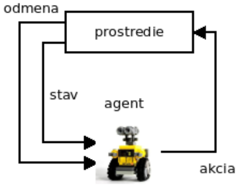
Aproximácia funkcie ohodnotení v algoritmoch Q-learning

Ing. Michal CHOVANEC
Fakulta riadenia a informatiky

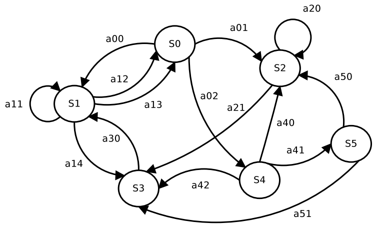
August 2016

Reinforcement learing

"A way of programming agents by reward and punishment without needing to specify how the task is to be achieved."
[Kaelbling, Littman, Moore, 96]



1. Zistenie stavu
2. Výber akcie
3. Vykonanie akcie
4. Prechod do ďalšieho stavu
5. Získanie odmeny alebo trestu
6. Učenie sa zo získanej skúsenosti



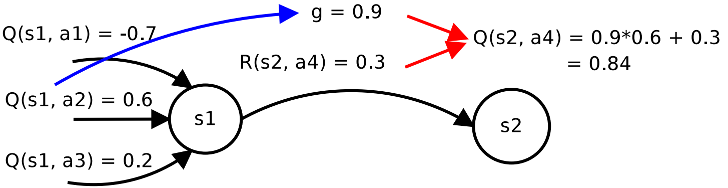
Funkcia ohodnotení

Daná je funkcia ohodnotení (Watkins, 1989)

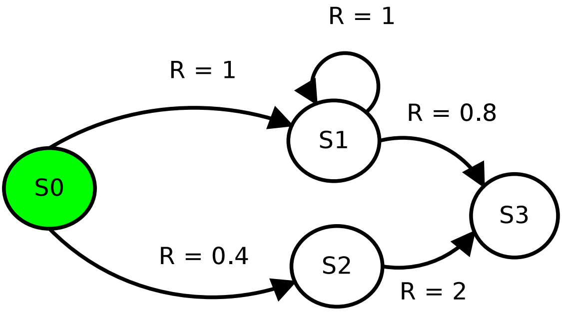
$$Q(s(n), a(n)) = R(s(n), a(n)) + \gamma \max_{a(n-1) \in \mathbb{A}} Q(s(n-1), a(n-1))$$

kde

- $R(s(n), a(n))$ je odmeňovacia funkcia
- $Q(s(n-1), a(n-1))$ je funkcia ohodnotení v stave $s(n-1)$ pre akciu $a(n-1)$,
- γ je konštanta zabúdania a platí $\gamma \in (0, 1)$.



Konštanta zabúdania



Ohodnotenie ciest :

- $S(S0, S2, S3) = 2 + 0.9 * 0.4 = 2.36$
- $S(S0, S1, S3) = 1 + 0.9 * 1 = 1.9$
- $S(S0, S1, S1, S1) = 1 + 0.9 * (1 + 0.9 * 1) = 2.71$
- $S(S0, S1, S1, S1, S1) = 1+0.9*(1+0.9*(1+0.9*1)) = 3.439$
- $S(S0, S1, S1, S1, S1, S1, ...) = 10 <\!\!-\!\!-$
- $S(S0, S1, S1, S1, S1, S1, ..., S3) = 10.8 <\!\!-\!\!-$

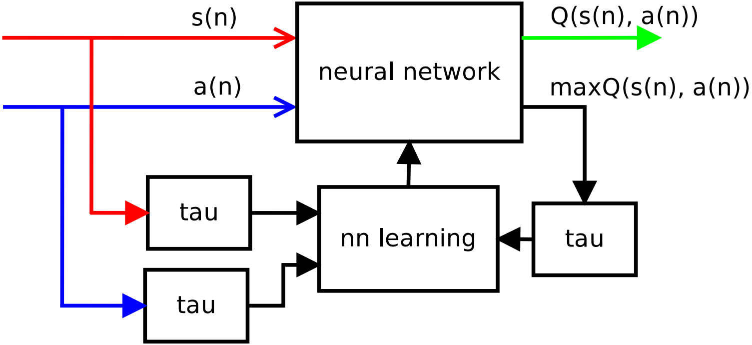
Implementačné problémy

Problémy tabuľkovej interpretácie $Q(s(n), a(n))$:

