مدل یادگیری Q سلولی و کاربردهای آن

رضا رستگار محمدرضا میبدی

آزمایشگاه محاسبات نرم
دانشکده مهندسی کامپیوتر و فناوری اطلاعات
دانشگاه صنعتی امیرکبیر
تهران ایران
(Rastegar,meybodi)@ce.aut.ac.ir

چکیده: اتوماتای سلولی برای مدل کردن سیستمهایی مناسب است که قطعیت در تغییر حالات سیستم وجود داشته باشد. در حالیکه اغلب سیستمهای واقعی پیچیده بوده و ویژگی نویزی بودن و عدم قطعیت و احتمالی بودن در آنها دیده می شود و به همین دلیل برای مدل کردن چنین سیستم هایی استفاده از اتوماتای سلولی با قوانین قطعی منطقی به نظر نمی رسد. در این مقاله از ترکیب یادگیری Q با اتوماتای سلولی مدل جدیدی به نام یادگیری Q سلولی[1] معرفی میگردد. این مدل جدید با استفاده از قابلیت های یادگیری Q مشکل نبود عدم قطعیت در تغییر حالات در اتوماتای سلولی را تا حدودی مرتفع می سازد. کاربرد این مدل ترکیبی در مساله تخصیص کانال در شبکه های سلولی مخابراتی مورد بررسی قرار میگیرد.

کلمات کلیدی: اتوماتای سلولی، یادگیری Q، یادگیری Q سلولی ، شبکه سیار سلولی ، تخصیص کانال، یادگیری

۱ - مقدمه

یادگیری می تواند به عنوان یک راه کار برای ایجاد تطبیق پذیری در اکثر سیستمهایی که دارای فرایندهای تصمیم گیری بر اساس عدم قطعیت و اطلاعات ناقص می باشند مورد استفاده گیرد. با استفاده از یادگیری در جایگاههای مناسب در سیستم، هرجز سازنده سیستم می تواند حتی با دریافت اطلاعات ناقص و غیر قطعی، به صورت تدریجی و بر اساس معیارهای تعریف شده در سیستم به استراتژی بهینه کنترلی مورد نیاز خود دست یابد. در یادگیری تقویتی که یکی از انواع مهم مدلهای یادگیری میباشد، یک عامل یادگیرنده در طی فرایند یادگیری با تعاملات[2] مکرر با محیط، به یک سیاست کنترل بهینه می‌رسد. کارایی این تعاملات با محیط بوسیله بیشینه(کمینه) بودن پاداشی (جریمه ای) که از محیط گرفته می‌شود، ارزیابی می‌گردد. از آنجاییکه این روش یادگیری در محیطی بلادرنگ انجام میگیرد، می‌توان آنرا همزمان با فعالیت محیط انجام دادکه در این صورت با تمام رخدادهای پیش‌بینی نشده بصورت یک تجربه جدید برخورد می‌شود و می‌توان از آنها برای بهبود کیفیت یادگیری استفاده کرد. مزیت عمده یادگیری تقویتی نسبت به سایر روشهای یادگیری عدم نیاز به اطلاعات بجز سیگنال تقویتی از محیط میباشد.

یکی از مدلهایی که در شبیه سازی و یا مدل کردن سیستمها مورد استفاده قرار می گیرد، اتوماتای سلولی است که اجزای آن به صورت مکانی توزیع شده اند و اطلاعات از طریق قوانین محلی حاکم بر سیستم به صورت جزئی بین اجزا رد و بدل می شوند[3] . در اتوماتای سلولی، فضا بصورت یک شبکه ای از سلولها تعریف می‌گردد، زمان بصورت گسسته پیش می‌رود و قوانین آن بصورت سرتاسری است که از طریق آن در هر مرحله هر سلول، وضعیت جدید خود را با در نظر گرفتن وضعیت همسایه‌های خود بدست می‌آورد. قانون اتوماتای سلولی، نحوه تاثیر پذیرفتن یک سلول از سلولهای همسایه خود را مشخص می کند. یک سلول، همسایه سلول دیگر گفته میشود اگر بتواند آن سلول را در یک مرحله و براساس قانون حاکم تحت تاثیر قرار دهد. ویژگی‌های اساسی اتوماتای سلولی، فضای گسسته، زمان گسسته، محدود یت تعداد وضعیتهای ممکن هر سلول، یکسان بودن تمام سلولها، قطعی بودن قوانین، وابستگی قانون در هر سلول به مقادیر سلولهای اطراف آن و وابستگی قانون به مقادیر تعداد محدودی از مراحل قبل همسایه ها و خود سلول می باشند. در اتوماتای سلولی همگام[3] عمل بروز در آوردن سلولها به صورت همگام و در اتوماتای سلولی نا همگام[4] عمل بروز در آوردن سلولها به بصورت ناهمگام انجام میگیرد.

