ENGINEERING MECHANICS 2004
NATIONAL CONFERENCE
with international participation
Svratka, Czech Republic, 10 - 13 May 2004

CONTINUOUS Q-LEARNING APPLICATION

S. Věchet[1], J. Krejsa[2], P. Miček[3]

Standard algorithm of Q-Learning is limited by discrete states and actions and Q-function is usually represented as discrete table. To avoid this obstacle and extend the use of Q-learning for continuous states and actions the algorithm must be modified and such modification is presented in the paper. Straightforward way is to replace discrete table with suitable approximator. A number of approximators can be used, with respect to memory and computational requirements the local approximator is particularly favorable. We have used Locally Weighted Regression (LWR) algorithm. The paper discusses advantages and disadvantages of modified algorithm demonstrated on simple control task.

1. Úvod

Standardní algoritmu Q-učení je omezen používáním diskrétních stavů a akcí. V tomto případě je Q-funkce nejčastěji representována jako diskrétní tabulka Q-hodnot. Diskretizace spojitých hodnot je často spojena s četnými problémy. Vyhnutí se problému s diskretizací stavů a akcí je možno dosáhnout použitím spojitého Q-učení. Tento článek se snaží přiblížit metody použité za účelem kontinualizace diskrétního algoritmu Q-učení a to pomocí lokálních aproximátorů. Jako vhodný aproximátor byl zvolen algoritmus patřící do skupiny aproximátorů s lokálním modelem (Locally Weighted Learning - LWL)(Atketson &Moore & Shaal, 1996). Z této skupiny tzv. paměťově orientovaných aproximátorů byla pro jednoduchou implementaci zvolena metoda lokálně vážené regrese(Locally Weighted Regression - LWR)

2. Lokálně vážená regresse

Mezi nejjednodušší algoritmy s lokálním lineárním modelem patří lokálně vážená regrese (Locally Weighted Regression – LWR). Tato metoda je založena na známé metodě nejmenších čtverců. V tomto případě není regresní přímka konstruována na základě celého datového souboru, ale pouze z bodů v určité definované oblasti a závisí na vzájemné poloze jednotlivých bodů dané oblasti. Metoda nejmenších čtverců pouze minimalizuje jednoduché kritérium, kterým je součet čtverců vzdálenosti: $C = \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2$ Kde n je počet bodů, y_i reprezentuje danou souřadnici i-tého bodu, $\hat{y}_i$ je odhad. Při použití lineárního modelu jako polynomu prvního stupně $y = p_1 x + p_2$, kde p_1,p_2 jsou neznámé parametry, dostaneme pro n bodů o souřadnicích x_i,y_i, n rovnic o dvou neznámých.

[1] Ing. Stanislav Věchet: Institute of Mechanics, Mechatronics and Biomechanics, FME, BUT, Technická 2; 616 69 Brno; e-mail: vechet@umtn.fme.vutbr.cz
[2] Ing. Jiří Krejsa PhD: Institute of Thermomechanics, CAS, Mechatronics Centre Brno, Technická 2, 616 69, Brno; e-mail: jkrejsa@umt.fme.vutbr.cz
[3] Ing. Pavel Miček: Institute of Mechanics, Mechatronics and Biomechanics, FME, BUT, Technická 2; 616 69 Brno; e-mail: micek@umtn.fme.vutbr.cz

Maticově lze tedy psát

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}$$

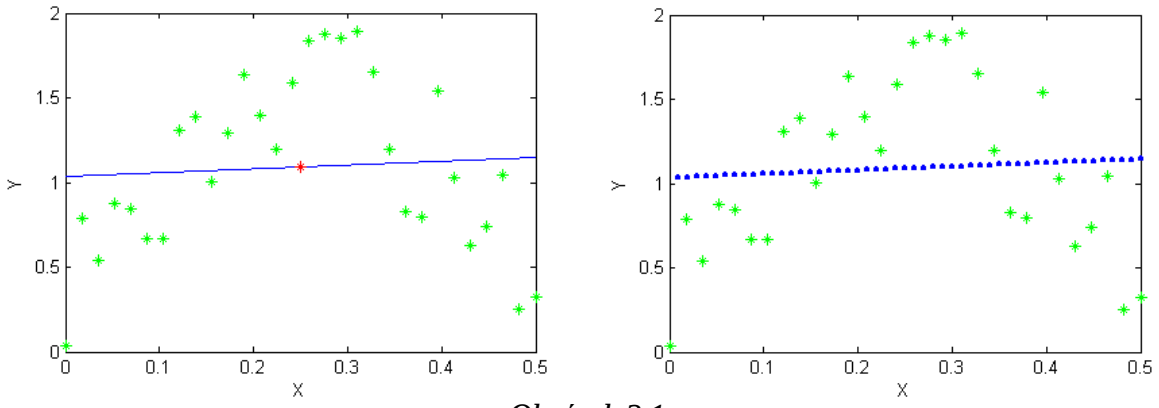
$$\begin{bmatrix} y_1 \\ y_2 \\ \mathrm{M} \\ y_n \end{bmatrix} = \begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ \mathrm{M} & \mathrm{M} \\ x_n & 1 \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \end{bmatrix}$$