1. pre veľké počty stavov, mnohorozmerné stavové priestory alebo veľký počet akcií narastajú pamäťové nároky
   - robot s 20 senzormi kde každý má 256 hodnôt sa môže nachádzať v $1.46 * 10^{48}$ stavoch
   - odhadovaný počet atómov v pozorovateľnom vesmíre je 10^{80}
2. o nevyplnených $Q(s(n), a(n))$ nevieme povedať nič,
3. pre rozsiahle stavové priestory ťažko vypočítateľné,
4. ako aproximovať $Q(s(n), a(n))$?

Aproximácia doprednou perceptronovou neurónovou sieťou

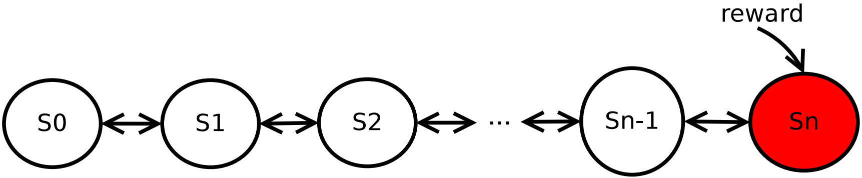
Utopická predstava :



Prečo nedáva správne výsledky?

Hypotéza

Na základe experimetov - Snowball problém



Pre korektné vyplnenie hodnôt v s_{n-1} sa vyžaduje korektá hodnota v s_n

$$Q(s(1), a(1)) = R(s(1), a(1)) + \gamma \max_{a(0) \in \mathbb{A}} Q(s(0), a(0))$$

$$Q(s(2), a(2)) = R(s(2), a(2)) + \gamma \max_{a(1) \in \mathbb{A}} Q(s(1), a(1))$$

$$\ldots$$

Rozklad $Q(s(n), a(n))$ na bázické funkcie

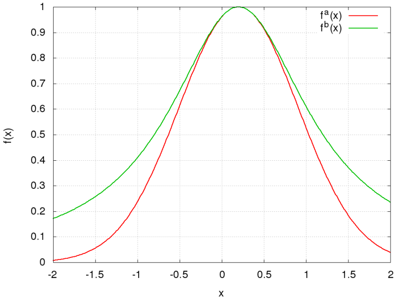
Vzhľadom na charakter učiaceho algoritmu

$$Q(s(n), a(n)) = R(s(n), a(n)) + \gamma \max_{a(n-1) \in \mathbb{A}} Q(s(n-1), a(n-1))$$

boli zvolené bázické funkcie

$$f_j^1(s, a) = e^{-\sum_{i=1}^{n_s} \beta_{aji}(s_i - \alpha_{aji})^2}$$

$$f_j^2(s, a) = \frac{1}{1 + \sum_{i=1}^{n_s} \beta_{aji}(s_i - \alpha_{aji})^2}$$



