سومین کنفرانس بین المللی سالانه کامپیوتر انجمن کامپیوتر ایران
۲ الی ۴ دیماه ۱۳۷۶
دانشکده مهندسی کامپیوتر - دانشگاه علم و صنعت ایران

تنظیم پارامترهای مدل یادگیری Q با استفاده از بازی اتوماتونها[1]

سیامک حجت — کارشناس ارشد
محمدرضا میبدی — دانشیار دانشکدهٔ مهندسی کامپیوتر
دانشکدهٔ مهندسی کامپیوتر، دانشگاه صنعتی امیرکبیر

چکیده

در این مقاله بجای استفاده از روشهای سعی و خطا در یافتن مقادیر مناسب پارامترهای مدل یادگیری Q، از اتوماتونهای یادگیر جهت تعیین و تنظیم این پارامترها استفاده شده است. برای این منظور برای هر کدام از پارامترها یک اتوماتون یادگیر جداگانه در نظر گرفته شده است. این اتوماتونها در کنار هم یک بازی اتوماتون را تشکیل می‌دهند. هدف از این بازی تغییر مقادیر پارامترهای مدل Q جهت عملکرد بهینهٔ این مدل در محیطهای ناشناخته است.

واژه‌های کلیدی:

Reinforcement Learning, Q_Learning, Statistical Clustering, Learning- Automata, Game of Automata

۱. مقدمه

مدل Q[1] که یک نوع مدل یادگیری تقویتی[2] است[4] یک شمای شناخته شده در یادگیری ماشین[3][2] و زندگی مصنوعی[4][3] می‌باشد. در یادگیری تقویتی روشهای انتخاب بهینهٔ فعالیتها توسط یک مأمور یادگیرنده مطالعه می‌شود. انتخاب فعالیتها بر اساس مقادیر کنونی و گذشتهٔ حواس مأمور است و باید به گونه‌ای باشد تا مقدار پسخورهای[5] دریافتی مأمور در طول زمان ماکزیمم شود. مدل Q یک استراتژی انتشار زمانی مقادیر پسخورهای بلافاصله بر روی فعالیتها را در اختیار می‌گذارد. این مدل معمولاً با روشهای دسته بندی آماری[6][5] همراه می‌شود.

---

[1] Game of Automata

[2] Reinforcement Learning

[3] Machine Learning

[4] Artificial Life

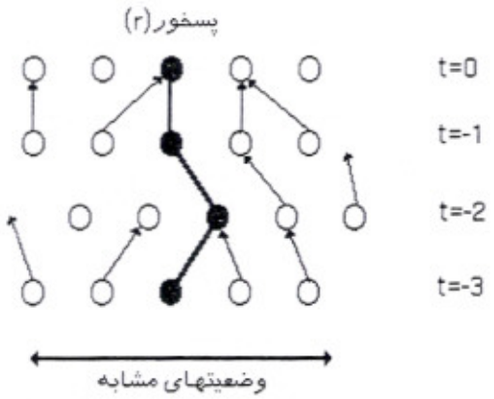
[5] Feedback

[6] Statistical Clustering

"مدل Q با دسته‌بندی آماری" پارامترهای متعددی دارد و عملکرد بهینهٔ مأمور به انتخاب مناسب این پارامترها وابسته است. مقادیر این پارامترها معمولاً با سعی و خطا و توسط طراح مدل انتخاب می‌شوند. انتخاب پارامترها به این شکل برخلاف ماهیت *یادگیری بدون معلم*[7] این روش است و بعلاوه دقت و انعطاف لازم را ندارد. در [6] تنظیم یکی از پارامترهای مدل Q بنام پارامتر انتخاب تصادفی توسط نمونهٔ دیگری از مدلهای یادگیری تقویتی بنام اتوماتون یادگیر[8][8,7] بررسی شده است و نشان داده شده که تنظیم این پارامتر می‌تواند موجب عملکرد بهتر، استقلال از محیط و افزایش انعطاف و اکتشاف شود. در این مقاله تنظیم همزمان بیش از یک پارامتر مدل Q بررسی خواهد شد. از آنجایی که استفاده از یک اتوماتون یادگیر جهت تنظیم چند پارامتر موجب کاهش سرعت پاسخ اتوماتون می‌شود[7]، در اینجا هر پارامتر توسط یک اتوماتون یادگیر جداگانه تنظیم خواهد گردید. استفاده از دو یا بیشتر اتوماتون یادگیر را در یک سیستم، بازی اتوماتون می‌نامند. در این مقاله ابتدا یادگیری تقویتی، مدل یادگیری Q با دسته‌بندی آماری، اتوماتونهای یادگیر و بازی اتوماتونها معرفی خواهند شد و سپس نتایج حاصل از استفاده از بازی اتوماتونها در تنظیم پارامترهای مدل Q با دسته‌بندی آماری ارئه و بررسی می‌شوند.

۲. یادگیری تقویتی

یادگیری تقویتی به بررسی مسئلهٔ یادگیری یک "سیاست انتخاب فعالیتها" بر اساس وضعیتهای تجربه شده و با بکارگیری تکنیک سعی و خطا می‌پردازد، به گونه‌ای که این سیاست موجب ماکزیمم سازی یک معیار اندازه‌گیری عملکرد (پسخور) شود[3,9,10,11]. این روش مشابه مدلهای استفاده شده در برخی مطالعات روانشناسی بر روی رفتارهای یادگیری حیوانات و انسانهاست. این نوع یادگیری در چند مورد با روشهای قبلی متفاوت است[10]. اولاً این یک روش یادگیری بدون معلم است، یعنی مثالها با دقت توسط یک معلم انتخاب نشده‌اند، بلکه توزیع مثالها کاملاً بستگی به فعالیتهای مأمور یادگیرنده دارد. ثانیاً تنها هدف مأمور انتخاب فعالیتهایی است که موجب ماکزیمم سازی پسخورهای دریافتی در طول زمان باشند، در صورتیکه در سایر مدلهای یادگیری هدفهای دیگری مانند مینیمم سازی تعریف یک مفهوم یاد گرفته شده نیز مهم است. ثالثاً در این نوع یادگیری، مأمور با مشکل انتشار پسخورهای دریافتی بر روی فعالیتهای مؤثر در دریافت آنها مواجه می‌باشد.



شکل ۱. انتشار زمانی و ساختاری پسخور

[7]Unsupervised Learning

[8]Learning Automata

شکل (۱) به دو مشکل در انتشار پسخورهای دریافتی اشاره می‌کند: انتشار زمانی و ساختاری. این شکل چند توالی مختلف اعمال فعالیتها برای رسیدن به وضعیت فعلی ($t=0$) را نشان می‌دهد. آنچه واقعاً اتفاق افتاده با پیکانهای پررنگتر مشخص شده است. مشکل انتشار زمانی پسخورها این است که پسخور دریافتی چگونه در زمان به عقب منتشر شود (از $t=0$ به $t=-3$). مشکل انتشار ساختاری این است که پسخور دریافتی چگونه در فضا پخش شود تا وضعیتهای مشابه موجب اعمال فعالیتهای مشابه توسط مأمور شود.

بنا بر [10] ما می‌توانیم بیشتر مدلهای یادگیری تقویتی را نمونه‌هایی از یک شمای کلی بدانیم. هر مأمور در این شما شامل این اجزاء است: یک وضعیت داخلی S، یک تابع بهنگام‌سازی U که موجب تغییر S می‌شود و یک تابع ارزیابی V برای انتخاب فعالیتها. شمای کلی به این شکل عمل می‌کند:

۱. یک مقدار اولیه (S_0) برای وضعیت درونی مأمور (S) در نظر بگیرید.

۲. برای همیشه:

الف) وضعیت فعلی محیط را مشاهده کنید (I).

ب) یک فعالیت $a=V(I,S)$ را با استفاده از تابع ارزیابی V انتخاب کنید.

پ) فعالیت a را در محیط اِعمال کنید. فرض کنید پسخور بلافاصلهٔ اِعمال a در I r باشد.