kde **y** je matice výstupů $\begin{bmatrix} y_1 & y_2 & \mathbf{L} & y_n \end{bmatrix}^T$, **X** je matice vstupů $\begin{bmatrix} x_1 & x_2 & \mathbf{L} & x_n \\ 1 & 1 & \mathbf{L} & 1 \end{bmatrix}^T$ a b je matice hledaných parametrů. Rovnici lze tedy řešit následujícím způsobem

$$\mathbf{X}^T\mathbf{y} = \left(\mathbf{X}^T\mathbf{X}\right)\boldsymbol{\beta}$$

$$\boldsymbol{\beta} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$$

Po těchto úpravách dostaneme matici parametrů $\boldsymbol{\beta} = \begin{bmatrix} p_1 & p_2 \end{bmatrix}^T$ a můžeme tedy pro libovolný neznámý bod q[x,y] psát: $y = xp_1 + p_2$. Množinou bodů leze tedy proložit regresní přímku, případně lze na základě parametrů p_1,p_2 odhadovat neznáme body q jednotlivě jak ukazuje obrázek 2.1.



Obrázek 2.1.

Oproti tomu metoda LWR bere v úvahu i vzájemnou polohu bodů. Parametry regresní přímky jsou přímo závislé na vzdálenosti od odhadovaného bodu q. Nejčastějším kritériem vzdálenosti je Euklidovská vzdálenost $d_E([x_i,y_i],\mathbf{q})$. V tomto jednoduchém ilustračním případě je vzdálenost daná pouze jednoduchým vztahem

$$d_E\left(\left[x_i, y_i\right], \mathbf{q}\right) = \sqrt{\left(x_i - x\right)^2}$$

Na základě takto získané vzdálenosti je poté vypočítána váha každého bodu. Váha určuje jaký vliv bude mít daný bod na parametry regrese. Tedy blízký bod bude regresi ovlivňovat více než bod vzdálený. Jako nejvhodnější funkce pro výpočet váhy danného bodu bývá označována funkce založena na gausově jádru. V našem případě by pak taková funkce mohla vypadat takto:

$$K\left(d_E\right) = e^{-d_E^2}$$

Postup výpočtu neznámých parametrů u metody lokálně vážené regrese se tedy změní následujícím způsobem.

Nejdříve je třeba vypočítad vzdálenost a váhu jednotlivých bodů

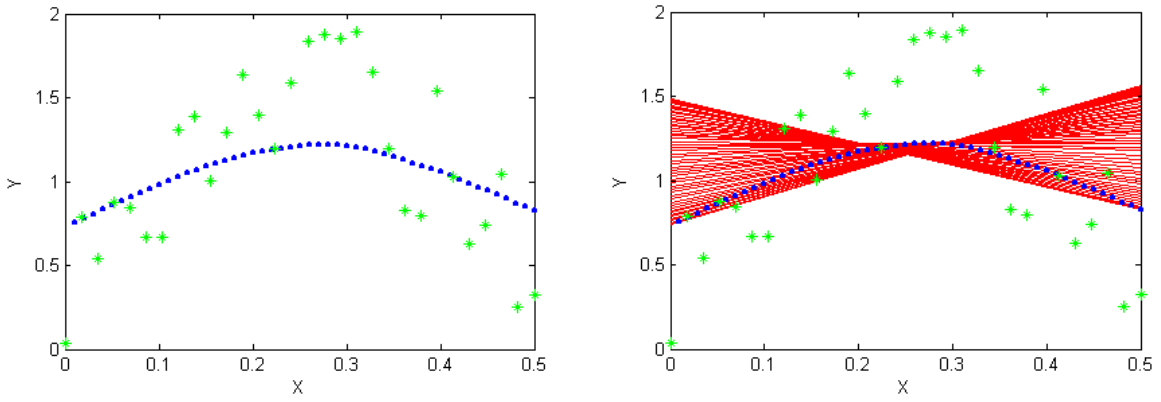
$$w_i = \sqrt{K\left(d_E\left([x_i, y_i], \mathbf{q}\right)\right)}$$