a ich lineárna kombinácia

$$Q^x(s(n), a(n)) = \sum_{j=1}^{l} w_{ja(n)} f_j^x(s(n), a(n))$$

Nová bázická funkcia

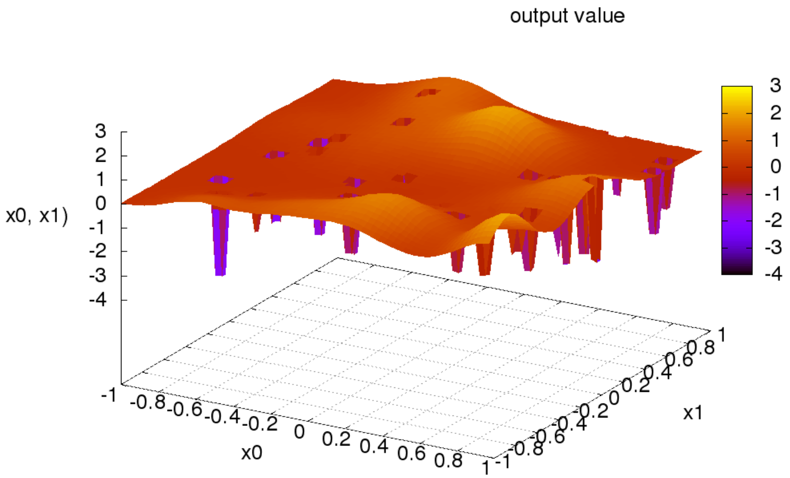
Tabuľka pre vybrané hodnoty - umožní zachytit skokovú zmenu
Gaussova krivka - dokáže pokryť nenulovými hodnotami celý definičný obor

$$P_i(s(n), a(n)) = \begin{cases} r_{ai} & ak \;\; s(n) = \alpha_i^1 \\ 0 & inak \end{cases} \tag{1}$$

$$H_j(s(n), a(n)) = w_{aj} e^{-\beta_{aj} \sum_{i=1}^{n_s} (s_i(n) - \alpha_{aji}^2)^2} \tag{2}$$

$$Q(s(n), a(n)) = \sum_{i=1}^{I} P_i(s(n), a(n)) + \sum_{j=1}^{J} H_j(s(n), a(n)) \tag{3}$$

Nová bázická funkcia



Bloková schéma syntézy testovaného riešenia

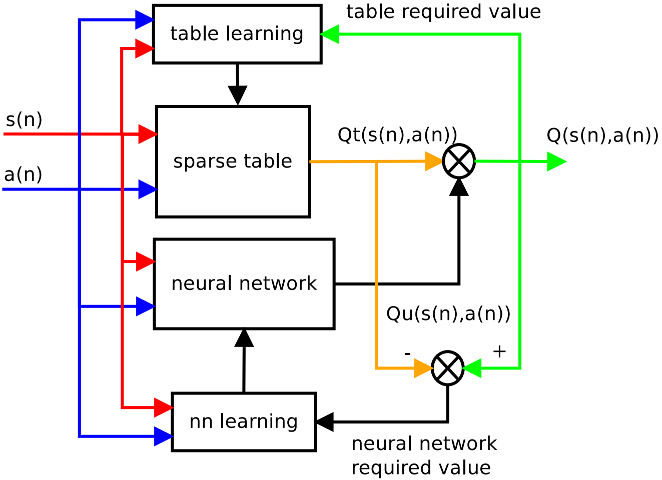
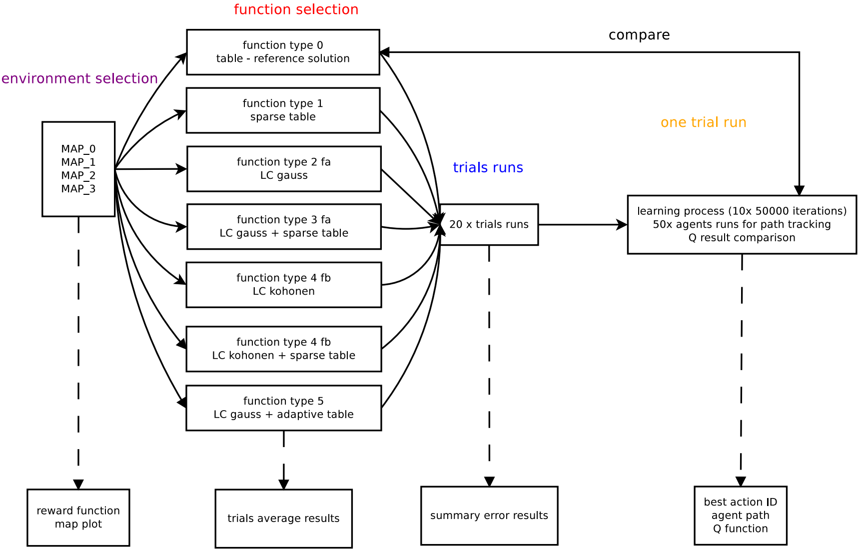