ت) وضعیت درونی مأمور را توسط تابع U بهنگام کنید: $S_{new} = U(S_{old}, I, a, r)$

۳. مدل Q

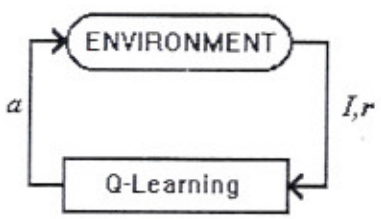
مدل Q یک تکنیک انتشار زمانی پسخورها می‌باشد. در این مدل از یک ساختمان داده بنام Q برای تخمین سودمندی اِعمال فعالیت a در وضعیت حس شدهٔ S استفاده می‌شود: $Q(S,a)$. ابتدا $Q(S,a)$ برای تمام فعالیتهای a و وضعیتهای S برابر با صفر فرض می‌شود. سپس با اِعمال هر فعالیت a در وضعیت X و دریافت پسخور بلافاصلهٔ r مقدار $Q(X,a)$ با فرمول زیر بهنگام میشود:

$$Q(X,a) \leftarrow Q(X,a) + \lambda\,(r + \gamma\ e(Y) - Q(X,a)) \qquad (۱)$$

در فرمول بالا Y وضعیت بعدی محیط (پس از اِعمال فعالیت a در وضعیت X است) و $e(Y)$ سودمندی وضعیت Y می‌باشد که با فرمول زیر محاسبه می‌شود: (m تعداد فعالیتهاست)

$$e(Y) \leftarrow \text{maximum } Q(Y,i) \text{ over all actions i, } (\text{i} = 1,2,\ldots,\text{m}) \qquad (۲)$$

پارامتر λ ($0 \le \lambda \le 1$) میزان اصلاح خطا برای Q را تعیین می‌کند و پارامتر γ ($0 \le \gamma \le 1$) میزان صرفنظر کردن از سودمندی وضعیت نتیجه شده را مشخص می‌کند .



شکل ۱. *رابطهٔ مدل Q با محیط.*

شکل (۱) به دو مشکل در انتشار پسخورهای دریافتی اشاره می‌کند: انتشار زمانی و ساختاری. این شکل چند توالی مختلف اعمال فعالیتها برای رسیدن به وضعیت فعلی ($t=0$) را نشان می‌دهد. آنچه واقعاً اتفاق افتاده با پیکانهای پررنگتر مشخص شده است. مشکل انتشار زمانی پسخورها این است که پسخور دریافتی چگونه در زمان به عقب منتشر شود (از $t=0$ به $t=-3$). مشکل انتشار ساختاری این است که پسخور دریافتی چگونه در فضا پخش شود تا وضعیتهای مشابه موجب اعمال فعالیتهای مشابه توسط مأمور شود.

بنا بر [10] ما می‌توانیم بیشتر مدلهای یادگیری تقویتی را نمونه‌هایی از یک شمای کلی بدانیم. هر مأمور در این شما شامل این اجزاء است: یک وضعیت داخلی S یک تابع بهنگام‌سازی U که موجب تغییر S می‌شود و یک تابع ارزیابی V برای انتخاب فعالیتها. شمای کلی به این شکل عمل می‌کند:

۱. یک مقدار اولیه (S_0) برای وضعیت درونی مأمور (S) در نظر بگیرید.

۲. برای همیشه:

الف) وضعیت فعلی محیط را مشاهده کنید (I).

ب) یک فعالیت $a=V(I,S)$ را با استفاده از تابع ارزیابی V انتخاب کنید.

پ) فعالیت a را در محیط اِعمال کنید. فرض کنید پسخور بلافاصلهٔ اِعمال a در I r باشد.

ت) وضعیت درونی مأمور را توسط تابع U بهنگام کنید: $S_{new} = U(S_{old}, I, a, r)$

۳. مدل Q

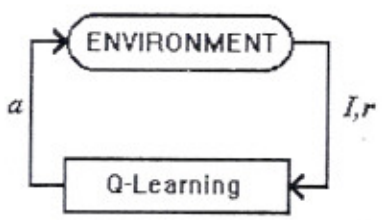
مدل Q یک تکنیک انتشار زمانی پسخورها می‌باشد. در این مدل از یک ساختمان داده بنام Q برای تخمین سودمندی اِعمال فعالیت a در وضعیت حس شدهٔ S استفاده می‌شود: $Q(S,a)$. ابتدا $Q(S,a)$ برای تمام فعالیتهای a و وضعیتهای S برابر با صفر فرض می‌شود. سپس با اِعمال هر فعالیت a در وضعیت X و دریافت پسخور بلافاصلهٔ r مقدار $Q(X,a)$ با فرمول زیر بهنگام میشود:

$$Q(X,a) \leftarrow Q(X,a) + \lambda\,(r + \gamma\ e(Y) - Q(X,a)) \qquad (۱)$$

در فرمول بالا Y وضعیت بعدی محیط (پس از اِعمال فعالیت a در وضعیت X است) و $e(Y)$ سودمندی وضعیت Y می‌باشد که با فرمول زیر محاسبه می‌شود: (m تعداد فعالیتهاست)

$$e(Y) \leftarrow \text{maximum } Q(Y,i) \text{ over all actions i, } (\text{i} = 1,2,\ldots,\text{m}) \qquad (۲)$$

پارامتر λ ($0 \le \lambda \le 1$) میزان اصلاح خطا برای Q را تعیین می‌کند و پارامتر γ ($0 \le \gamma \le 1$) میزان صرفنظر کردن از سودمندی وضعیت نتیجه شده را مشخص می‌کند .



شکل ۱. رابطهٔ مدل Q با محیط.

در مدل Q تابع ارزیابی برای انتخاب بهترین فعالیت در وضعیت S باید فعالیتی را انتخاب کند که مقدار $Q(S,a)$ را ماکزیمم کند. این سیاستِ انتخاب فعالیتها تمام فعالیتهای ممکن را کنکاش نمی‌کند و درنتیجه در بسیاری از مواقع منجر به انتخاب غیر بهینهٔ فعالیتها می‌شود. بنا بر این لازم است در درصدی از مواقع (θ) انتخاب فعالیت بطور تصادفی انجام گیرد[4].

۳-۱. دسته‌بندی آماری

مدل Q معمولاً با تکنیکی برای انتشار ساختاری پسخورها همراه می‌شود. یک تکنیک برای این منظور نگهداری کلّ وضعیتهای مشاهده شده توسط مأمور یادگیرنده (تجربیات) و استفاده از فاصلهٔ همینگ[9] برای تشخیص وضعیتهای مشابه و انتشار پسخورهای دریافت شده بر روی آنهاست. در مواقعی که تعداد وضعیتهای قابل تجربه در محیط زیاد باشد بهتر است بجای این روش از روشهای دسته‌بندی آماری استفاده شود. در دسته‌بندی آماری تمام تجربیات مشابه در یک دسته قرار می‌گیرند و بجای ذخیره کردن همهٔ آنها تنها اطلاعاتی آماری بعنوان نمایندهٔ آنها نگهداری می‌شود. در این تکنیک هر تجربهٔ جدید با دسته‌های موجود مقایسه شده و در دسته (یا دسته‌های) مشابه ادغام می‌شود. در صورتیکه تجربهٔ جدید مشابه هیچکدام از دسته‌های موجود نباشد یک دستهٔ جدید برای آن تجربه ایجاد خواهد شد.

