Reinforcement Learning in the Artificial Intelligence

Prep. Horia-Ioan TOMA
Catedra de Informatică Economică, A.S.E. Bucureşti

Reinforcement Learning represents a machine-learning paradigm that is particularly well suited for use on mobile systems as well as an approach on how an agent deals with context (environment). This paper gives a short review on reinforcement learning model and surveys techniques of action selection and the reward associated with state transition. The Markov Decision Process Model is also presented together with two algorithms for selecting the optimal policy.
Keywords: *reinforcement learning, agent, state transition, action selecting, optimal policy, reward function.*

Introducere

În acest articol vom insista asupra noţiunii de "instruire prin întărire", metodă specifică inteligenţei artificiale orientată exclusiv asupra învăţării prin atingerea de obiective (scopuri), diferenţiindu-se astfel de alte abordări legate de instruire.
Instruirea prin întărire înseamnă învăţarea a ceea ce trebuie făcut, alegerea acţiunilor în scopul maximizării rezultatelor obţinute, [2]. Cel care învaţă nu este îndrumat spre acţiunile pe care trebuie să le efectueze, în schimb trebuind să descopere care dintre aceste acţiuni ar aduce cel mai important rezultat prin încercarea lor. Problema instruirii prin întărire aparţine unui agent care trebuie să înveţe un comportament prin interacţiuni de tipul încercare-eşec (*trial-and*-error) într-un mediu dinamic.
Astfel, caracteristicile definitorii ale instruirii prin întărire sunt, pe de o parte, căutarea şi încercarea diferitelor acţiuni din spaţiul celor posibil de efectuat, iar pe de altă parte obţinerea rezultatelor în urma selectării acţiunilor, rezultate care pot să se manifeste nu doar imediat ci în urma parcurgerii unor paşi următori.
Instruirea prin întărire este diferită de metoda instruirii supervizate utilizată curent în inteligenţa artificială (de exemplu, în cazul reţelelor neuronale artificiale). Instruirea supervizată înseamnă învăţarea pe baza unor exemple disponibile a-priori. În cazul problemelor interactive este deseori dificil de obţinut astfel de exemple corecte şi reprezentative pentru comportamentele care trebuie urmate, mai ales în mediul incert.
Pentru rezolvarea problemelor utilizând metoda instruirii prin întărire există, în principal, două strategii conform [4]. Prima dintre ele constă într-o căutare în spaţiul comportamentelor posibile pentru a-l determina pe cel care se potriveşte cel mai bine mediului. Această abordare este specifică în cazul algoritmilor genetici. A doua strategie, tratată în continuare, presupune utilizarea unor tehnici statistice şi a unor metode de programare dinamică pentru estimarea utilităţii acţiunilor efectuate.

Modelul instruirii prin întărire

Se presupune că indiferent de detaliile senzoriale, memorare, control şi de obiectivul general care trebuie atins în cadrul unei probleme, comportamentul unui agent poate fi redus la trei tipuri de semnale (informaţii) transmise între agent şi mediul înconjurător. Primul semnal este reprezentat de alegerile făcute de agent (acţiuni). Al doilea semnal oferă informaţii asupra deciziilor de a selecta o acţiune (stări), iar al treilea semnal defineşte scopul agentului (rezultatul sau utilitatea). În mod obişnuit, se presupune că agentul recepţionează o imagine corectă şi completă asupra mediului. Ca în atâtea domenii ale inteligenţei artificiale, reprezentările celor trei semnale ţin mai degrabă de calităţile umane ale celui care încearcă să rezolve problema şi mai puţin de fundamentarea ştiinţifică.
În modelul standard de instruire prin întărire,

un agent este conectat cu mediul înconjurător prin intermediul percepției şi al acțiunii. La fiecare pas de interacțiune cu mediul, agentul primeşte din partea mediului un input şi indicații despre starea curentă. Pe baza acestor informații, poate fi selectată o acțiune care va genera un output. Acțiunea modifică starea mediului iar valoarea acestei tranziții este retransmisă agentului. Ceea ce trebuie să realizeze, în ansamblu, un agent este să descopere o anumită "politică" de urmat care constă în maparea stărilor în acțiuni care să maximizeze utilitatea pe termen lung. Acest lucru trebuie realizat în condițiile unui mediu nondeterministic, însemnând că asumarea unor aceleaşi acțiuni la diferite momente de timp poate avea consecințe diferite.