یکی از مشکلات اتوماتای سلولی تعیین فرم قطعی قوانین مورد نیاز برای یک کاربرد خاص است. اتوماتای سلولی برای مدل کردن سیستمهایی مناسب است که قطعیت در تغییر حالات سیستم وجود داشته باشد در

[1] Cellular Q-Learning
[2] Interaction
[3] Synchronous Cellular Automata
[4] Asynchronous Cellular Automata

صورتیکه اغلب سیستمها نویزی و دارای عدم قطعیت می باشند و وضع قوانین برای آنها به صورت قطعی، منطقی به نظر نمی رسد. روشهای متفاوتی برای حل این مشکل پیشنهاد شده است. یکی از این روشها احتمالاتی کردن قوانین می باشد اما مشکل این رهیافت، محاسبه این احتمالات برای سیستمهای ناشناخته می باشد. با معرفی اتوماتای یادگیر سلولی گامی در حل این مساله برداشته شده است.[۱][۱۹][۲۰][۱۸][21].

در این مقاله مدل دیگری با نام "یادگیرنده Q سلولی" برای مشکل فوقالذکر پیشنهاد می شود. یادگیرنده Q سلولی یک اتوماتای سلولی است که هر سلول آن به یک یا چند یادگیرنده Q که یکی از انواع یادگیری تقویتی میباشد مجهز است که وضعیت سلول را مشخص میکند. مانند اتوماتای سلولی، قانون محلی در محیط حاکم است و این قانون تعیین می کند که آیا عمل انتخاب شده توسط یک یادگیرنده Q در سلول باید پاداش داده شود ویا اینکه جریمه شود. دادن پاداش ویا جریمه منجر بروز درآوردن ساختار یادگیرنده Q سلولی بمنظور نیل به یک هدف مشخص می گردد. برای نشان دادن کاربرد یادگیرنده Q سلولی، از آن برای حل مسئله تخصیص کانال در شبکه های سلولی مخابراتی استفاده میشود.

ادامه مقاله بدین صورت سازماندهی شده است. در ابتدا در بخش ۲ به بررسی اتوماتای سلولی می پردازیم. در بخش ۳ یادگیریQ شرح داده میشود. در بخش ۴ مدل یادگیرنده Q سلولی ارائه می شود. در بخش ۵، کاربرد یادگیرنده Q سلولی در حل مسئله تخصیص کانال در شبکه های سلولی مخابراتی مورد بررسی قرار می گیرد. بخش پایانی نتیجه گیری میباشد.

۲- اتوماتای سلولی

اتوماتای سلولی شبکه‌ای سلولی است که هر سلول می‌تواند k حالت (وضعیت) داشته باشد. در هر سلول یک اتوماتا با حالات محدود[5] قرار دارد. در حالت یک بعدی، هر سلول دو همسایه نزدیک به خود دارد. در این حالت، وضعیت سلول i در زمان $t+1$ یعنی $a_i^{(t+1)}$ مطابق فرمول زیر بدست می‌آید.

$$a_i^{(t+1)} = \phi(a_{i-1}^{(t)}, a_i^{(t)}, a_{i+1}^{(t)}) \qquad (۱)$$

تابع ϕ انون اتوماتای سلولی نامیده میشود. همسایگی در اتوماتای سلولی یک بعدی را می‌توان بگونه‌ای بسط داد که از دو همسایه بیشتر را نیز شامل شود. یعنی می‌توان شعاع r را برای همسایگی در نظر گرفت. البته معمولاً نزدیک‌ترین همسایه‌ها را در نظر می‌گیریم. همچنین سلولها در اتوماتای سلولی می‌توانند در شبکه‌ای با هر ابعادی قرار گیرند که متناسب با بعد، تعاریف مربوط به همسایگی و قانون تغییر می یابند. متداولترین اتوماتای سلولی، اتوماتای سلولی دو بعدی است. چند نوع همسایگی مهم در این نوع اتوماتای سلولی همسایگی مور[6]، ون نیومن، کول[7] و اسمیت[8] می باشند [۳].