۳-۱-۱. دسته ها:

هر دسته نمایانگر گروهی از وضعیتهای *مشابه* است. یک دسته را میتوان با $n+2$ تایی زیر نشان داد:

$$C = < (z_1, o_1), (z_2, o_2), \ldots, (z_n, o_n), Q_c, M_c > \qquad (۳)$$

z_i و o_i تعداد دفعاتی است که بیت iام از وضعیت S در دستهٔ C، 0 یا 1 بوده است. n تعداد بیتهای یک وضعیت است، Q_c مقدار Q دسته را مشخص می‌کند و M_c نمایانگر تعداد تجربیاتی است که در این دسته قرار گرفته‌اند. با این نمایش میتوان احتمال شرطی یک بودن بیت iام از وضعیت S در این دسته را با فرمول زیر محاسبه کرد: (s_i بیت iام S است)

$$p(s_i = 1 | S \in C) = \frac{o_i}{o_i + z_i} \qquad (٤)$$

۳-۱-۲. مقایسهٔ وضعیتها با دسته ها:

برای مقایسهٔ وضعیت S با دستهٔ C می‌توان از احتمال شرطی قرار گرفتن S در دستهٔ C استفاده کرد: ($v_i = 0$ یا 1)

$$p(S \in C | s_1 = v_1, s_2 = v_2, \ldots, s_n = v_n) = \frac{p(s_1 = v_1, \ldots, s_n = v_n | S \in C) p(S \in C)}{p(s_1 = v_1, \ldots, s_n = v_n)} \qquad (۵)$$

با فرض مستقل بودن بیتهای یک وضعیت (که فرضی نادرست ولی با تقریب خوبی قابل قبول است) می‌توان مخرج کسر بالا را بصورت زیر نوشت:

---

[9] Hamming distance

$$p(s_1 = v_1, \ldots, s_n = v_n) = \prod_{i=1}^{n} p(s_i = v_i) \qquad (۶)$$

برای صورت کسر نیز میتوان از فرمول زیر استفاده کرد:

$$p(s_1 = v_1, \ldots, s_n = v_n | S \in C) = \prod_{i=1}^{n} p(s_i = v_i | S \in C) \qquad (۷)$$

سمت راست معادلهٔ (۶) را می‌توان با نگهداری اطلاعاتی آماری از حواس محاسبه کرد. مقدار $p(S \in C)$ و سمت راست معادلهٔ (۷) را نیز می‌توان با استفاده از اطلاعات دسته‌ها بدست آورد. در نتیجه سمت چپ معادلهٔ (۵) قابل محاسبه خواهد بود. حال برای اینکه وضعیت S در دستهٔ C قرار گیرد باید داشته باشیم:

$$p(S \in C | s_1 = v_1, s_2 = v_2, \ldots, s_n = v_n) > \varepsilon \qquad (۸)$$

$$|Q_c - Q_s| < \delta \qquad (۹)$$

نامعادلات (۸) و (۹) تضمین می‌کنند که اولاً مشابهت وضعیت S با دستهٔ C از یک مقدار آستانه‌ای (ε) بیشتر باشد و ثانیاً Q محاسبه شده برای وضعیت S نسبت به مقدار Q ذخیره شده در دستهٔ C از یک مقدار ثابت آستانه‌ای (δ) کمتر باشد.

۳-۱-۳. ادغام وضعیتها با دسته ها:

بعد از اینکه مشخص شد که وضعیت S در دستهٔ C قرار می‌گیرد از آن برای بهنگام سازی دسته استفاده خواهد شد. فرض کنید:

$$C = < (z_1, o_1), (z_2, o_2), \ldots, (z_n, o_n), Q_c, M_c > \qquad (۱۰)$$

اگر C_u نمایانگر دستهٔ C پس از بهنگام‌سازی باشد و داشته باشیم:

$$C_u = < (z_{1u}, o_{1u}), (z_{2u}, o_{2u}), \ldots, (z_{nu}, o_{nu}), Q_{cu}, M_{cu} > \qquad (۱۱)$$

برای هر بیت i از وضعیت S که برابر با 1 باشد خواهیم داشت:

$$z_{iu} = \mu z_i \ , \ o_i = 1 + \mu o_i \quad (s_i = 1) \qquad (۱۲)$$

و برای هر بیت i از وضعیت S که برابر با 0 باشد خواهیم داشت:

$$z_{iu} = 1 + \mu z_i \ , \ o_i = \mu o_i \quad (s_i = 0) \qquad (۱۳)$$

در اینجا μ عددی حقیقی و بین صفر و یک است که برای افزایش اهمیت تجارب جدید بکار می‌رود. اگر μ=1 باشد اهمیت تجارب جدید در نظر گرفته نخواهد شد. μ را معمولاً از فرمول $\mu = (2K-1)/2K$ بدست می‌آورند که در آن K عددی صحیح است (از K برای ایجاد دسته های جدید استفاده خواهد شد).

فرض کنید در وضعیت S فعالیت a إعمال شده باشد و مقدار محاسبه شدهٔ Q برای آن برابر با $Q(S,a)$ باشد آنگاه برای ساختن Q_{cu} می‌توان از مجموع Q_c و $Q(S,a)$ استفاده کرد. معمولاً در این جمع از M_c بعنوان وزن استفاده می‌شود:

$$Q_{cu} = Q_u(\frac{M_c}{M_c+1}) + Q(S,a)(\frac{1}{M_c+1}) \quad (۱۴)$$

همچنین تعداد تجربیات دستهٔ C_u بصورت زیر بهنگام می‌شود:

$$M_{cu} = M_c + 1 \quad (۱۵)$$

۳-۱-۴. ایجاد دسته‌های جدید:

اگر یک وضعیت S مشابه هیچ یک از دسته‌های موجود نباشد باید یک دستهٔ جدید C_{new} برای آن ساخته شود. برای اینکار ابتدا یک دستهٔ خالی به شکل زیر ایجاد می‌گردد:

$$C =< (z_1,o_1),(z_2,o_2),\ldots,(z_n,o_n),Q_c,M_c > \quad (۱۶)$$

که در آن:

$$z_i = o_i = K \ , \ Q_c = 0 \ , \ M_c = 0 \quad (۱۷)$$

سپس دستهٔ خالی فوق با وضعیت S ادغام می‌شود و دستهٔ C_{new} را می‌سازد.

۳-۱-۵. ادغام دسته‌های موجود:

گاهی دو دسته به‌اندازه‌ای مشابه یکدیگرند که می‌توانند در هم ادغام شوند. برای محاسبهٔ تشابه دو دسته می‌توان از اندازه‌گیری فاصلهٔ بین دو دسته استفاده کرد:

$$distance(C_1,C_2) = \sum[p(s_i = 1|S \in C_1) - p(s_i = 1|S \in C_2)] \quad (۱۸)$$

دو دستهٔ C_1 و C_2 تنها زمانی با هم ادغام می‌شوند که اولاً فاصلهٔ آنها کمتر از مقدار ثابت ρ باشد و ثانیا مقادیر Q دو دسته اختلافی کمتر از δ داشته باشد. یعنی:

$$distance(C_1,C_2) < \rho \quad (۱۹)$$

$$|Q_{c_1} - Q_{c_2}| < \delta \quad (۲۰)$$

حال اگر دو دستهٔ زیر را داشته باشیم:

$$C_a =< (z_{1a},o_{1a}),(z_{2a},o_{2a}),\ldots,(z_{na},o_{na}),Q_{ca},M_{ca} > \quad (۲۱)$$

$$C_b =< (z_{1b},o_{1b}),(z_{2b},o_{2b}),\ldots,(z_{nb},o_{nb}),Q_{cb},M_{cb} > \quad (۲۲)$$

و این دو دسته به اندازهٔ کافی مشابه باشند (یعنی روابط ۱۹ و ۲۰ برای آنها صادق باشد) آنگاه دو دسته در هم ادغام می‌شوند و دستهٔ جدید C را بوجود می‌آورند:

$$C =< (z_1,o_1),(z_2,o_2),\ldots,(z_n,o_n),Q_c,M_c > \quad (۲۳)$$

عناصر دستهٔ C بصورت زیر ساخته می‌شوند:

$$z_{ic} = z_{ia}\left(\frac{M_a}{M_a + M_b}\right) + z_{ib}\left(\frac{M_b}{M_a + M_b}\right) \quad (۲۴)$$

$$o_{ic} = o_{ia}\left(\frac{M_a}{M_a + M_b}\right) + o_{ib}\left(\frac{M_b}{M_a + M_b}\right) \quad (۲۵)$$

$$Q_c = Q_a\left(\frac{M_a}{M_a + M_b}\right) + Q_b\left(\frac{M_b}{M_a + M_b}\right) \quad (۲۶)$$

$$M_c = M_a + M_b \quad (۲۷)$$

۳-۱-۶. انتخاب فعالیت با استفاده از دسته‌های موجود:

در یادگیری Q باید در هر وضعیت S فعالیت a را به گونه‌ای انتخاب کرد که مقدار $Q(S,x)$ را به ازای تمام فعالیتهای ممکن ماکزیمم کند ($x = 1,2,\ldots,m$). برای محاسبهٔ $Q(S,x)$ از روی دسته‌های موجود میتوان از فرمول زیر استفاده کرد:

$$Q(S,x) = \frac{\sum_{C \in C_x} [Q_c \times P(S \in C | S_1 = V_1, \ldots, S_n = V_n)}{\sum_{C \in C_x} [P(S \in C | S_1 = V_1, \ldots, S_n = V_n)} \quad (۲۸)$$

در عبارت فوق C_x مجموعهٔ دسته‌هایی می‌باشد که در آنها فعالیت x انتخاب شده است. صورت کسر بالا مجموع وزن‌دار مقادیر Q برای عناصر Cx است (از احتمال قرار گرفتن وضعیت S در دستهٔ C بعنوان وزن این جمع استفاده شده است). مخرج کسر نیز برای نرمال کردن عبارت می‌باشد.

۳-۲. جمعبندی روش یادگیری مدل Q

مدل Q با دسته بندی را می‌توان به صورت زیر خلاصه کرد:

۱. مقادیر ثابتی برای پارامترهای Q (θ,γ,λ) و پارامترهای دسته بندی $(K,\rho,\varepsilon,\delta)$ در نظر بگیرید.

۲. برای همیشه:

الف) وضعیت فعلی محیط را مشاهده کنید (X).

ب) در θ درصد از مواقع فعالیتی را به طور تصادفی انتخاب کنید. در مواقع دیگر فعالیتی را انتخاب کنید که مقدار $Q(X,a)$ را ماکزیمم کند.

پ) فعالیت a را در محیط اِعمال کنید. فرض کنید وضعیت جدید Y باشد و پسخور بلافاصلهٔ اِعمال این فعالیت r باشد.

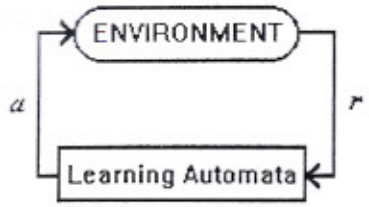
ت) میزان $Q(X,a)$ را با معادلهٔ (۱) بهنگام کنید.

ث) اگر دسته‌ای مانند C وجود داشت که به همراه X در نامعادلات (۸) و (۹) صدق کند، وضعیت X را در دستهٔ C ادغام کنید. در غیر اینصورت دستهٔ جدیدی از روی X ایجاد نمایید.

ج) هر دو دستهٔ C_1 و C_2 که در نامعادلات (۱۹) و (۲۰) صدق می‌کنند را در هم ادغام کنید.

۴. اتوماتونهای یادگیری (LA)

در اتوماتونهای یادگیری یک یادگیرنده استرتژی انتخاب فعالیت خود را تنها بر اساس پسخورهای دریافت شده از محیط ایجاد می‌کند (شکل ۳). در اتوماتونهای یادگیری به هر فعالیت a یک *احتمال انتخاب* $p(a)$ نسبت داده می‌شود. هرگاه اتوماتون فعالیت a را در زمان t انتخاب کند و پسخور محیط موفقیت آمیز باشد آنگاه $p(a)$ افزایش داده می‌شود و احتمال انتخاب سایر فعالیتها کاهش پیدا می‌کند (در صورتیکه پسخور محیط نشانگر عدم موفقیت باشد عکس این عمل اتفاق خواهد افتاد).



شکل ۳. رابطهٔ اتوماتون یادگیری با محیط.

در حالت کلی اتوماتون یک فعالیت را بر اساس احتمال انتخاب فعالیتها انتخاب می‌کند (بعنوان مثال فعالیت a) و پس از دریافت پسخور حاصل از إعمال این فعالیت مقادیر $p(i)$ ها را بر اساس فرمول زیر بهنگام می‌سازد:

$$p(i) \leftarrow p(i) + \frac{(1-r)\times\beta\times p(a)}{m-1} - \frac{r\times a\times(1-p(a))}{m-1} \quad , \text{for all } i \neq a \qquad (۳۰)$$

$$p(a) \leftarrow p(a) - (1-r)\times\beta\times p(a) + r\times a\times(1-p(a)) \qquad (۳۱)$$

در فرمول بالا m تعداد فعالیتها و α و β به ترتیب پارامترهای پاداش و جزا می‌باشند که مقادیری بین صفر و یک اختیار می‌کنند. *پارامتر پاداش* نرخ افزایش احتمال انتخاب فعالیتی را نشان می‌دهد که پسخور موفقیت آمیز دریافت کرده و *پارامتر جزا* نرخ کاهش احتمال انتخاب فعالیتی را مشخص می‌کند که پسخور غیر موفق دریافت نموده است. در رابطهٔ بالا مقدار r باید بین ۰ و ۱ باشد، در غیر اینصورت می‌توان از رابطهٔ زیر مقدار r را به عددی بین ۰ و ۱ نگاشت کرد (در این رابطه فرض شده است که مقدار ماکزیمم و مینیمم پسخورهای دریافت شده، از پیش مشخص نمی‌باشند):

$$r_{norm} = \frac{r - Min(r_i)}{Max(r_i) - Min(r_i)} \qquad (۳۲)$$

اتوماتونهای یادگیر بر حسب مقدار پارامتر جزا به سه اتوماتون اصلی تقسیم‌بندی می‌شوند. اگر $\alpha \neq 0$ و $\beta = 0$ اتوماتون را LRI[۱۰] و در صورتیکه $\alpha = \beta \neq 0$ باشد اتوماتون را LRP[۱۱] می‌نامیم. اگر $\alpha \neq 0$ و β به اندازهٔ کافی کوچک باشد اتوماتون را $LR\varepsilon P$[۱۲] می‌نامیم.

---

[۱۰] Linear-Reward-Inaction

[۱۱] Linear-Reward-Penalty

[۱۲] Linear-Reward-Epsilon Penalty

۵. بازی اتوماتونها

یک بازی اتوماتون شامل دو یا بیشتر اتوماتون می‌باشد و نتیجهٔ این بازی بستگی به رفتار اتوماتونها دارد. میزان ارتباط بین اتوماتونها، میزان اطلاعاتی که در دسترس هر اتوماتون است و فعالیتهایی که هر اتوماتون در اختیار دارد قوانین بازی را تشکیل می‌دهند. پیچیدگی یک بازی بستگی به تعداد اتوماتونها، تعریف تابع ارزیابی هر اتوماتون و نوع ارتباط (همکاری یا رقابت) اتوماتونها دارد.

فرض کنید اتوماتونهای $A_g,\ldots,A_2,A_1$ در یک بازی شرکت داشته باشند و به ترتیب از مجموعهٔ فعالیتهای $\{a_{m_1}^1,\ldots,a_2^1,a_1^1\}$، $\{a_{m_1}^2,\ldots,a_2^2,a_1^2\}$، ... و $\{a_{m_1}^g,\ldots,a_2^g,a_1^g\}$ استفاده کنند. در اینصورت بازی را می‌توان با ماتریس D به ابعاد $m_1\times m_2\times\ldots\times m_g$ نمایش داد که عنصر d_{ij_k} آن احتمال موفقیت انتخاب فعالیتهای $(a_k^g,\ldots,a_j^2,a_i^1)$ را مشخص می‌کند. در ماتریس D عنصر d_{ij_k} یک ماکزیمم محلی خوانده می‌شود اگر ماکزیمم مقادیر تمام ابعادش (i، j ... و k) باشد. ماکزیمم کلی، ماکزیمم تمام ماکزیمم‌های محلی است. مطالعات انجام شده [7]، نشان می‌دهد که در یک محیط استاتیک (وقتی D مستقل از زمان باشد) بازی اتوماتون به سمت یک ماکزیمم محلی همگرا خواهد داشت. برای رسیدن به ماکزیمم کلی، ماتریس پسخور D باید تنها یک نقطهٔ سکون[۱۳] داشته باشد. راه حل دیگر این است که محدودیتی بر نوع اطلاعاتی که در اختیار اتوماتونها قرار می‌گیرد وجود نداشته باشد تا بتوان از الگوریتمهای تخمینی جهت همگرایی به ماکزیمم کلی استفاده کرد.