Schéma priebehu experimentov



Návrh experimentov - podmienky

- 50000 iterácií učenia
- rozmer s je $n_s = 2$, rozmer a je $n_a = 2$
- predpis funkcie ohodnotení

$$
\begin{aligned}
&Q(s(n), a(n)) = \\
&\alpha Q(s(n-1), a(n-1)) \\
&(1-\alpha)(R(s(n), a(n)) + \gamma \max_{a(n-1) \in \mathbb{A}} Q(s(n-1), a(n-1))
\end{aligned}
$$

- $R(s(n), a(n)) \in \langle -1, 1 \rangle$ náhodné prostredie (mapa) s 1 cieľovým stavom
- $\gamma = 0.98$ a $\alpha = 0.7$
- hustota referenčného riešenia = 1/32 (4096 stavov)
- počet akcií v každom stave = 8
- hustota riedkej tabuľky = 1/8 (1:16 pomer)
- počet bázických funkcií $I = 64$
- rozsah parametrov $\alpha_{ja}(n)$, $\beta_{ja}(n)$, $w_{ja}(n)$

Návrh experimentov - podmienky

$Q_{rt}(s(n), a(n))$ referenčná funkcia Q (funkcia 0), kde $t \in \langle 0, 19 \rangle$ je číslo trialu
$Q_{jt}(s(n), a(n))$ testované funkcie Q a $j \in \langle 1, 5 \rangle$.
Celková chyba behu trialu t je

$$e_{jt} = \sum_{s,a} (Q_{rt}(s, a) - Q_{jt}(s, a))^2$$

priemerná, minimálna, maximálna chyba a smerodatná odchylka

$$\bar{a}_j = \frac{1}{20} \sum_t e_{jt}$$

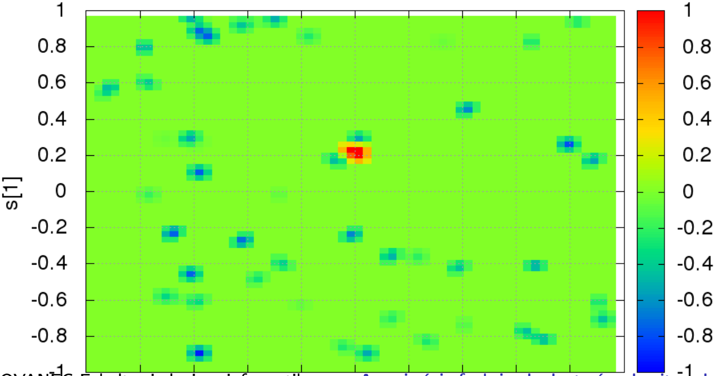
$$e_j^{min} = \min_t e_{jt}$$

$$e_j^{max} = \max_t e_{jt}$$

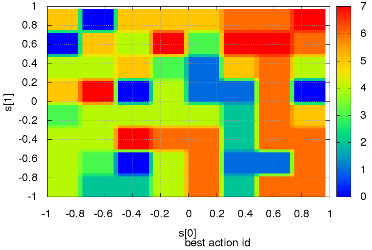
$${\sigma_j}^2 = \frac{1}{20} \sum_t (\bar{a}_j - e_{jt})^2$$

Funkcia R(s, a), prostredie 1 - Výsledky experimentov

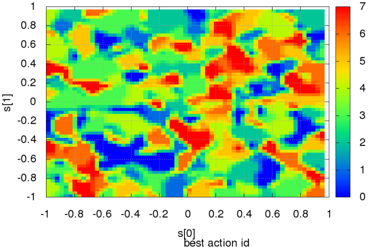
pre každý stav je zvolená rovnaka množina akcií.
Ďalej platí $s = (s[0], s[1])$.



Funkcia voľby najlepšej z 8 akcií v stave $s = (s[0], s[1])$.