۳- یادگیری Q

یک سیستم با مجموعه حالات محدود و قابل شمارش، S، را در نظر بگیرید. یک کنترل کننده در هر حالت $s \in S$، یک عمل a را از میان مجموعه اعمال مجاز، $A(s)$، انتخاب و انجام می دهد. پس از آن سیستم با احتمال $p(s,a,s')$، از حالت s به s' می رود و پاداش $r(s,a,s')$ را به کنترل کننده می دهد. هدف کنترل کننده ماکزیمم کردن تابع زیر در تمامی حالات است :

$$J(s) = E\{\int_0^{\infty} e^{-\beta t} r(t)\,dt \,|\, s\} \qquad (۳)$$

که $E\{.|s\}$ امید ریاضی پاداش در یافت شده برای تمام مسیرهای ممکن آغاز شونده از s، $r(t)$ نرخ عایدی کل در زمان t، و β نرخ تخفیف[9] می باشند. در حالتی که زمان پیوسته نباشد می توانیم فرمول (۳) را به صورت زیر بنویسیم :

$$J(s) = E\{\sum_{t=0}^{T} \gamma^t r(t)dt \,|\, s\} \qquad (۴)$$

به طوری که γ نرخ تخفیف گسسته است. مقدار بهینه تابع J ، J^*، طبقه معادله بلمن[10] به صورت زیر تعریف می شود:

$$J^*(s) = \max_{a \in A(s)} [E_{\Delta t, s'}\{c(s,a,s') + \gamma(\Delta t) J^*(s')\}] \qquad (۵)$$

که Δt زمان تصادفی تا رویداد بعدی، و $\gamma(\Delta t)$ نرخ تخفیف موثر برای حالت بعدی s'، $\gamma(\Delta t) = e^{-\beta t}$ می باشند.

تخمین J^* می تواند از طریق یادگیری Q انجام گیرد. معادله بلمن را می توان با استفاده از مقادیر Q به صورت زیر بازنویسی کرد:

$$J^*(s) = \max_{a \in A(s)} Q^*(s,a) \qquad (۶)$$

در آغاز مقادیر به صورت تصادفی مقدار دهی می شوند و در ادامه با هر انتقال حالت در سیستم مقادیر Q بروز رسانی می شوند. و در حالت s عملی که بیشترین مقدار $Q(s,.)$ را داشته باشد انتخاب می گردد. اگرعامل یادگیرنده با انتخاب عمل a باعث انتقال سیستم به s' شود و

[5] Finite State Automata

[6] Moore
[7] Cole
[8] Smith
[9] Discount Factor
[10] Bellman Equation

پاداش $r(s,a,s')$ را دریافت کند، مقدار $Q(s,a)$ به صورت زیر تنظیم می شود:

$$Q(s,a) = (1-\alpha)Q(s,a) + \alpha(r(s,a,s') + \gamma(\Delta t) \max_{b \in A(s')} Q^*(s',b)) \quad (۷)$$

به طوریکه $0 \le \alpha \le 1$ نرخ یادگیری می باشد. برای اطلاعات بیشتر در باره یادگیری Q میتوان به [۵ و۱۶] مراجعه کرد.

۴- یادگیرنده Q سلولی

"یادگیرنده Q سلولی" که از این پس آن را با نام CQL می‌شناسیم مدلی است که ازترکیب اتوماتای سلولی و یادگیری Q حاصل میشود. این مدل برای سیستم‌های طراحی شده که اجزای آنها از طریق تعامل با یکدیگر از تجربیات گذشته همدیگر اطلاع پیدا میکنند واز این طریق می‌توانند رفتار خود را اصلاح کنند. یک CQL را می توان به صورت یک شش تایی $< E, A, N, \Phi, Q, C >$ تعریف کرد که $E=\{e_1,\cdots,e_n\}$ مجموعه مکانهای تعریف شده در اتوماتای سلولی هستند که می‌توانند در بشکلهای مختلفی مانند خطی، دو بعدی و سه بعدی درکنار هم قرار گیرند. $A = \{a_1,\ldots,a_k\}$ مجموعه مقادیر مجازی است که یک سلول میتواند اختیار کنند. $A^t(e_i)$ نشان دهنده مقدار سلول e_i درزمان t است. Φ قوانین حاکم بر جامعه سلولی است که پاداشها و جریمه ها براساس آن تعیین میشود. $N(e_i)$ مجموعه همسایه‌های سلول e_i راتعریف می‌کند که این مجموع دارای این ویژگیهاست:

$$e_i \notin N(e_i) \quad \forall e_i \in E \quad (۸)$$

$$e_i \in N(e_j) \leftrightarrow e_j \in N(e_i) \quad \forall e_i, e_j \in E \quad (۹)$$

