
بسم الله الرحمن الرحيم

دانشکده مهندسی کامپیوتر

پایان‌نامه کارشناسی ارشد مهندسی کامپیوتر- هوش مصنوعی

پیکربندی شبکه‌های بی‌سیم شامل مواد مبتنی‌بر نرم‌افزار با کمک یادگیری تقویتی

دانشجو: فاطمه علیان نژادی

استاد راهنما

دکتر اسماعیل طحانیان

اساتید مشاور

دکتر منصور فاتح

دکتر محسن رضوانی

مهر ۹۹

شماره: ۷/۷/د/۰۰۰
تاریخ: ۹۹/۸/۱۹

باسمه‌تعالی

مدیریت تحصیلات تکمیلی

فرم شماره (۳) صورتجلسه نهایی دفاع از پایان نامه دوره کارشناسی ارشد

با نام و یاد خداوند متعال، ارزیابی جلسه دفاع از پایان نامه کارشناسی ارشد خانم / آقای ..... با شماره دانشجویی ..... رشته ..... گرایش ..... تحت عنوان .... که در تاریخ ۹۹/۷/۱۶ ..... با حضور هیأت محترم داوران در دانشگاه صنعتی شاهرود برگزار گردید به شرح ذیل اعلام می گردد:

الف) درجه عالی: نمره ۲۰-۱۹ ☐ ب) درجه خیلی خوب: نمره ۱۸/۹۹ – ۱۸ ☑

ج) درجه خوب: نمره۱۷/۹۹– ۱۶ ☐ د) درجه متوسط: نمره ۱۵/۹۹-۱۴ ☐

ه) کمتر از ۱۴ غیر قابل قبول و نیاز به دفاع مجدد دارد ☐

نوع تحقیق: نظری ☐ عملی ☐

عضو هیأت داوران	نام ونام خانوادگی	مرتبهٔ علمی	امضاء
۱ـ استادراهنمای اول	اسماعیل طهانیان	استادیار	
۲- استادراهنمای دوم			
۳- استاد مشاور			
۴- نماینده تحصیلات تکمیلی			
۵- استاد ممتحن اول			
۶ـ استاد ممتحن دوم			

نام و نام خانوادگی رئیس دانشکده

تاریخ و امضاء و مهر دانشکده

۴

ماحصل آموخته‌هایم را تقدیم می کنم به آنان که مهر آسمانی شان آرام بخش آلام زمینی ام است.

به استوارترین تکیه گاهم، دستان پرمهر پدرم

به سبزترین نگاه زندگیم، چشمان سبز مادرم

که هرچه آموختم در مکتب عشق شما آموختم و هرچه بکوشم، قطره ای از دریای بی کران مهربانیتان را

سپاس نتوانم بگویم.

امروز، هستی من به امید شما است و فردا کلید باغ بهشتم، رضای شما.

باشد که حاصل تلاشم، نسیم گونه، غبار خستگیتان را بزداید.

بوسه بر دستان پرمهرتان.

اکنون که به یاری پروردگار و راهنمایی اساتید بزرگ موفق به پایان این پایان نامه شده ام،

وظیفه خود دانسته که نهایت سپاسگزاری را از تمامی عزیزانی که در این راه

به من کمک کرده اند را به عمل آورم:

در آغاز از استاد گرامی جناب آقای دکتر طحانیان که راهنمایی این پایانامه را به عهده داشته اند؛

و به دلیل یاری ها و راهنمایی های بی دریغشان

بسیاری از سختی ها را برایم آسان نمودند، کمال تشکر را دارم.

همچنین سپاس فراوان از

آقایان دکتر فاتح و دکتر رضوانی که زحمت تقبل مشاوره این پژوهش را در اختیار داشتند؛

و کمال تشکر را جهت زحمات و خدمات دلسوزانه شان دارم

و تمامی آن هایی که مرا راهنمای زندگی بوده اند.