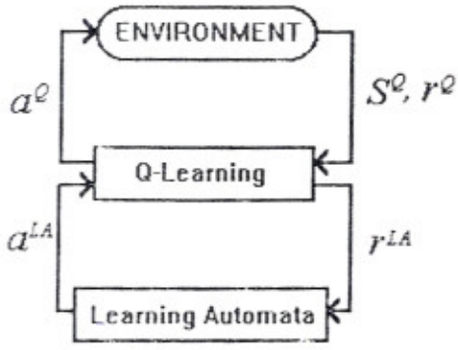
۶. استفاده از بازی اتوماتونهای یادگیر در تنظیم پارامترهای مدل Q (Q_LA)

برای تنظیم پارامترهای مدل Q با استفاده از اتوماتونهای یادگیر این اتوماتونها باید مجهز به فعالیتهایی جهت تغییر مقادیر پارامترهای مدل Q باشند. ایدهٔ کلی این است که اتوماتونها با اِعمال فعالیتهای خود مقدار پارامترهای مدل Q را تغییر دهند. مدل Q در طول یک پریود زمانی(*period*) از این مقادیر استفاده می‌کند. سپس در انتهای پریود میزان کارا بودن هر یک از پارامترها توسط یک تابع ارزیابی تخمین زده می‌شود. این تخمینها بصورت پسخورهایی در اختیار اتوماتونهای یادگیر قرار می‌گیرند. اتوماتونها با استفاده از پسخورهای دریافتی احتمال انتخاب فعالیتهای خود را بهنگام کرده و مجدداً فعالیتهایی را برای تغییر مقادیر پارامترهای مدل Q اِعمال می‌کنند (شکل ۳). به این ترتیب مدل Q نقش محیط اتوماتون یادگیر را ایفا می‌کند.

به عنوان مثال اگر از یک اتوماتون یادگیر برای تنظیم مقدار θ استفاده شود. برای تعیین مقدار مناسب این پارامتر که عددی بین صفر و یک است بازهٔ $[0,1]$ به بازه‌های کوچکتری تقسیم می‌شود. فرض کنید بازه‌ها با فاصلهٔ یک دهم انتخاب شوند (یعنی ده بازه) و هر بازه با کوچکترین عدد آن بازه مشخص شود. وظیفهٔ اتوماتون یادگیر انتخاب بازه‌ای است که عدد مشخص کنندهٔ آن مناسبترین مقدار را برای پارامتر انتخاب تصادفی یادگیری Q داشته باشد. برای این منظور باید از یک اتوماتون یادگیر که مجهز به ده فعالیت (برای انتخاب هر کدام از بازه‌ها) است استفاده کرد. اتوماتون یادگیر پس از انتخاب یک مقدار برای این پارامتر آنرا در اختیار مدل Q می‌گذارد و مدل Q در یک پریود زمانی (*period*) از این مقدار استفاده خواهد کرد. پس از انقضای این پریود میزان کارایی پارامتر تنظیم شده توسط یک تابع ارزیابی به اتوماتون یادگیر پسخور می‌گردد .

---

[۱۳] Equilibrium



شکل ۴. رابطهٔ مدل Q با اتوماتون یادگیری و محیط.

استفاده از اتوماتونهای یادگیر برای تنظیم پارامترهای مدل Q نیاز به تعیین پارامترهای جدیدی را ایجاب می‌کند اما آزمایشهای انجام شده [6] نشان می‌دهد که استفاده از مقادیر متعارف پارامترهای α و β (α=β 0.1) در اتوماتونهای یادگیری کافی است. بعلاوه طول بازه‌ها را میتوان متغییر در نظر گرفت، به این ترتیب که مثلاً در ابتدا می‌توان از بازه‌هایی به طول 0.1 استفاده کرده و پس از همگرا شدن پارامتر به یکی از این بازه ها، بازهٔ مورد نظر را به بازه‌های کوچکتری تقسیم کرد و از سایر بازه‌ها صرفنظر نمود. به این ترتیب مشکل انتخاب طول بازهٔ مناسب تا حدودی مرتفع خواهد شد.

بازی‌هایی که در این مقاله بررسی می‌شوند شامل اتوماتونهایی می‌باشند که جهت تنظیم پارامترهای θ، ε، ρ و σ با یکدیگر همکاری می‌کنند. در این بازیها هر اتوماتون بطور کاملاً مستقل و بدون در نظر گرفتن سایر اتوماتونها فعالیتهای خود را انتخاب می‌کند و هر اتوماتون فقط مسئول تنظیم یک پارامتر است.

در این مقاله اتوماتون تنظیم کنندهٔ θ شامل ۱۰ فعالیت می‌باشد که به ترتیب مقادیر 0، 0.1، 0.2، ... و 0.9 را برای θ انتخاب می‌کنند. تابع ارزیابی این اتوماتون نیز به صورت زیر تعریف شده است: (این روابط پسخور اتوماتون در زمان t را مشخص می‌کنند، فرض شده که اتوماتون در زمان T پسخور ماکزیمم را دریافت کرده باشد)

$$1)\ \text{IF} \sum_{i=t-p}^{t} r_i^Q \geq \sum_{i=T-p}^{T} r_i^Q \ \text{THEN}\ r^{LA} = MAX_{i=1}^{t}(r_i^Q)$$

$$2)\ \text{ELSE IF} \sum_{i=t-p}^{t} \frac{r_i^Q}{p} \leq \sum_{i=1}^{t} \frac{r_i^Q}{t} \ \text{THEN}\ r^{LA} = MIN_{i=1}^{t}(r_i^Q)$$

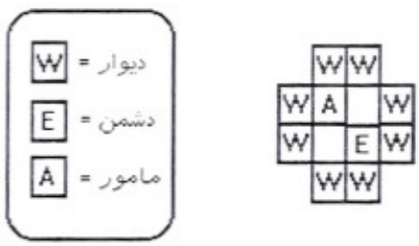
$$3)\ \text{ELSE}\ r^{LA} = MIN_{i=1}^{t}(r_i^Q) + \frac{\sum_{i=t-p}^{t} \frac{r_i^Q}{p} - \sum_{i=1}^{t} \frac{r_i^Q}{t}}{\sum_{i=T-p}^{T} \frac{r_i^Q}{p} - \sum_{i=1}^{t} \frac{r_i^Q}{t}} \times \left(MAX_{i=1}^{t}(r_i^Q) - MIN_{i=1}^{t}(r_i^Q) \right)$$

در تابع ارزیابی اتوماتونهای تنظیم کنندهٔ پارامترهای ε، ρ و σ درصورتیکه تعداد دسته‌ها در یک پریود تغییر نکند، پسخور اتوماتونها با روابط بالا محاسبه خواهد شد. اما اگر تعداد دسته‌ها در یک پریود زمانی افزایش یابد و میانگین پسخورهای دریافتی در این پریود نسبت به میانگین کل بهتر نشود، آنگاه این اتوماتونها پسخور مینیمم را دریافت خواهند کرد و درصورتیکه تعداد دسته ها در یک پریود زمانی کاهش یابد و میانگین پسخورهای دریافتی در این پریود نسبت به پریود قبلی کاهش نیابد آنگاه این اتوماتونها پسخور ماکزیمم را دریافت می‌کنند. در آزمایشهای انجام شده اتوماتون تنظیم کنندهٔ ε دارای ۱۰ فعالیت برای انتخاب یکی از مقادیر 0، 0.00000001،

0.000001، 0.0001، 0.001، 0، 0.2، 0.4، 0.6 و 0.8 برای ε است، اتوماتون تنظیم کنندهٔ ρ یکی از ۵ مقدار 0، 1، 2، 3 و 4 را برای ρ و اتوماتون تنظیم کنندهٔ σ یکی از ۱۰ مقدار 1، 5، 10، 15، ... و 45 را برای σ انتخاب می‌کند. مقادیر اولیهٔ پارامترهای λ و γ در این آزمایشها به ترتیب برابر با 1 و 0 در نظر گرفته شده است که برای آزمایشهای انجام شده مناسبترین مقادیر می‌باشند. همچنین مقدار اولیهٔ پارامتر K برابر با 5 و هرجا پارامترهای θ، ε، ρ و σ تنظیم نشده باشند مقادیر اولیهٔ آنها به ترتیب برابر با 10، 0.000001، 2 و 20 قرار داده شده است (این مقادیر برای آزمایش مشابهی در [4] پیشنهاد شده‌اند). تمام اتوماتونهای مورد استفاده در آزمایشها از نوع L_RP می‌باشد و مقدار *period* برای تمام آنها برابر با 10 واحد زمانی در نظر گرفته شده است.