هر قانون $\phi \in \Phi$ را می‌توان با توجه به مفاهیم قوانین عمومی و جمعی و یا وزن دار به یکی از فرمهای زیر تعریف کرد:

قوانین عمومی:

$$< a_1,\ldots,a_h > \rightarrow r \quad (۱۰)$$

برای مثال قانون $< 1,0,0,0,0 > \rightarrow -1$ با فرض همسایگی ون نیومن به معنای آن است که در صورت ۱ بودن سلول مرکزی و صفر بودن بقیه سلولهای همسایه، این خانواده جریمه شود. که این جریمه بر اساس تابع تخصیص پاداش C بین آنها تقسیم می شود.

قوانین جمعی:

$$< a_1 \times n_1,\ldots,a_k \times n_k > \rightarrow r \quad (۱۱)$$

$$\sum_1^k n_i = |N| + 1 \quad (۱۲)$$

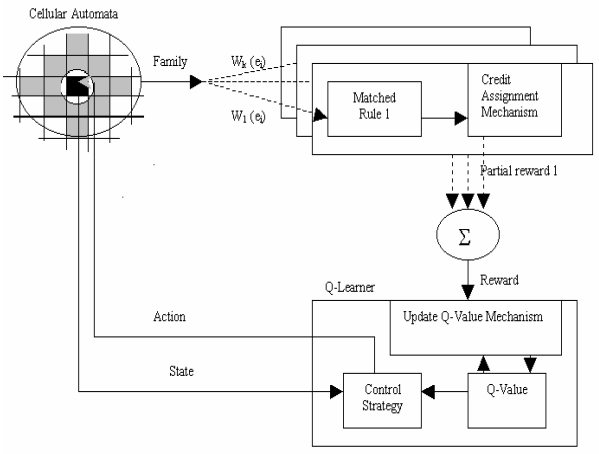
قوانین وزن دار:

$$< a_1 \times n_1,\ldots a_k \times n_k > \rightarrow \sum_1^k n_i \times w_i \quad (۱۳)$$

که $a_i \in A$ ، r مقادیر مجاز برای پاداش و تنبیه و w_i وزنی است که به هر وضعیت داده می شود.. اگر $W(e_i) = N(e_i) \bigcup \{e_i\}$، در این صورت $C = C(W(e_i))$ وظیفه تقسیم پاداش بدست آمده از $\phi = \phi(W(e_j))$ مابین اعضای $W(e_i)$ را دارد. بر طبق تابع تخصیص پاداش C به هر سلول براساس کارایی آن پاداش داده می‌شود. حال که با این مدل آشنا شدیم به شرح عملکرد آن می‌پردازیم. در نتی‌جه در ی‌افت ی‌ک درخواست توسط سلول e_i، ای‌ن سلول فعال می‌شود (در مدل آسنکرون تمام سلولها به طور همزمان فعال نمی شوند) پس از فعال شدن، سلول، e_i ، از میان از مجموعه A یک مقدار انتخاب می‌کند و سپس اعمالی بدی‌ن شرح را انجام می‌دهد. تابع ϕ برای سلول ارزی‌ابی می‌شود و سپس طبق تابع C، مقدار پاداش بدست آمده برای مجموعه $W(e_i)$ را بین اعضای آن تقسیم می‌کنیم. با چنین مکانیسمی در صورت سنکرون بودن اتوماتای سلولی، هر سلول e_i در گام t ام $|W(e_i)|$ خرده پاداش دریافت می‌کند که با جمع این خرده پاداش‌ها، پاداش کل سلول طبق (۱۴) بدست می‌آید که به یادگیرنده Q داده می شود و بر اساس آن مقادیر Q بهنگام می‌شوند.

$$sum = \sum_{\substack{\forall W(e_j) \\ \exists e_i \in W(e_j)}} C(W(e_j); e_i) \quad (۱۴)$$

اگر اتوماتای سلولی به صورت آسنکرون عمل کند پس از دریافت سیگنال تقویتی، تنها سلول فعال شده و همسایه های آن بر طبق تابع تخصیص پاداش C پاداش را دریافت می کنند. سپس مقادیر پاداشها محاسبه شدند و به یادگیرنده‌های هر سلول داده می‌شود. نهایتا سلول عمل بروز رسانی را انجام می‌دهد. مراحل فوق تا رسیدن سیستم به تعادل که در انجا مقدار هر سلول بهینه میباشد ادامه پیدا میکند.