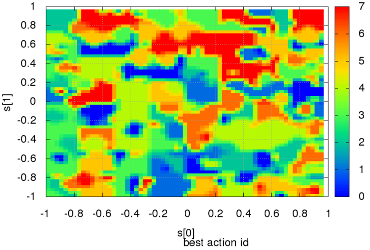


Obr.: sparse table

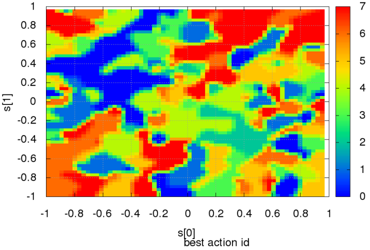


Obr.: linear combination Gauss

Funkcia voľby najlepšej z 8 akcií v stave $s = (s[0], s[1])$.



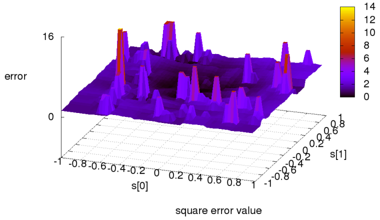
Obr.: sparse table + linear combination Gauss



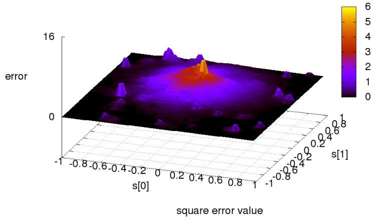
Obr.: linear combination Kohonen function