۷. محیط آزمایشها

بمنظور مقایسهٔ رفتار 'مدل Q با پارامترهای ثابت' (Q) و 'مدل Q با پارامترهای تنظیم شده' (Q_LA) از این دو مدل برای کنترل یک مأمور یادگیرنده در محیطهایی مصنوعی استفاده شده است. در این آزمایشها محیط (شکل ۵)، یک صفحهٔ شطرنجی شکل است که در هر خانهٔ آن ممکن است یک شیء وجود داشته باشد. مأمور یادگیرنده مجهز به چهار حس برای مشاهدهٔ اجسام خانه‌های مجاور خود (بالا، راست، پایین و چپ) و چهار فعالیت برای حرکت به این خانه‌ها است. این فعالیتها تنها زمانی موجب حرکت به یک خانهٔ مجاور می‌شود که در آن خانه شیئی وجود نداشته باشد. در این صورت حرکت به خانهٔ مجاور انجام می‌گردد و مأمور پسخور صفر ($r^0 = 0$) را دریافت می‌کند. در هر خانه از محیط یکی از اشیاء دیوار یا دشمن می‌تواند وجود داشته باشد. اِعمال یک فعالیت برای حرکت به خانه‌ای که در آن این اشیاء وجود دارند به ترتیب پسخورهای 40- و 100- را ایجاد می‌کند. مکان دیوار در محیط ثابت است اما دشمن می‌تواند در هر واحد زمانی به یکی از چهار خانهٔ مجاور (بالا، راست، پایین یا چپ) حرکت کند. حرکت دشمن به خانه‌های مجاور بطور تصادفی انجام می‌گیرد.

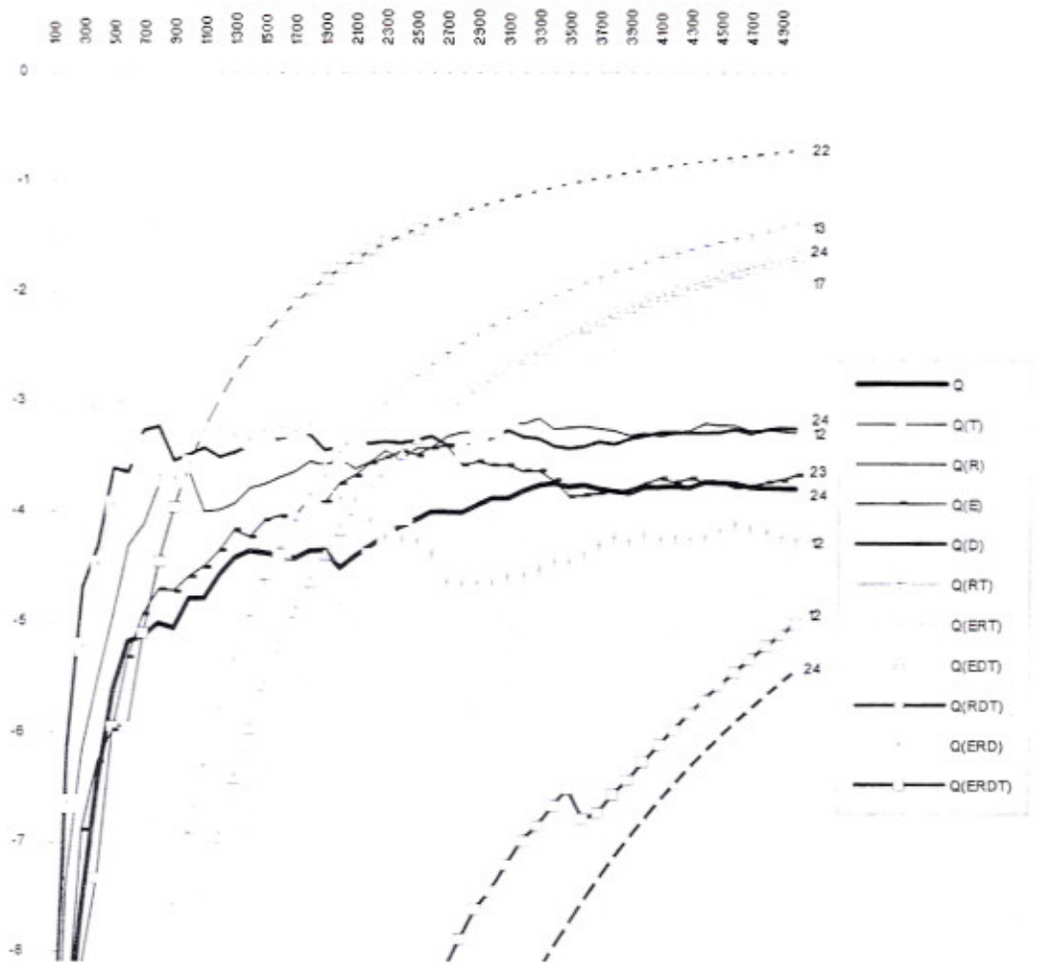


شکل ۵. محیط

۸. نتایج آزمایشها

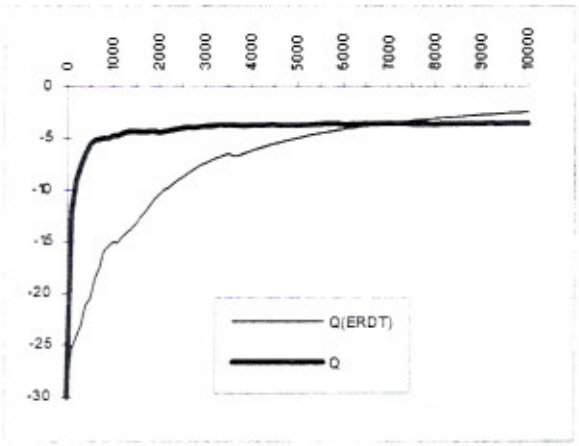
در نمودار ۱ نحوهٔ تغییر پسخورهای کل دریافتی از محیط در طول ۵۰۰۰ واحد زمانی برای ۱۱ مدل مختلف مقایسه شده‌اند. در اینجا برای تفکیک نام مدلها در داخل پرانتز حرف اول نام پارامتر یا پارامترهایی که تنظیم می‌کنند آورده شده است. به این ترتیب حروف E، R، D و T به ترتیب به پارامترهای ε، ρ، δ و θ اشاره می‌کنند. به عنوان مثال در مدل Q(RT) دو پارامتر ρ و θ تنظیم شده‌اند. مقایسهٔ جداگانهٔ هر مدل با مدل Q در نمودارهای ۴ تا ۱۳ آورده شده است. مزایای تنظیم پارامتر θ قبلاً در [8] بررسی شده است. با محاسبه می‌توان نشان داد که یک مدل با پارامتر ثابت $\theta = 10$ حداکثر قادر است به پسخور متوسط 2.83- برسد، اما مدلها با تنظیم θ می‌توانند در تئوری به پسخور متوسط 0 برسند (که بهترین عملکرد ممکن در این محیط است). در نمودار ۱ همگرا شدن پسخور متوسط مدلهای Q(T)، Q(RT) و Q(ERT) به سمت 0 قابل مشاهده است. گرچه افزایش تعداد پارامترهای

باعث شده تا سرعت همگرایی در مدلهای Q(EDT)، Q(RDT) و Q(ERDT) تا حدودی کاهش یابد ولی آزمایشها نشان داده‌اند که این مدلها نیز با صرف زمان بیشتر عملکرد بهتری را نسبت به Q نشان می‌دهند (نمودار ۲ تغییرات پسخور متوسط مدل Q(ERDT) را در ۱۰۰۰۰ واحد زمانی دنبال می‌کند).