شکل۱ : معماری یک سلول در مدل CQL

همانطوریکه که گفته شد در مدل ارائه شده، پس از مشخص پاداش مربوط به یک $W(e_i)$ ، سیگنال پاداش بین سلولهای آن تقسیم می‌شود که این روش تقسیم توسط، $C(W(e_i))$ تعریف می‌شود و مقدار بازگشتی آن سهم پاداش هر سلول $e_j \in W(e_i)$ از مقدار $\phi(W(e_i))$ را نشان می‌دهد. روشهای متفاوتی برای پیاده‌سازی تابع تخصیص پاداش C وجود دارد که از مهمترین آنها می‌توان به تقسیم مساوی، تقسیم تصادفی، تقسیم براساس ویژگیهای سلول از جمله خبرگی [۲] سلول یا موقعیت جغرافیایی اشاره کرد.

۵- حل مساله تخصیص کانال در شبکه های سیار سلولی توسط یادگیرنده Q سلولی

مسئله تخصیص کانال در شبکه های سیار سلولی[۶]یک مسئله-- NP complete میباشد و بهمین دلیل الگوریتمهای تقریبی متعددی برای حل این مساله از جمله الگوریتمهای مبتنی بر الگوریتم‌های ژنتیکی[۸]، تابکاری فلزات[۱۱] [۸]، جستجوی[۱۲]TABU [۸] ، شبکه‌های عصبی [۹]، برنامه سازی پویای نورونی[۱۳] [۴]و یادگیری Q [۱۰] گزارش شده اند.

در تمامی استراتژی‌های تخصیص کانال به صورت پویا، دانش و تجربه بدست آمده در طول کار سیستم به دست فراموشی سپرده می‌شود. اگر چه استراتژی‌های مبتنی بر شبکه‌های عصبی از آموزش بهره میبرند ولی در همه آنها داشتن یک ناظر خوب (یک استراتژی شناخته شده تخصیص پویای کانال) ضروری است. به نظر می‌رسد که استفاده از روشهای یادگیری بدون نیاز به داشتن استراتژی معین و شناخته شده بتواند گره‌گشای این مشکل باشد [۱۰]. در [۴ و۱۰] دو استراتژی تخصیص کانال مبتنی بر الگوریتم های یادگیری تقویتی که در آنها نیازی به داشتن استراتژی معین و شناخته شده نیست ارائه گردیده است. در [۴] از یک معماری خطی تخمینی به همراه $TD(0)$ استفاده شده است. در این استراتژی هدف ماکزیمم کردن $E[\int_{t=0}^{\infty} e^{-\beta t} n(t)dt]$ می باشد که $n(t)$ تعداد مکالمات در جریان در زمان t و β نرخ تخفیف است. قابلیت توزیع شدگی و کارایی بالا از ویژگی های این الگوریتم است. در الگوریتم ارائه شده در [۱۱] از یادگیری Q برای تخصیص کانال به صورت تمرکز یافته[۱۴] استفاده شده است. آنچه الگوریتم پیشنهادی در مقاله را از الگوریتم متمایز متمایز می سازد، مستقل شدن فرایند یادگیری در هرکدام از سلولهای شبکه می باشد که سبب می شود هر سلول بتواند بر اساس موقعیت جغرافیایی خود (به دلیل شرایط مرزی پوچ[۱۵]، سیستم متقارن نمی باشد.) استراتژی مناسب خود را یاد بگیرد. همچنین الگوریتم پیشنهادی کاملا توزیع شده میباشد که این خود از حجم پیغامهای کنترلی بر روی شبکه میکاهد. الگوریتم پیشنهادی با دو الگوریتم فوقالذکر مقایسه خواهد شد.

در ادامه این قسمت به توصیف یک الگوریتم پویای توزیع شده تخصیص کانال مبتنی بر یادگیرنده Q سلولی اسنکرون می‌پردازیم. در الگوریتم پیشنهادی هر سلول در شبکه سلولی به یک سلول در یادگیرنده Q سلولی نگاشت می‌شود. زمانیکه یک سلول فعال شود(زمانیکه یک درخواست مکالمه به سلول میرسد)، یادگیرنده Q این سلول یک عمل انتخاب و سپس قانون حاکم بر یادگیرنده Q سلولی ارزیابی میشود. براساس نتیجه ارزیابی قانون، فرایند یادگیری انجام می شود. در ادامه این بخش پارامترهای یادگیرنده Q سلولی تعریف و سپس الگوریتم تخصیص کانال پویای مبتنی بر یادگیرنده Q سلولی ارایه میشود.

