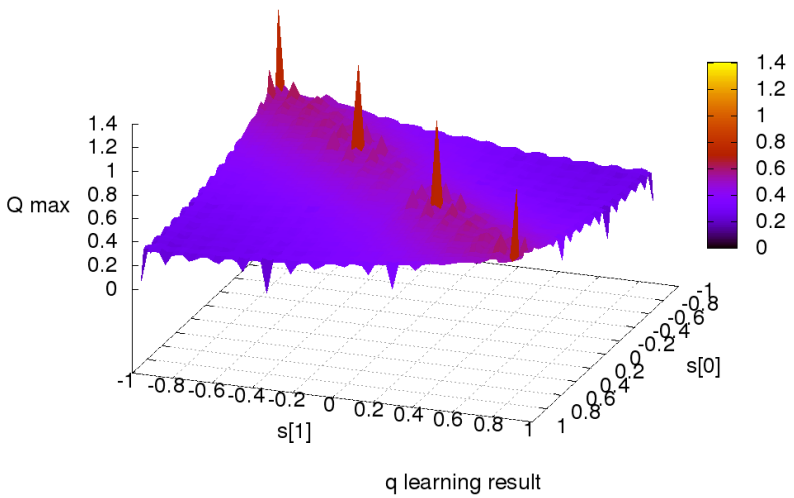
best action id

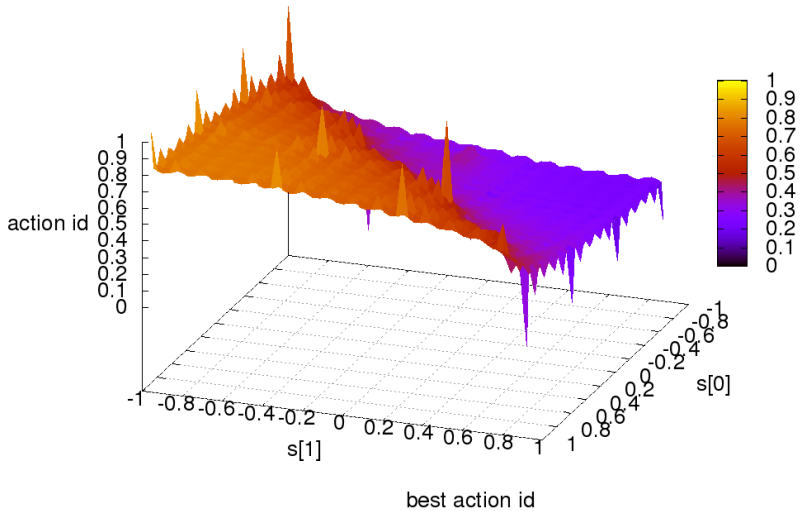
Experiment, k = 2.0, testovaný neurón



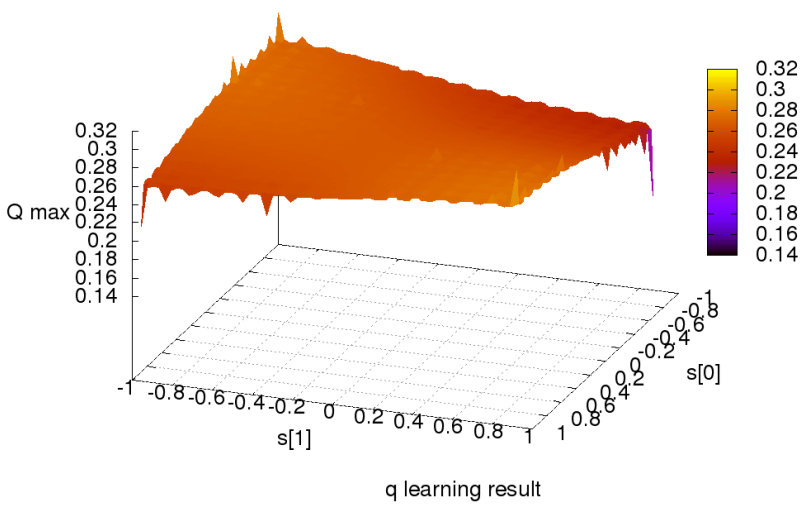