Chybové funkcie - Výsledky experimentov

$$e_{jt}(s) = (Q_{rt}(s,a) - Q_{jt}(s,a))^2$$



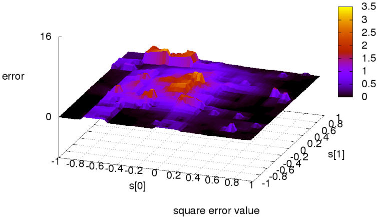
Obr.: sparse table



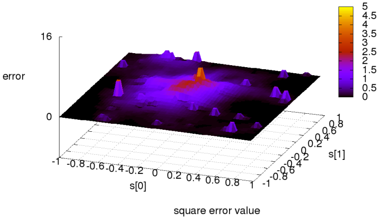
Obr.: linear combination Gauss

Chybové funkcie - Výsledky experimentov

$$e_{jt}(s) = (Q_{rt}(s,a) - Q_{jt}(s,a))^2$$

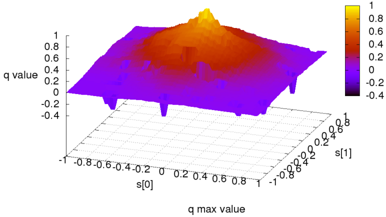


Obr.: sparse table + linear combination Gauss

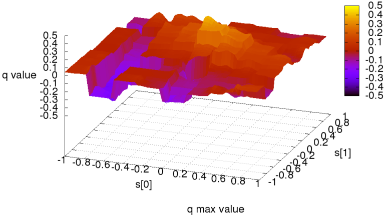


Obr.: linear combination Kohonen function

max Q(s, a) - Výsledky experimentov

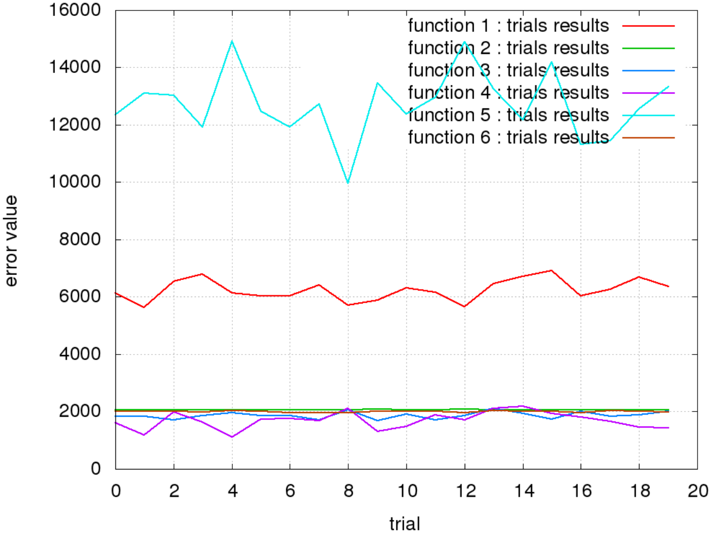


Obr.: reference table

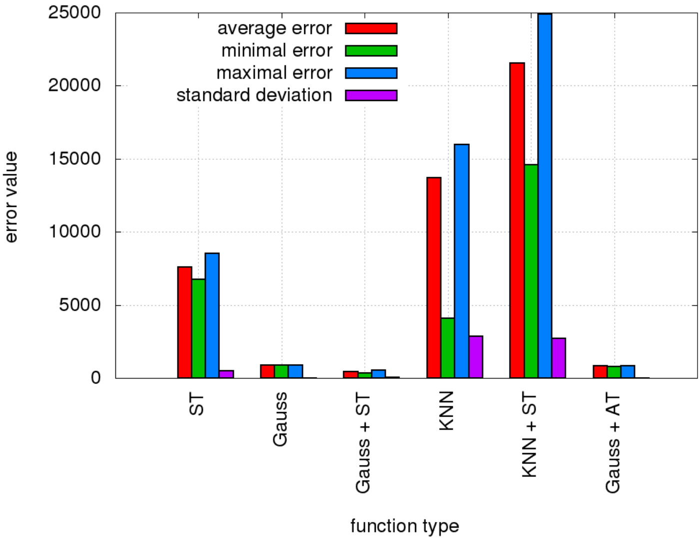