مدل همسایگی: مدل همسایگی، مدل همسایگی مور توسعه یافته[۱۶] میباشد. مجموعه همسایه های هر سلول به دو دسته تقسیم می‌شوند: همسایه‌های ثابت و همسایه‌های متغیر. با فرض اینکه فاصله استفاده مجدد هم کانال R باشد، R حلقه حول سلول، سلولهای غیر همکانال یا همسایه‌های ثابت سلول می‌باشند که با سلول مرکزی تعامل محلی انجام می‌دهند. هر سلول در حالت کلی دارای $6(R+1)$ همسایه ثابت است که از این تعداد، حداکثر ۶ سلول همسایه مجاور می باشند و بقیه همسایگان غیر مجاور این سلول را تشکیل می‌دهند. هر سلول دارای یک مجموعه سلول هم کانال نیز می‌باشد که حالت آنها نیز در حالت فعلی سلول تاثیر دارد. مجموعه سلولهای همکانال یک سلول، سلولهایی هستند که در فاصله R+1 از سلول واقع هستند و می توانند از یک مجموعه کانال استفاده کنند. تعداد همسایه‌های متغیر یک سلول حداکثر ۶ تا می‌باشد. نکته‌ای که باید به آن توجه داشته باشیم آن است که با توجه به ویژگیهای فیزیکی شبکه مخابراتی سلولی اتوماتای سلولی استفاده شده در این الگوریتم دارای شرایط مرزی پوچ می‌باشد.

حالات سلول: در مدل استفاده شده هر سلول دارای متغیرهای حالت(U_i)متفاوتی است که تعداد آنها برابر تعداد کانالهای شبکه مخابراتی موبایل است و هر متغیر میتواند یکی از مقادیر *free* و *used* را اختیار کند.

[11] Simulated Annealing
[12] TABU Search
[13] Neuro-Dynamic Programming
[14] Centralized
[15] Null Boundary Conditions
[16] Extended Moore

قانون: قانون از نوع قانون وزن دار می‌باشد و تابعی از همسایه‌های متغیر وثابت سلول می‌باشد. پاداش $r(s,a,s')$ پاسخ آنی سیستم در برابر تخصیص کانال a درحالت s می‌باشد که به صورت زیر تعریف میشود:

$$r(s,a,s') = n_1(k)w_1 + n_2(k)w_2 \quad (۱۵)$$

که $n_1(k)$ تعداد سلولهای همکانال سلول iام می‌باشد که در آنها از کانال K استفاده شده است و $n_2(k)$ تعداد سلولهای همکانال می‌باشد که در لایه سوم همسایگی (با شرط این که فاصله استفاده مجدد برابر دو باشد) قرار دارند و کانال K در آنها موجود است. ضرایب $w_1 = -1, w_2 = +1$ مقادیر ثابتی هستند.

تابع تخصیص پاداش: تابع تخصیص پاداش C به صورت زیر تعریف می شود :

$$C(W(e_i),e_j) = \begin{cases} 0 & otherwise \\ \phi(W(e_i)) = r(s,a,s') & e_j \, is \, centralelb \, of \, W(e_i) \end{cases} \quad (۱۶)$$

یادگیرنده Q: در زیر پارامترهای یادگیرنده Q هر سلول e_i تعریف میشود.

- **حالت:** با فرض وجود N سلول و M کانال، حالت s به صورت زیر تعریف می‌شود(این تعریف مربوط به ساختار درونی یادگیرنده بوده و با تعریف حالات اتوماتای سلولی متفاوت است) :

$$s = \sum_{i=1}^{M} H(i)2^{i-1} \quad (۱۷)$$

به طوری که:

$$H(i) = \begin{cases} 1 & \text{Channel } i \text{ is not available} \\ 0 & \text{Otherwise} \end{cases} \quad (۱۸)$$

- **عمل**: تخصیص یک کانال a از میان کانالهای $A(i)$ به درخواست رسیده به سلول i ام :

$$a = k \quad , \quad k \in A(i) \quad (۱۹)$$

که $A(i)$ ، مجموعه کانالهای موجود[17] در سلول i است.