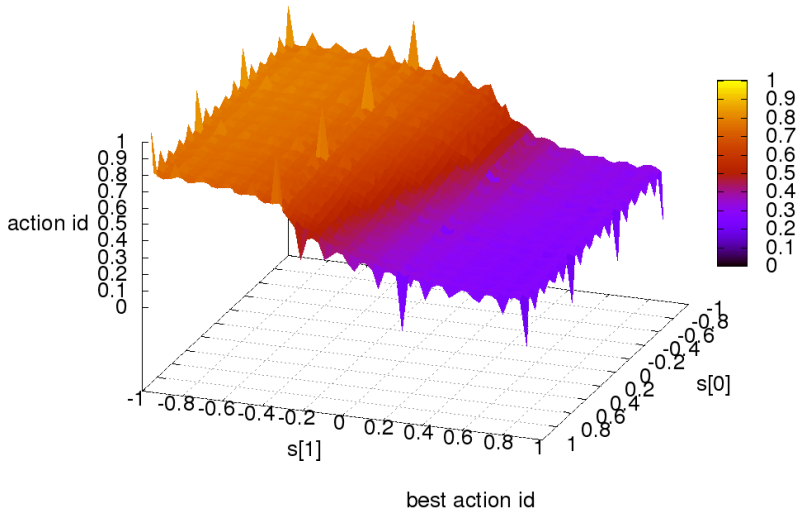
Experiment, k = 10.0, optimálne riešenie



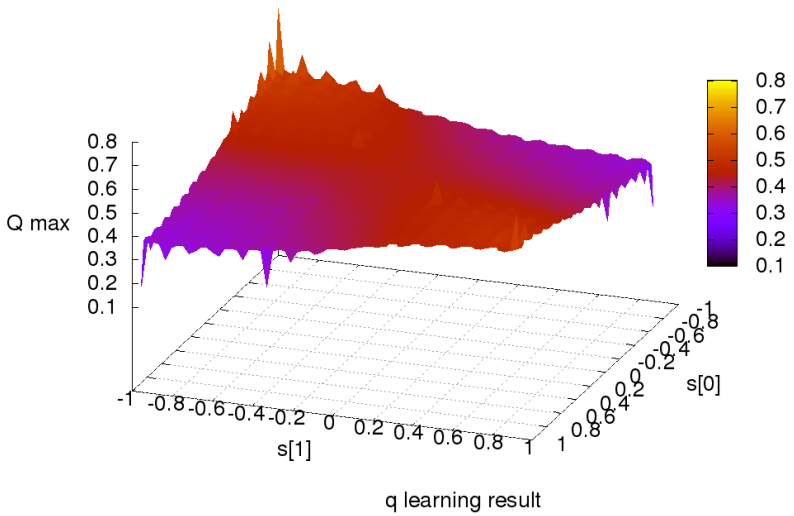


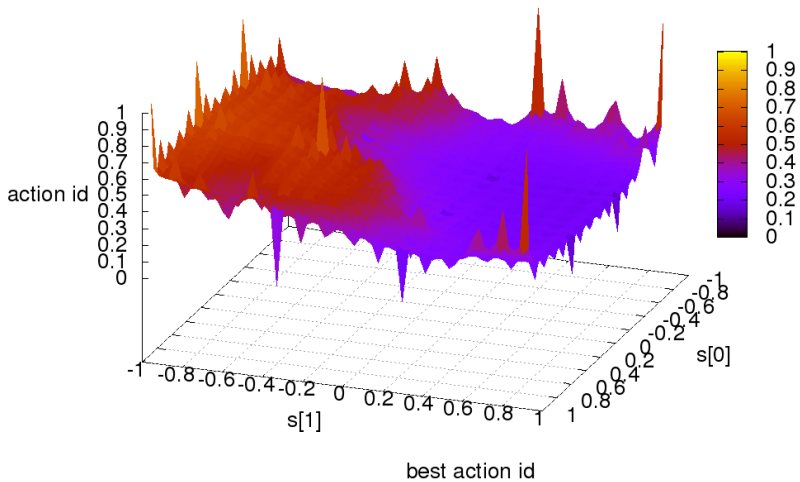
Experiment, k = 10.0, mcculloch pitts neurón