Pro další výpočty je nutné tyto váhy umístit do diagonalní matice **W** kde jednotlivé prvky na digonále odpovídají jednotlivým vypočítaným vahám $W_{ii} = w_i$. Na základě takto vypočítané matice vah přepočítáme známou matici vstupů **X**, získáme tak váženou matici vstupů **Z** = **W X**. Vlastní výpočet neznámých parametrů lze poté zapsat jako

$$\mathbf{Z}^T\mathbf{y} = \left(\mathbf{Z}^T\mathbf{X}\right)\boldsymbol{\beta}$$

$$\boldsymbol{\beta} = \left(\mathbf{Z}^T\mathbf{X}\right)^{-1}\mathbf{Z}^T\mathbf{y}$$

Takto vypočítanou matici parametrů $\boldsymbol{\beta} = [\,p_1 \quad p_2\,]^T$ lze opět stejným způsobem použít pro proložení regresní přímky, případně odhadovat jednotlivě neznáme body. Obrázek 2.2 nalevo ukazuje metodu LWR použitou pro aproximaci stejného vstupního datového souboru jaký byl použit při demonstraci metody nejmenších čtverců. Na první pohled je jasně vidět výrazně věrnější aproximaci dat. Obrázek 2.2 napravo ukazuje jak se liší parametry jednotlivých regresních přímek pro jednotlivé odhady, což bylo podrobněji popsáno v předcházejících odstavcích.



Obrázek 2.2

3. Spojité Q-učení

Standardní model Q-učení je tvořen agentem a prostředím. Prostředí je tvořeno množinou diskrétních stavů, které jsou většinou získány diskretizací spojitých stavových veličin. V každém časovém okamžiku *t* se prostředí nachází ve stavu s_t . Agent má k dispozici množinu akcí, kterými stav prostředí ovlivňuje. Poté co agent provede akci a_t způsobí změnu stavu prostředí na stav s_{t+1}. Jednou z možností jak specifikovat požadované chování agenta je definovat funkci okamžitého posílení $r(s_t,s_{t+1},a_t)$, která určuje konkrétní odměnu/pokutu za přechod ze stavu s_t do stavu s_{t+1} při provedení akce a_t. Dlouhodobý cíl agenta je možné definovat jako úkol maximalizovat funkci okamžitého posílení. Tato funkce může být representována jako diskrétní tabulka *Q(s,a)*, kde každé vzájemné kombinaci stavů a akcí odpovídá jedna Q-hodnota. Pokud je tedy Q-funkce implementována jako tabulka, je přepočtový vztah dán jako

$$Q(s_t, a_t) = Q(s_t, a_t) + a\left[r(s_t, a_t, s_{t+1}) + g \max_{a_t} Q(s_{t+1}, a_t) - Q(s_t, a_t)\right]$$

kde α je učící poměr, γ srážkový faktor. Při tomto způsobu implementace Q-funkce je také důležité správné rozdělení mřížky tabulky. Velmi časté je použití nelineárního rozdělení

mřížky, kdy v okolí důležitých oblastí je mřížka vzorkována hustěji než v méně důležitých oblastech. Tabulková implementace Q-funkce je vhodná pro soustavy s menším počtem stavových proměnných (se zvyšujícím se počtem stavových proměnných neúměrně narůstá výpočetní náročnost), nebo tam kde postačuje omezený počet akcí, kterými lze ovlivňovat chování soustavy. Pokud je ovšem potřeba pracovat s větší počtem stavových proměnných, popř. je nutné pracovat se spojitými stavy a akcemi, je nutné Q-funkci implementovat jiným způsobem než klasickou tabulkou. Jako možnost se jeví nahradit tabulku Q-hodnot vhodným aproximátorem. Aproximátor je poté používán pro odhady Q-hodnot v neznámých stavech prostředí. V našem případě byl jako vhodný aproximátor použita metoda lokálně vážené regrese. V tomto případě již přepočet Q-hodnot v jednotlivých krocích učení není dán jednoduchým vztahem, i když jeho základ je možné vysledovat i v modifikovaném algoritmu. Tuto metodu již nelze jednoduše popsat jedním vztahem a proto je popsána kódem pseudoalgoritmu 3.1.