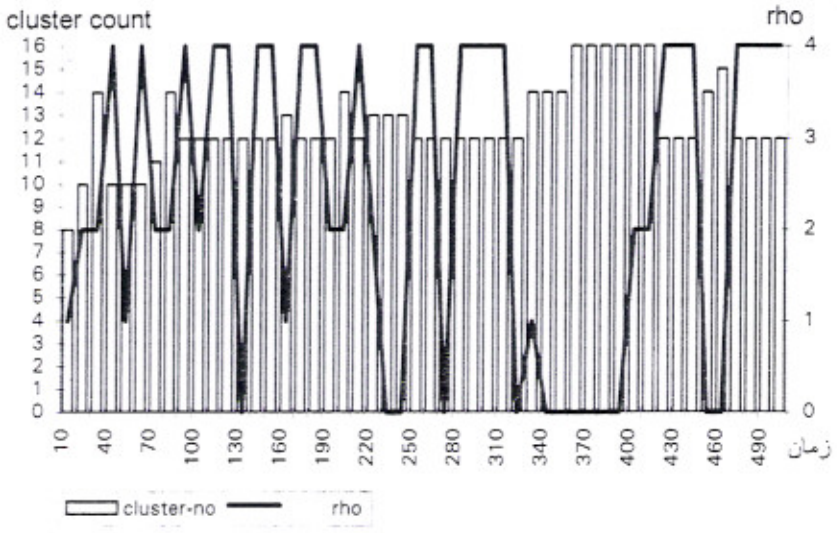


نمودار۱- مقایسهٔ تغییرات پسخور متوسط مدلهای مختلف در زمان.

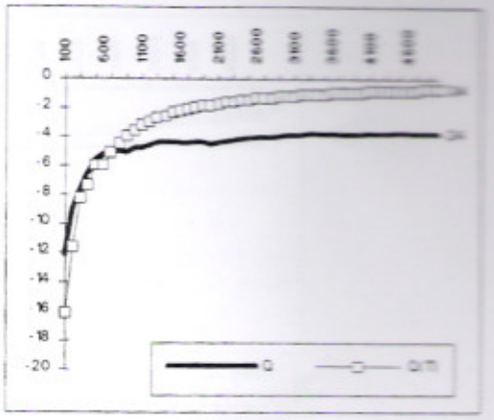
در کنار هر کدام از منحنی‌های نمودار ۱ تعداد نهایی دسته‌هایی که مدل پس از خبرگی به آنها رسیده نوشته شده است. می‌توان نشان داد که حداقل تعداد دسته‌هایی که مأمور در این محیط احتیاج دارد ۱۲ است (این دسته‌ها معنی‌دارترین دسته‌ها برای یادگیری این محیط را نیز تشکیل می‌دهند). همانطور که مشاهده می‌شود مدلهایی که پارامتر ρ در آنها تنظیم شده است اکثراً قادر به خبرگی با ۱۲ دسته شده‌اند.در نمودار ۳ نحوهٔ تغییر تعداد دسته‌ها با مقدار ρ در طول زمان آورده شده است. همانطور که در نمودار مشخص است مقدار $\rho = 4$ موجب کاهش تعداد دسته‌ها در طول یادگیری گشته است. در آزمایش دیگری از مدل Q با مقدار پارامتر ثابت $\rho = 4$ استفاده شد. در این آزمایش مدل با ۱۲ دسته به خبرگی رسیده است. نمودار ۱ همچنین نشان می‌دهد که مدلهای $Q(R)$، $Q(E)$ و $Q(D)$ عملکرد بهتری نسبت به مدل Q داشته‌اند.



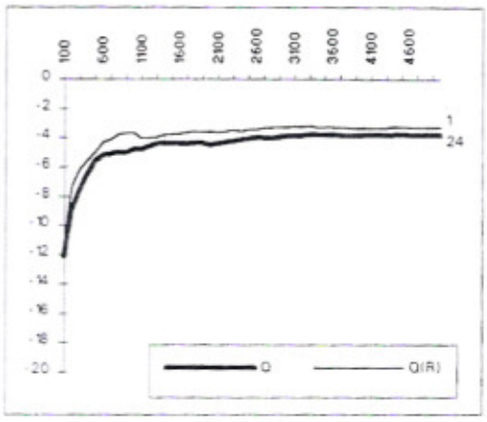
نمودار ۲- مقایسهٔ تغییرات پسخور متوسط مدلهای Q(ERDT) و Q در ۱۰۰۰۰ واحد زمانی.



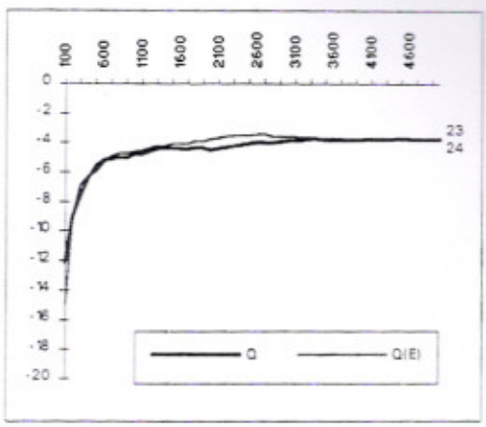
نمودار ۳. تغییرات ρ و تعداد دسته با زمان در مدل Q(R).