Acțiunile de efectuat pot fi caracterizate în funcție de nivelul de decizie implicat. Astfel, deciziile la nivel redus sunt cele care implică doar transmiterea unor informații, iar cele de nivel înalt implică o selecție dificilă. Stările pot, de asemenea, avea diferite niveluri de complexitate de la determinarea valorii unor senzori până la, spre exemplu, descrierea obiectelor existente într-o cameră.

Alegerea acțiunilor de efectuat pe baza utilității obținute

Una dintre provocările instruirii prin întărire constă în legătura dintre *explorarea* unor situații şi *exploatarea* lor pentru maximizarea utilității. Pentru a obține un rezultat important în urma unei acțiuni desfăşurate, un agent care utilizează instruirea prin întărire ar trebui să prefere un comportament similar celui ales în trecut şi despre care se cunoaşte că a adus un beneficiu important dar este la fel de necesară şi încercarea altor acțiuni care s-ar putea dovedi mai eficiente. Astfel, agentul desfăşoară o acțiune de exploatare a ceea ce deja cunoaşte şi o acțiune de explorare pentru a putea alege mai bine în viitor. Pentru a putea face față acestor situații, agentul are definite scopuri precise, poate conştientiza diferite aspecte legate de mediul înconjurător şi poate face alegeri în urma cărora selectează acțiuni de efectuat care vor influența mediul.

Rezultatul obținut în urma desfăşurării unei acțiuni oferă informații despre cât de eficientă a fost respectiva acțiune, dar nu spune nimic despre faptul că alegerea a fost corectă sau incorectă (a fost sau nu selectată cea mai bună variantă posibilă). Corectitudinea alegerii unei acțiuni este o proprietate relativă care poate fi determinată cu exactitate doar dacă ar fi încercate toate acțiunile şi apoi comparate rezultatele în ansamblu. Acest procedeu poartă numele de învățare prin selecție, în contrast cu învățarea prin instruire pe baza unor exemple, [3].

În abordarea situațiilor pe baza perechii acțiune-rezultat, utilitatea unei acțiuni reprezintă media rezultatelor obținute în trecut, ceea ce presupune menținerea unei statistici asupra selecțiilor efectuate anterior. Atunci când agentul reîntâlneşte o acțiune, el poate să o evalueze pe baza cunoştințelor precedente. Problema care apare în acest caz este cantitatea mare de informație care trebuie memorată şi care nu are, practic, o limită superioară.

Formalizarea rezultatelor obținute în urma selecțiilor efectuate

Dacă mediul înconjurător este caracterizat de instabilitate înseamnă că o aceeaşi acțiune poate avea consecințe diferite. În această situație, agentul trebuie să acorde o importanță sporită utilității curente rezultate în urma selectării acțiunii, în defavoarea utilității observate în trecut. Cum poate însă un agent să recunoască valoarea unui rezultat obținut în urma unei acțiuni? Valoarea utilității curente trebuie comparată cu un nivel de referință standard, numit "utilitatea de referință". Acest nivel poate fi construit pe baza mediei rezultatelor precedente. Metoda din cadrul instruirii prin întărire care se bazează pe o astfel de analiză din partea agentului se numeşte "instruirea prin întărire utilizând comparația". Mai mult, metoda presupune menținerea, alături de rezultatele acțiune-utilitate, a preferințelor pentru a anumită acțiune ca o măsură de probabilitate cu care fiecare acțiune este selectată într-o anumită stare. În acest fel, acțiunea care corespunde, pe baza informațiilor din trecut, unei utilități maximale va avea cea mai mare probabilitate de a fi selectată atunci când este întâlnită.