Algoritmus 3.1. Modifikovaný algoritmus Q-učení

1. $q \leftarrow Q_{estim}(s_t, a_t)$ metoda LWR
2. $q' \leftarrow \max_a Q(s_{t+1}, a_t, p)$
3. $K \leftarrow$ použité body pro aproximaci q
4. $W_{ii} \leftarrow \exp(-(\mathbf{q} - \mathbf{k}_i)^2 / h^2)$
5. $\Delta q \leftarrow a\,(r + g\,q' - q)$
6. $Q(s,a) \leftarrow Q(s,a) + \Delta q$
7. aktualizace pro každý bod $Q(s_i, a_i)$ v K; $Q(s_i, a_i) \leftarrow Q(s_i, a_i) + \Delta q w_i$

kde p je šířka pásma, ze kterého se vybírají platné Q-hodnoty, h označuje efektivní okolí, tj. označuje velikost oblasti ve které mají okolní body významný vliv na výpočet váhy jednotlivých bodů. Právě volba těchto parametrů má klíčový význam na kvalitu naučení.

4. Provedené experimenty

Pro experimenty se spojitým Q-učením byl použit jednoduchý matematický model inversního kyvadla daný těmito rovnicemi

$$(M+m)\ddot{x} + ml\ddot{\varphi}\cos\varphi - ml\dot{\varphi}^2\sin\varphi = F$$

$$\left(I + ml^2\right)\ddot{\varphi} + mgl\sin\varphi = ml\ddot{x}\cos\varphi$$

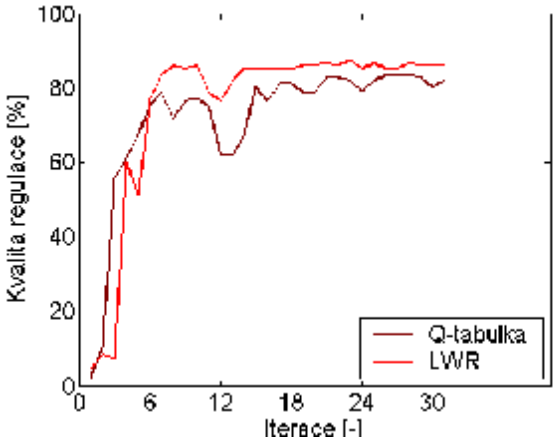
kde M je hmotnost vozíku, m hmotnost kyvadla, l délka kyvadla, I moment setrvačnosti kyvadla, F síla působící na vozík, g tíhové zrychlení, x souřadnice vozíku, φ úhel odklonu kyvadla od vertikální osy.

Byla zanedbána hmotnost vozíku a tření. Jako akce bylo použito zrychlení se kterým se pohybuje vozík a jako stavu úhel odklonu kyvadla od vertikální osy. Cílem agenta-řídícího členu je udržet kyvadlo ve vymezeném intervalu $\langle -\varphi_{max}, \varphi_{max} \rangle$. Akce byly voleny z intervalu $\langle -a, a \rangle$. Posilovací funkce stanovuje odměnu/trest z množiny $\{-1,0,1\}$ následujícím pravidlem:

$$r = \begin{cases} 1 & \varphi \le |\varphi_{\min}| \\ 0 & \varphi \in \left(|\varphi_{\min}|, |\varphi_{\max}|\right\rangle \\ -1 & jinak \end{cases}$$

Pro experimenty bylo použito těchto parametrů modelu: $m=0.5kg$, $g=9.81kgms^2$, $l=0.2m$ a parametrů učení $a=10ms^2$, $j_{min}=0.1°$, $j_{max}=20°$, $a=0.2$, $g=0.99$. Testován byl algoritmus klasického Q-učení oproti spojitému Q-učení s LWR.