- **حالت بعدی:** با توجه به تعریف حالت که قبلا به ان اشاره شد، حالت بعدی از طریق واکنش الگوریتم به رویداد رسیده و حالت فعلی قابل محاسبه است.
- **استراتژی کنترل**: در اینجا از استراتژی کنترلی انتخاب Soft-Max[18] استفاده شده است که در آن احتمال انتخاب کانال k در حالت s بصورت زیر تعریف می شود:

$$e^{Q(s,k)/T} / \sum_{l=1}^{M} e^{Q(s,l)/T} \quad (۲۰)$$

که T پارامتر دما است که با گذشت زمان کاهش می یابد. مقدار مینیمم T برابر ۱ میباشد.

- **نرخ یادگیری:** در الگوریتم پیاده سازی شده از نرخ یادگیری ثابت[19] و متغیر کاهش یابنده[20] استفاده شده است. نرخ یادگیری کاهش یابنده به صورت زیر تعریف می شود:

$$\alpha(s,a) = 1/visit(s,a) \quad (۲۱)$$

که $visit(s,a)$ تعداد دفعاتی است که عامل یادگیرنده، عمل a در حالت s را انتخاب می کند.

راههای متفاوتی برای ذخیره سازی مقادیر Q (شبکه عصبی ، درخت تصمیم گیری[21] و جدول) وجود دارد. در این تحقیق از روش ذخیره سازی به شکل جدول استفاده شده است. از آنجا که حافظه مورد نیاز برای جدول، نمایی نسبت به تعداد کانالها در شبکه افزایش میابد و با توجه به اینکه در عمل تعداد کانالها زیاد است، حافظه موردنیاز برای ذخیره سازی مقادیر Q بسیار بالا خواهد بود. برای حل این مشکل فضای حالت را به چند مجموعه افراز کرده و برای هر مجموعه یک سطر در جدول Q ذخیره شده است.

حال به شرح الگوریتم تخصیص کانال پیشنهادی میپردازیم. هر گاه یک درخواست مکالمه به سلول i میرسد مراحل زیر انجام میگیرد.

۱- برای تمام سلولهای در همسایگی ثابت، پیغام give_used_channels فرستاده می‌شود. سلولهای همسایه پس از دریافت این پیغام با ارسال پیغام get_used_channels که شامل لیست کانالهای اشغال شده خود می‌باشد به سلول i پاسخ می‌دهند.

۲- سلول i، براساس پیغامهای دریافتی، حالت s را محاسبه می‌کند و از یادگیرنده Q خود بهترین کانال را درخواست می‌کند. یادگیرنده نیز براساس استراتژی کنترل Soft-Max، کانال k را انتخاب می‌کند. در صورتی که کانالی موجود نباشد، درخواست جدید مسدود می‌شود.

۳- پس از انتخاب کانال و انتساب آن، سلول i بارسال پیغام lock_channel(k)به تمام سلولهای در همسایگی ثابت خود از آنها می‌خواهد کانال k را در خود قفل کنند.

۴ - پیغام give_channel_status(k) به تمام همسایگان سلول متغیر ارسال می‌شود. هر همسایه پس از دریافت پیغام، وضعیت کانال k را در پیغام get_channel_status(k) به سلول i ارسال می‌کند.

[17] Available
[18] Soft-Max Selection Strategy Control

[19] Constant Learning Rate
[20] Constant Learning Rate
[21] Decision Tree

۵ - سلول i با آگاهی از وضعیت کانال k در تمام سلولهای همسایه مقدار پاداش خود را محاسبه و به یادگیرنده می‌دهد و یادگیرنده خود را به روز می‌کند.

۶- پس از پایان مکالمه سلول i به تمام همسایگان ثابت خود پیغام unlock_channel(k) ارسال می‌کند و از آنها می‌خواهد کانال k را آزاد کنند.

برای ارزیابی کارایی الگوریتم پیشنهادی(CQL-CA)، یک شبکه سلولی ۷ در ۷ با تعداد کانال ۱۲ و شعاع استفاده مجدد ۲ ($۱۲^{۲}$حالت) در نظر گرفته شده است. متوسط زمان مکالمه و متوسط زمان تحویل کانال به ترتیب ۳ و ۲ دقیقه فرض شده است. در شبیه سازیها سه الگوریتم FA، Bert [۴] و Hykin [۱۱] پیاده سازی و با نتایج الگوریتم پیشنهادی مقایسه شده اند. همچنین کارایی الگوریتم پیشنهادی با نرخ یادگیری ثابت و متغیر کاهش یابنده مورد بررسی قرار دادیم. در تمام آزمایشها تحویل کانال در نظر گرفته شده است. نتایج آزمایشها که به تفصیل در [11] آمده است نشان دهنده برتری الگوریتم پیشنهادی در مقایسه با سه الگوریتم FA، Bert و Hykin میباشد.