Rezultatul unei acțiuni poate fi exprimat într-

o valoare numerică care variază de la un pas la altul. Utilizarea unei măsuri de utilitate a fiecărei acţiuni reprezintă o caracteristică distinctă a instruirii prin întărire. Deşi formularea unor scopuri care trebuie atinse în cadrul învăţării poate părea limitată, în practică s-a dovedit că permite flexibilitate şi aplicabilitate. De exemplu, pentru a învăţa un robot cum să se deplaseze, este transmis un semnal de utilitate proporţional cu mişcarea robotului. Pentru a învăţa acelaşi robot cum să iasă dintr-un labirint, rezultatul acţiunilor este zero cât timp robotul nu reuşeşte să găsească soluţia şi 1 atunci când a fost descoperită ieşirea. Dacă robotul se află într-o cameră unde trebuie să identifice un anumit obiect, rezultatul transmis este 1 în caz de reuşită, 0 în majoritatea timpului şi -1 dacă robotul a găsit obiectul nepotrivit.

Uneori, agentul se poate afla într-o situaţie de "percepţie incompletă". De exemplu, dacă un agent răspunde la telefon nu ne putem aştepta ca el să identifice în avans persoana apelantă. În acest caz există informaţii de tip ascuns în mediul înconjurător care ar putea să-l ajute pe agent, dar acesta nu are cum să le obţină. De aceea, rezultatul primit în urma selectării unei acţiuni nu trebuie să penalizeze agentul pentru necunoaşterea unor informaţii importante. Penalizarea ar putea apărea doar în cazul în care agentul s-a confruntat în trecut cu o aceeaşi situaţie, dar "a uitat" între timp.

Procesele de decizii de tip Markov. Algoritmi pentru determinarea politicii optimale

Cele mai multe metode de instruire prin întărire modelează mediul de învăţare sub forma unui proces de decizii de tip Markov (MDP). Un asemenea proces este definit, conform [4], ca o asociere de mai multe variabile: (S, A, T, R, y), unde:

- S reprezintă o mulţime de stări;
- A este mulţimea de acţiuni;
- $T(s'|s,a)$ reprezintă probabilitatea tranziţiei către starea s' când agentul desfăşoară acţiunea a în starea s;
- R este o funcţie stohastică, $R: SxA \to \Re$, care defineşte rezultatul imediat obţinut în urma desfăşurării acţiunii a în starea s. În mod obişnuit, R are o valoare scalară: $\{0,1\}$, dar poate lua valori şi în domeniul real.

În continuare, vom utiliza notaţia r_t care reprezintă valoarea scalară a rezultatului (utilităţii) obţinută după t paşi în viitor.

- y este un factor pozitiv, subunitar care poate fi interpretat în mai multe moduri: factor de actualizare a utilităţii obţinute în viitor, probabilitatea de a ajunge pe o nouă stare sau, pur şi simplu, ca un artificiu matematic de limitare a sumei utilităţii pentru un orizont de timp infinit.

Având în vedere notaţiile de mai sus, obiectivul instruirii prin întărire pentru modelul Markov continuu (orizont de timp infinit) este descoperirea unei politici: $\pi : S \to A$ care să maximizeze utilitatea totală: $\sum_{t=0}^{\infty} y^t r_t$.

Tranziţia între stări este determinată probabilistic pe baza informaţiilor despre starea curentă şi acţiunile agentului.

Markov a definit conceptul de "stări terminale" ca fiind acele stări în care s-a reuşit atingerea obiectivului general sau de unde nu se mai poate evolua spre alte stări. Când un agent ajunge pe una dintre aceste stări, toate acţiunile viitoare vor avea drept rezultat tranziţia spre o stare nulă şi utilităţile asociate vor avea valoarea zero.

Valoarea optimală a unei stări reprezintă suma rezultatelor obţinute în cazul în care agentul porneşte din respectiva stare şi execută o politică optimală. Considerând π politica ce caracterizează procesul de decizii, valoarea optimală a stării s se scrie:

$V^*(s) = \max_{\pi} E(\sum_{t=0}^{\infty} y^t r_t)$, unde $E(\sum_{t=0}^{h} y^t r_t)$ reprezintă valoarea aşteptată pentru următorii h paşi, în cazul modelului continuu $h \to \infty$.