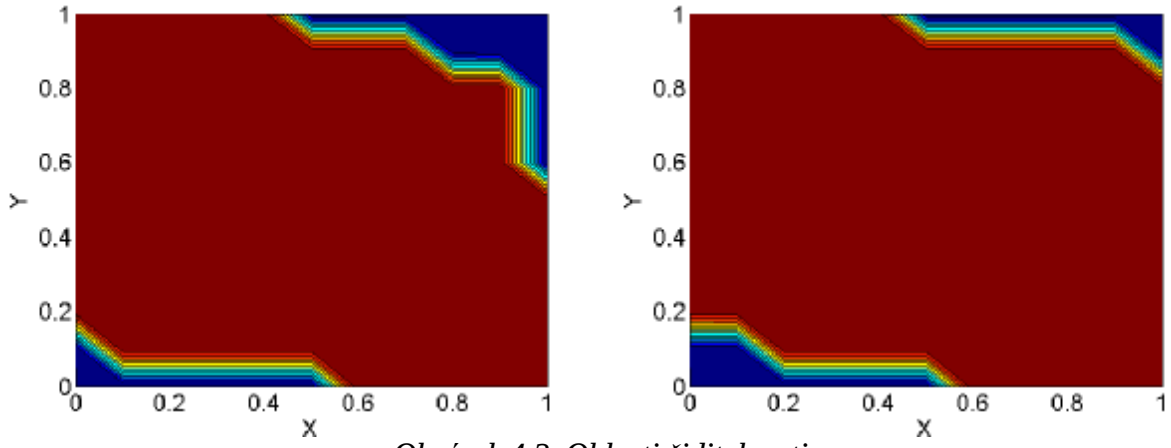
Systém byl testován následujícím způsobem: Nejprve bylo vygenerováno náhodně *n* počátečních stavů. Tyto stavy byly použity během testování a bylo sledováno, zda po stanovený počet kroků dokáže agent udržet výchylku kyvadla v určeném intervalu. Pokud to agent dokázal, byl pokus vyhodnocen jako úspěšný. Tím byla stanovena procentuální úspěšnost v průběhu učení(kvalita regulace).
Simulační výsledky jsou zobrazeny v na grafech 4.1 a 4.2. Graf 4.1 ukazuje průběh učení, graf 4.2 ukazuje oblasti řiditelnosti, levá část pro diskrétní, pravá pro spojité Q-učení. Na osách *x* a *y* jsou vyneseny normalizované hodnoty úhlu odklonu od vertikální osy a úhlové rychlosti. Při použití diskrétního Q-učení byla použita tabulka o velikosti (14,14,3), kde jednotlivá čísla označují počet diskrétních hodnot pro (postupně z leva) úhel, rychlost, zrychlení. Při použití spojitého Q-učení byly rozsahy jednotlivých stavů a akcí normalizovány tak že jejich rozsahy byli vždy v intervalu <0,1>. Tabulka 4.1 ukazuje srovnání procentuální velikosti oblastí řiditelnosti pro jednotlivé metody.



Obrázek 4.1

Metoda	Kvalita Regulace
Q-tabulka	85%
LWR	88%

tabulka 4.1



Obrázek 4.2. Oblasti řiditelnosti

5. Závěr

Abychom mohli použít spojitých stavů a akcí v konvenčním Q-učení je nutné použít vhodný aproximátor. Jako první krok při použití spojitých stavů a akcí jsme zvolili LWR. Algoritmus pro aproximaci Q-funkce patřící do skupiny algoritmů s lokálním modelem. Zvolená metoda byla testována na jednoduchém modelu inverzního kyvadla.

V experimentech byla Q-tabulka kompletně nahrazena aproximátorem. Takto upravená metoda Q-učení byla testována na zmíněném modelu inverzního kyvadla. Z experimentů je patrno že kvalita regulace takto modifikovaného algoritmu je srovnatelná se standardní metodou. Její mírnou nevýhodou je pomalejší konvergence v počátečních fázích učení, což je způsobeno počáteční absencí Q-hodnot v prostoru. Jelikož tato vlastnost nemá vliv na další doučování, je možné ji považovat za méně závažnou oproti problémům se správnou diskretizací tabulky u diskrétního Q-učení, špatná diskretizace má totiž negativní vliv na učení po celou dobu trénování, tedy nejen při počátečních fázích učení, ale i v dalších fázích, jako je například on-line doučování na reálném modelu.

Jako další stupeň, v poznávání vlastností spojitého Q-učení, se jeví nahrazení jednoduchého aproximátoru LWR, nějakým vhodnějším aproximátorem s lepšími vlastnostmi a také použití složitějšího simulačního modelu, jakým může být např. aktivní magnetického ložisko.

6. Poděkování

Tato práce je podporována projektem MSM 262100024 ministerstva vzdělávání „Výzkum a vývoj mechatronických soustav”

7. Literatura

Atkeson C.G., Moore A.W., Schaal S.: Locally Weighted Learning, Technical Report, ATR Human Information Processing Laboratories, Japan, 1996

Schaal, S. & Atkeson, C. G.. “Receptive Field Weighted Regression.” Technical Report TR-H-209, ATR Human Information Processing Laboratories, Japan, 1997