Obr.: sparse table + linear combination Gauss

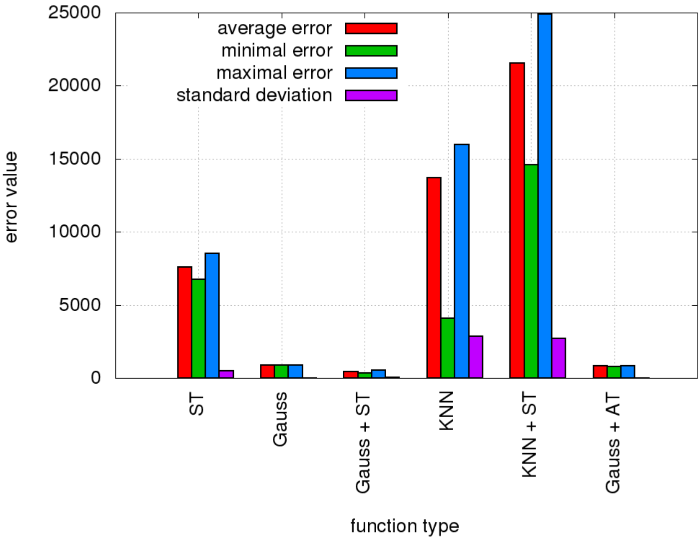
Priebeh trialov - Výsledky experimentov



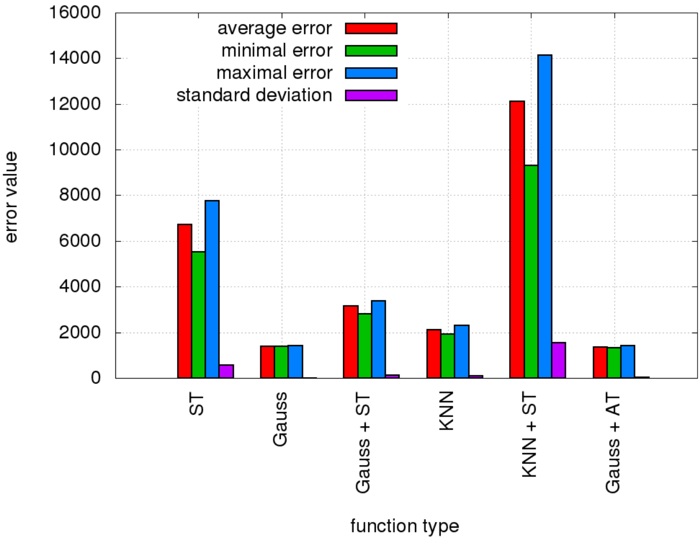
Prostredie 0 - Výsledky experimentov



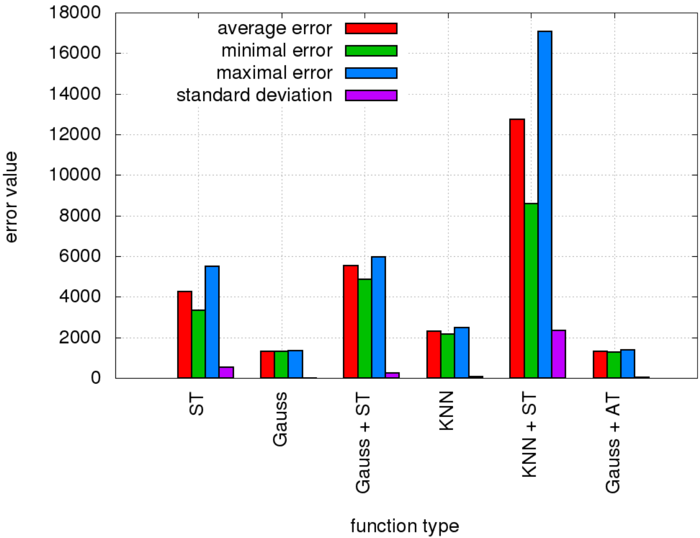
Prostredie 1 - Výsledky experimentov



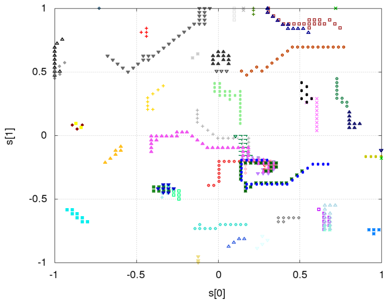
Prostredie 2 - Výsledky experimentov



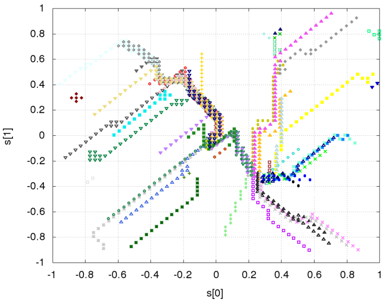
Prostredie 3 - Výsledky experimentov



Porovnanie s ostatnými

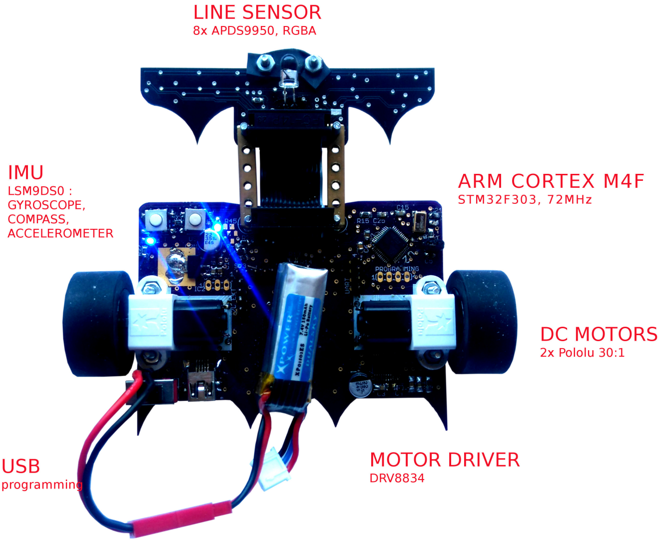


Obr.: Dráha robotov, funkcia 2 - Gauss



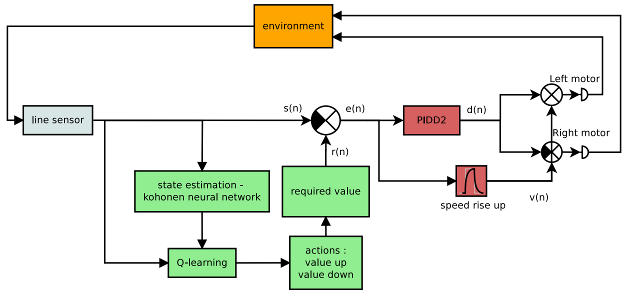
Obr.: Dráha robotov, funkcia 6 - Peak and Hill