Funcţia care determină valoarea optimală a stării s este unică şi se determină prin rezolvarea unui sistem de ecuaţii care specifică faptul că valoarea stării este dată de rezultatul imediat la care se adaugă valoarea stării următoare la care s-a ajuns prin selectarea celei mai bune acţiuni.

Găsirea unei politici optimale presupune determinarea valorilor optimale pentru stări.

Următorul algoritm iterativ converge către valorile asociate stărilor, V^*, care compun politica optimală, conform [3]:

Inițializarea aleatoare a valorilor stărilor, $V(s)$;
Repetă până la determinarea politicii optimale:
  Repetă pentru $s \in S$
  Repetă pentru $a \in A$

$$Q(s,a) = R(s,a) + y\sum_{s' \in S} T(s,a,s')V(s')$$

$$V(s) = \max_a Q(s,a)$$

Dificultatea în aplicarea acestui algoritm constă în alegerea momentului la care să fie oprită iterația. A fost propusă următoarea soluție: dacă diferența maximă între două valori succesive ale rezultatelor alegerii stărilor este mai mică decât o valoare predefinită e, atunci valoarea politicii curente diferă de valoarea politicii optimale cu mai puțin de $\frac{2ey}{1-y}$, indiferent de starea pentru care se face comparația între cele două politici, unde y reprezintă utilitatea alegerii unei stări conform modelului Markov. Algoritmul se oprește atunci când se ajunge la o politică care verifică criteriul specificat.

Politica optimală se poate determina printr-un algoritm care lucrează cu politica în mod direct, spre deosebire de cazul precedent în care politica optimală era determinată pe baza unei funcții de determinare a valorii optimale pentru fiecare stare.

Algoritmul pentru determinarea politicii optimale π are următoarea formă:

Alegerea unei politici arbitrare, π'.
Repetă:
$\pi = \pi'$
  Calculează valoarea politicii π:
  Rezolvă sistemul de ecuații liniare:

$$V_\pi(s) = R(s,\pi(s)) + y\sum_{s' \in S} T(s,\pi(s),s')V_\pi(s')$$

  Optimizează politica pentru fiecare stare:

$$\pi'(s) = \max_a (R(s,a) + y\sum_{s' \in S} T(s,a,s')V_\pi(s'))$$

Până când $\pi = \pi'$.

Valoarea politicii reprezintă utilitatea obținută, pentru ansamblul de stări, prin executarea respectivei politici. În momentul în care se cunoaşte valoarea fiecărei stări din politica curentă, se poate aplica un proces de optimizare prin schimbarea primei acțiuni de efectuat. Dacă se constată o îmbunătățire, politica se va orienta spre aceeaşi acțiune de fiecare dată când agentul reîntâlneşte starea. Dacă optimizarea nu conduce la rezultate superioare, atunci politica curentă este cea optimală. Numărul maxim de politici dintre care se alege cea optimală este $|A|^{|S|}$, unde $|X|$ reprezintă cardinalul mulțimii X.

Interacțiunile dintre un agent şi mediul în care îşi desfăşoară activitatea pot fi descompuse secvențial sau pot fi de tip continuu. Primul caz este cel mai uşor de modelat deoarece fiecare acțiune influențează doar un număr finit de rezultate anterior exprimate iar utilitatea totală poate fi calculată ca sumă a utilităților obținute pe parcurs.

Pentru metodele de instruire prin întărire bazate pe procesele de decizie de tip Markov este necesar ca agentul să dispună de o percepție perfectă şi completă asupra stării mediului.

Cele mai multe probleme sunt, însă, de tip continuu sau conțin spații foarte largi de stări discrete. Problema instruirii în spații largi se referă la tehnicile de generalizare care permit memorarea cunoştințelor acumulate şi transferul acestora către acțiuni şi stări similare. Starea unui agent joacă un rol central în selectarea acțiunilor cu utilitate maximă. Putem considera că starea curentă a unui agent reprezintă inputul său. Depinzând de arhitectura agentului, outputul poate fi considerat ca următoarea acțiune selectată sau ca o evaluare a stării curente pentru selectarea acțiunii. Nevoia de generalizare este pronunțată mai ales în spațiul acțiunilor de tip continuu deoarece numărul acestora este foarte mare şi nu se pot menține statistici separate.