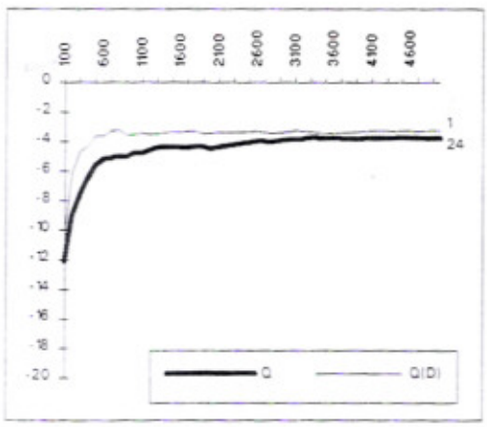
نمودار ۴



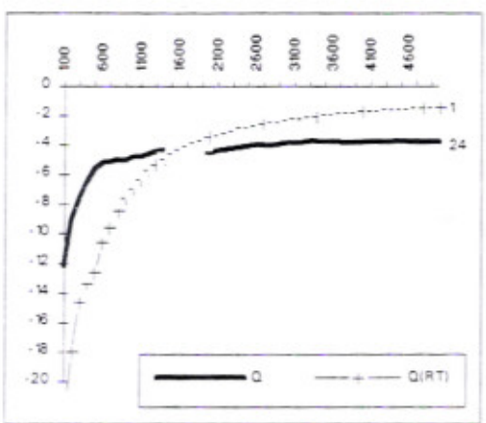
نمودار ۵



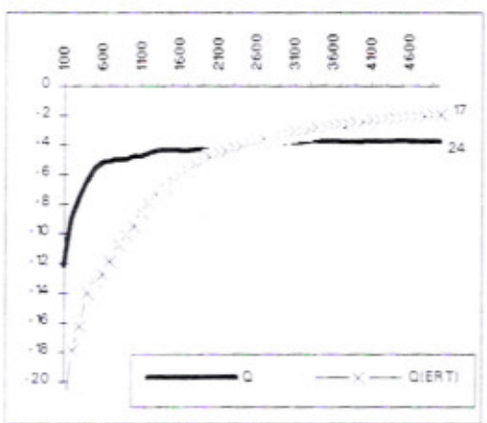
نمودار ۶



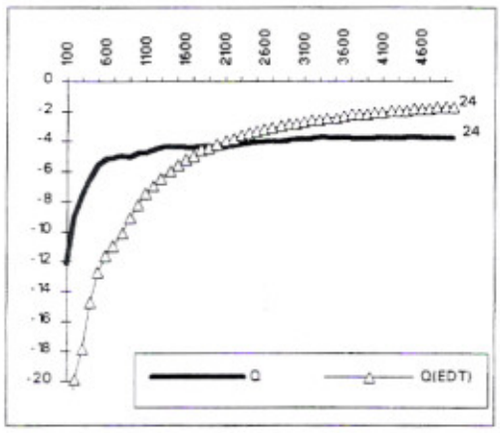
نمودار ۷



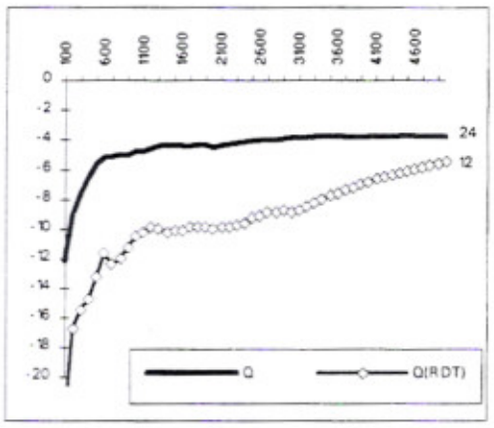
نمودار ۸



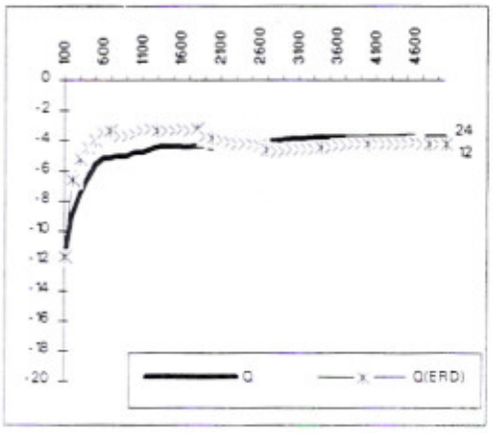
نمودار ۹



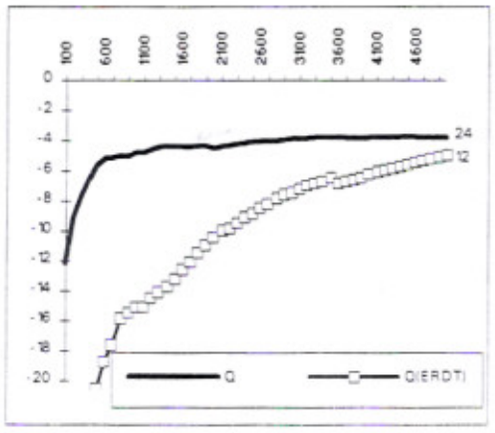
نمودار ۱۰



نمودار ۱۱



نمودار ۱۲



نمودار ۱۳

۶.جمعبندی

برای عملکرد بهینهٔ مدل Q در محیطهای مختلف باید مقادیر اولیهٔ مناسبی برای پارامترهای این مدل انتخاب شود. بعلاوه در محیطهای پویا با تغییر شرایط و قوانین محیطی مقادیر این پارامترها نیز باید تغییر کند. در این مقاله برای تنظیم پارامترهای مدل Q در حین یادگیری از بازی اتوماتونها استفاده شده است. آزمایشات اولیه در تنظیم پارامترهای θ، ρ، ε و δ نشان می‌دهد که تنظیم این پارامترها به این شکل علاوه بر مرتفع ساختن مشکلات انتخاب تجربی پارامترها توسط ناظر خارجی، باعث افزایش عملکرد مدل نیز می‌شود. در این بین سهم پارامتر θ در افزایش پسخور متوسط و پارامتر ρ در یافتن دسته‌های کمتر و معنی‌دارتر بیش از سایر پارامترها بوده است. آزمایشات همچنین نشان می‌دهند که افزایش تعداد پارامترهای تنظیم شده در مدلها تا حدودی باعث کاهش سرعت یادگیری می‌شود. در این مقاله کلیهٔ پارامترهای دسته‌بندی (ρ، ε و δ) برای تمام دسته‌ها مشترک بوده‌اند اما بنظر می‌رسد که استفاده از مقادیر متفاوت برای دسته‌های مختلف مناسبتر باشد. یافتن توابع ارزیابی مناسبتر برای اتوماتونها نیز می‌تواند باعث افزایش کارایی آنها در تنظیم پارامترها شود.

مراجع

[1] C.Watkins, Learning from delayed rewards, PhD. Thesis, Kings College, 1989.

[2] M Dorigo and H Bersini, A comparision of Q learning & classifier systems, proceedings of from Animats to Animals, International Conference on Simulation of Adaptive Behavior, SAB 1994.

[3] L. Kaelbling, Learning in embedded systems, PhD. Thesis, Stanford University, Stanford. CA, 1990.

[4] Sidhar Mahadevan and Jonathan Connel, Automatic programming of behavior-based robots using reinforcement learning, Artificial Intelligence Journal 55, pp 311-365, 1992.

[5] Robert Schalkoff, Pattern recognition, Wiley International Editions, 1991.

[6] S. Hodjat and M.R. Meybodi, Fine tuning of Q-learning parameters using learning automata, Proceedings of the 2nd Annual Conference of the Computer Society of Iran, pp 33-44, 1996.

[7] K. S. Narendra and M. A.L. Thathachar, Learning automata, Prentice Hall, 1989.

[8] M.R. Meybodi and S. Lakshmivarahan, ε-Optimality of a general class of absorbing barrier learning algorithms", Information Sciences, 28, pp. 1-20, 1982.

[9] A. Barto, R. Sutton and C. Anderson, Neuronlike adaptive elements that can solve difficult learning control problems, IEEE Trans. Syst. Man Cybern, 13(5), pp 834-846.

[10] R. Sutton, Integrated architectures for learning, planning and reacting based on approximating dynamic programming., Proceedings of Seventh International Conference on Machie Learning, Austin, TX, pp 216-224, 1990

[11] T. M. Mitchell, Generalization as search, Artif. Intell. 18(2) pp 203-226, 1988.

PROCEEDING OF THE

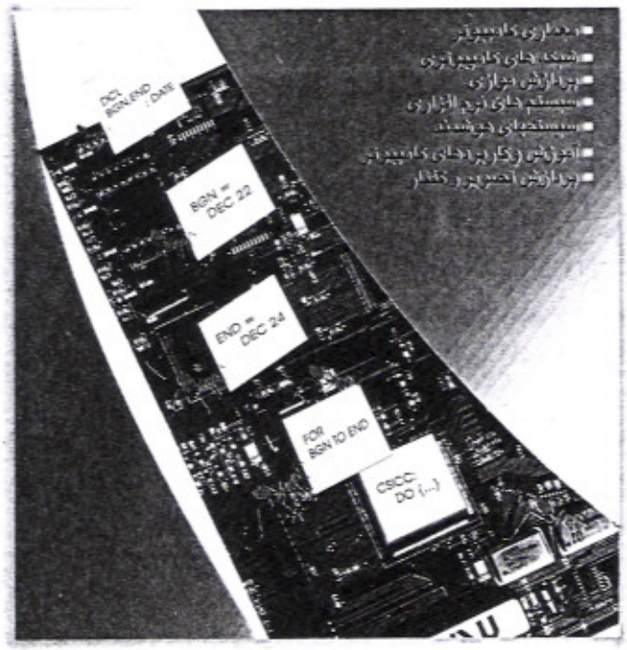
THIRD INTERNATIONAL

ANNUAL COMPUTER

SOCIETY OF IRAN

COMPUTER CONFERENCE

CSICC'97



COMPUTER CONFERENCE

Iran University Of Science And Technology

Computer Engineering Department

TEHRAN - I.R.IRAN

23-25 Dec. 1997