۶- نتیجه گیری

در این مقاله یک مدل یادگیری جدید به نام یادگیرنده Q سلولی پیشنهاد و کاربرد آن. برای حل مساله تخصیص کانال در شبکه های سلولی مخابراتی ارایه گردید. از طریق شبیه سازی نشان داده شد که الگوریتم پیشنهادی برای تخصیص کانال مبتنی بر این مدل ترکیبی از کارایی خوبی بر خوردار است.

مراجع

[1] Meybodi, M. R, Beigy, H. and Taherkhani, M., "Cellular Learning Automata", Proceedings of 6th Annual CSI Computer Conference, Computer Engineering Department, University of Isfahan, pp. 153 –163, 20-22 Feb. 2001.

[2] Ahmadabadi, M. N. and Asadpour, M., "Expertness Based Cooperative Q-Learning", IEEE Transactions on Systems, Man, and Cybernetics, Vol. 32, No. 1, 2002.

[3] Wolfram, S., Cellular Automata and Complexity, Perseus Books Group, 1994.

[4] Bertsekas, D. P. and Singh, S.,"Reinforcement Learning for Dynamic Channel Allocation in Cellular Telephone Systems", NIPS96 Proceeding, 1996.

[5] Bertsekas, D. P. and Tsitsiklis, J. N., Neuro-Dynamic Programming, Athena Scientific, Belmont, Massachuset, 1996.

[6] Brown, T. and Tong, H., "Adaptive Resource Allocation in Telecommunications", Denver, Proceeding of the SPIE, 1999.

[7] Boukerche, A. and Jacob, T., "A Distributed Algorithm for Dynamic Channel Allocation", Kluwer Academic Publishers, Netherlands, Mobile Networks Journal, Vol. 7, PP. 115-126, 2002.

[8] Chen, J., Seah, D. and Xu, W., "Channel Allocation for Cellular Networks Using Heuristic Methods", unpublished report, 1999.

[9] Funabiki, N., "A Neural Network Parallel Algorithm for Channel Assignment Problems in Cellular Radio Networks", IEEE Transactions on Vehicular Technology, Vol 41, No. 4, 1992.

[10] Hykin, S. and Nie, J., "A Dynamic Channel Assignment Policy through Q-learning", IEEE Transactions on Neural Networks, Vol. 10, No. 6, 1999.

[11] Rastegar, R. and Meybodi, M. R., "CQL and Its Applications", Technical Report, Computer Engineering Department, Amirkair University, 2005.

[12] Krumke, S., Marathe, M. and Ravi, S., "Approximation Algorithms Assignment in Radio Networks", Dallas, International Workshop on Discrete Algorithms and Methods for Mobile Computing and Communications, 1996.

[13] Lawrence, K. and Peter, Yum T., "Phantom Cell Analysis of Dynamic Channel Assignment in Cellular Mobile Systems", IEEE Transactions on Vehicular Technology, Vol. 47, No. 1, 1998.

[14] Lawrence Young, K. and Yum, K., "Compact Pattern Based Dynamic Channel Assignment for Cellular Mobile Systems", IEEE Transactions on Vehicular Technology, Vol. 43, No. 4, 1994.

[15] Smith K., "A Genetic Algorithm for the Channel Assignment Problem", unpublished report, 1998.

[16] Sutton, R., and Barto, A., Reinforcement Learning: An Introduction, MIT Press, 1998.

[17] Tong, H. and Brown, T. X., "Reinforcement Learning for Call Admission Control and Routing under Quality of Service Constraints on Multimedia Networks", Accepted in Machine Learning Journal, 2000.

[18] Meybodi, M. R. and Kharazmi, M. R., "Cellular Learning Automata and Its Application to Image Processing", Journal of Amirkabir, Vol. 14, No. 56A, pp. 1101-1126, 2004.

[19] Meybodi, M. R, Beigy, H. and Taherkhani, M., "Cellular Learning Automata and Its Applications", Journal of Science and Technology, University of Sharif, No. 25, pp.54-77, Autumn/Winter 2003-2004.

[20] Beigy, H. and Meybodi, M. R., "A Mathematical Framework for Cellular Learning Automata", Advances on Complex Systems, Vol. 7, Nos. 3-4, pp. 295-320, September/December 2004.

[21] Beigy, H. and Meybodi, M. R., "Open Synchronous Cellular Learning Automata", Journal of Computer Science and Engineering, ۲005, to appear.