Extinderea modelării instruirii prin întărire

Există o varietate de metode şi tehnici de instruire prin întărire care pot rezolva eficient

probleme specifice de dimensiune redusă. Problema apare în cazul extinderii şi generalizării acestor metode pentru aplicarea pe o scară mai largă deoarece, în acest caz, este dificilă soluționarea problemelor de tip arbitrar. Dintre adaptările propuse pentru îmbunătățirea metodelor merită atenția următoarele, conform [4]:

- Tehnica modelării care presupune prezentarea către agent a unor probleme inițiale uşor de rezolvat, după care se face trecerea graduală către situații complexe care vor fi soluționate treptat, pasul către următoarea etapă făcându-se abia după ce se constată că agentul are abilități suficiente. Această metodă dă informații agentului despre valoarea uneia sau alteia dintre stări dar nu poate să-l îndrume spre o anumită acțiune;
- Tehnica semnalelor locale de instruire prin întărire. Agentul primeşte informații locale despre eficacitatea selectării unei acțiuni şi nu despre utilitatea în ansamblu;
- Tehnica imitării prin care agentul urmăreşte modul în care un alt agent face alegerea unei acțiuni (al doilea agent poate fi chiar factorul uman care transmite informații prin intermediul unei interfețe);
- Descompunerea problemei de rezolvat. Prin detalierea problemei generale într-o colecție de subprobleme de dimensiune redusă şi transmiterea de semnale de tip feed-back agentului se obține o metodă eficientă de instruire prin întărire;
- Tehnica reflexelor care presupune o ameliorare a abilităților agentului prin programarea unor "reflexe" care îl vor determina să reacționeze într-un mod rezonabil mai degrabă decât selectarea inițială a acțiunilor de urmat.

Prin introducerea unor astfel de metode, instruirea prin întărire poate conduce la rezolvarea problemelor complexe. Este vorba, de fapt, de introducerea unor elemente ajutătoare prin programare sau prin instruirea de către un alt agent.

Concluzii

Metoda instruirii prin întărire a devenit populară deoarece serveşte ca instrument teoretic pentru studierea principiilor prin care agenții învață să se comporte într-un anumit mediu. Reprezintă, de asemenea, un mijloc de elaborare şi construcție a sistemelor autonome care se autoinstruiesc prin experiență. Domeniile de aplicabilitate variază de la robotică, jocuri pe computer la procesele industriale.

Instruirea prin întărire diferă semnificativ de clasa metodelor care utilizează instruirea supervizată deoarece nu face apel la exemple de tipul input-output. Un agent va cunoaşte utilitatea imediată a unei acțiuni dar nu va şti dacă respectiva acțiune îl conduce la rezultate finale optime. De aceea, pe tot parcursul evoluției sale, agentului îi este absolut necesară experiență acumulată prin cunoaşterea stărilor şi a tranzițiilor dintre acestea. Evaluarea performanțelor sistemului se face în paralel cu procesul de instruire.

Bibliografie

1. Humphrys, Mark – *Action Selection Methods using Reinforcement Learning*, Trinity Hall, Cambridge, 1997, varianta electronică disponibilă la: http://www.compapp.dcu.ie

2. Kulikova, Yevgeniya – *Reinforcement Learning*, varianta electronică disponibilă la: http://www.cs.helsinki.fi/reinforce.html

3. Kaelbling, Leslie Pack, Littman, Michael L. - *Reinforcement Learning: A Survey* , varianta electronică disponibilă la: http://www-2.cs.cmu.edu/afs/cs/project/jair/pub/volume4/kaelbling96a-html/

4. Sutton, Richard Andrew – *Reinforcement Learning, An Introduction*, MIT Press, Cambridge, 1998, varianta electronică disponibilă la: http://www.cs.umass.edu/~rich/book/9/node10.html