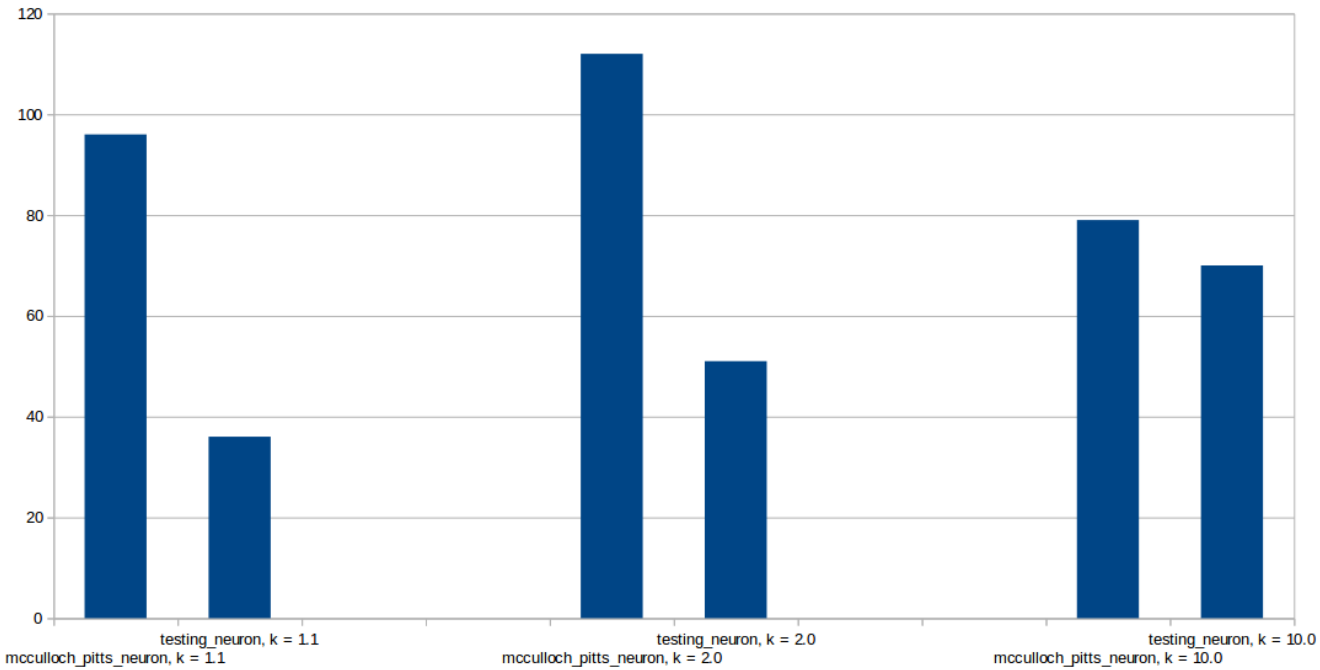


Experiment, k = 10.0, testovaný neurón





Experiment - zhrnutie



Nevyriešené otázky

... ktorými sa momentálne zaoberám

1. Zrýchliť učenie neurónovej siete
2. Urobiť experiment vo veľkom stavovom priestore
3. Implementácia do vyvíjaného multirobotického frameworku

Ďakujem za pozornosť

michal.chovanec@yandex.com