Doplnkový experiment - robot



Doplnkový experiment - robot

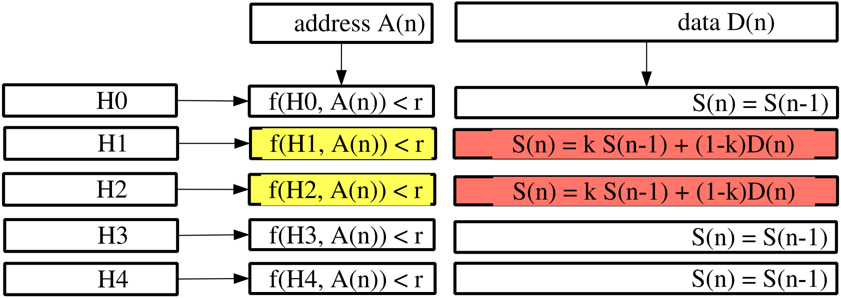
Bloková schéma riadiaceho bloku robota



Ďalšie smerovanie - sparse distributed memory (SDM)

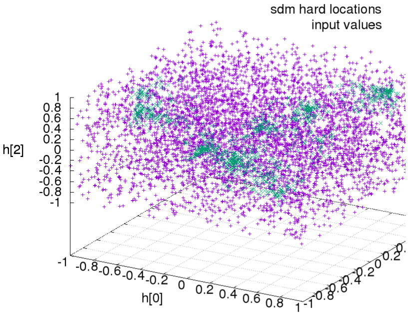
sparse distributed memory - Pentti Kenerva 1988
hierarchical temporal memory - Jeff Hawkins 2012
adaptive sparse distributed memory - Michal Chovanec 2016

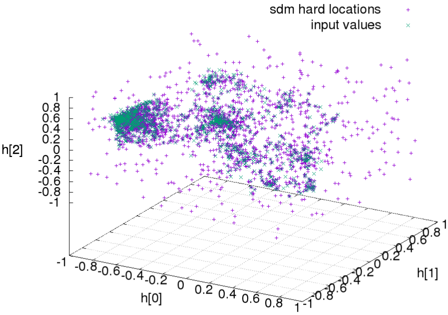
aproximácia $D(n) = f(A(n))$



Adaptive sparse distributed memory

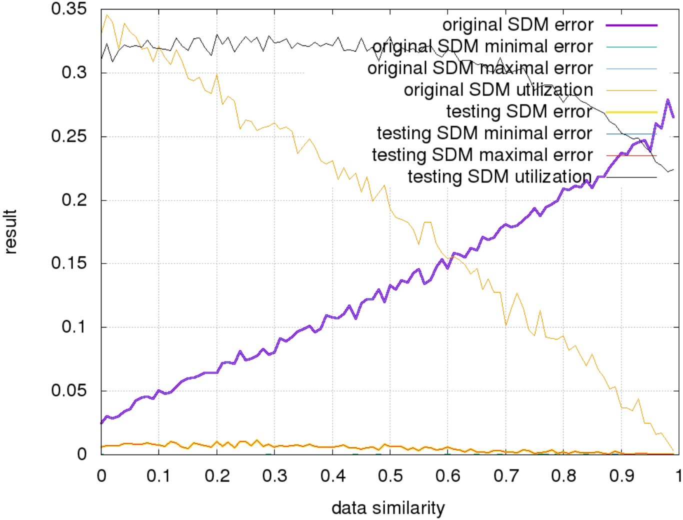
Pokrytie priestoru s hard locations





Experimentálne výsledky

- aproximácia funkcie, predikcia časových radov, MNIST - rukou písané číslice



Ďakujem za pozornosť

michal.chovanec@yandex.ru
https://github.com/michalnand/q_learning