فاطمه علیان نژادی

مهر ۱۳۹۹

تعهدنامه

اینجانب فاطمه علیان نژادی دانشجوی دوره کارشناسی ارشد رشته کامپیوتر گرایش هوش مصنوعی

دانشکده کامپیوتر دانشگاه صنعتی شاهرود نویسنده پایان نامه " پیکربندی شبکه‌های بی‌سیم شامل مواد مبتنی‌بر نرم‌افزار با کمک یادگیری تقویتی" تحت راهنمائی دکتر اسماعیل طحانیان متعهد می شوم.

- تحقیقات در این پایان نامه توسط اینجانب انجام شده است و از صحت و اصالت برخوردار است.
- در استفاده از نتایج پژوهشهای محققان دیگر به مرجع مورد استفاده استناد شده است.
- مطالب مندرج در پایان نامه تاکنون توسط خود یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچ جا ارائه نشده است.
- کلیه حقوق معنوی این اثر متعلق به دانشگاه صنعتی شاهرود می باشد و مقالات مستخرج با نام « دانشگاه صنعتی شاهرود » و یا « Shahrood University of Technology» به چاپ خواهد رسید.
- حقوق معنوی تمام افرادی که در به دست آمدن نتایج اصلی پایان نامه تأثیرگذار بوده اند در مقالات مستخرج از پایان نامه رعایت می گردد.
- در کلیه مراحل انجام این پایان نامه ، در مواردی که از موجود زنده (یا بافتهای آنها) استفاده شده است ضوابط و اصول اخلاقی رعایت شده است.
- در کلیه مراحل انجام این پایان نامه، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته یا استفاده شده است اصل رازداری ، ضوابط و اصول اخلاق انسانی رعایت شده است.

.

تاریخ

امضای دانشجو

چکیده

با افزایش استفاده از دستگاه‌های بی‌سیم هوشمند و ظهور ایده اینترنت اشیا، سیستم‌های ارتباطی بی‌سیم با تقاضای نرخ داده بالاتر و همچنین کیفیت بهتر خدمات روبرو شده‌اند. برای دستیابی به چنین میزان داده بالا، شبکه‌های بی‌سیم باید در فرکانس‌های بالا مانند باندهای میلی‌متر و تراهرتز، کار کنند. از طرفی با افزایش فرکانس، امواج قادر به عبور از موانع مانند دیوار نیستند و در محیط تلف می‌شوند که باعث عمق نفوذ کمتر می‌شوند و همچنین محوشدگی چند مسیره، به دلیل برهم کنش سیگنال‌هایی که از چند مسیر مختلف در هوا از فرستنده به گیرنده می‌رسند، استفاده از چنین فرکانس‌های بالایی از امواج الکترومغناطیسی را محدود می‌کنند. محققان برای کنترل این موارد، ایده پوشش اشیاء موجود در محیط با فراسطوح مبتنی‌بر نرم افزار را ارائه کردند تا محیط انتشار را به یک محیط قابل برنامه‌ریزی تبدیل کنند. در این روش تمام دیوارها توسط قطعاتی با اندازه‌های یکسان از فرا سطوح که به آن‌ها کاشی گفته می‌شود، پوشانده می‌شوند. چالش اصلی این است که محیط در طول ارتباط می‌تواند دست‌خوش تغییرات شود، به عنوان مثال یک مانع ناگهانی مسیر سیگنال را مسدود کند. از این رو در این پایان‌نامه الگوریتم آموزش یادگیری تقویتی (RL) برای ایجاد چنین محیطی ارائه شده است. از آن‌جا که یک عامل در الگوریتم RL به تدریج با محیط در تعامل است می‌تواند تصمیم بهتری بگیرد، به خصوص زمانی که محیط تغییر می‌کند. نتایج روش پیشنهادی با کار گذشته که از الگوریتم ژنتیک برای ایجاد محیط قابل برنامه‌ریزی استفاده کرده، مقایسه شده است. در کار گذشته، میانگین توان دریافتی برای ۱۲ گیرنده، ۲۵.۳۸ dBmW می‌باشد و روش مبتنی‌بر یادگیری تقویتی در این پایان‌نامه، میانگین توان دریافتی را ۱۲ درصد افزایش داده است.

کلمات کلیدی: فراسطوح مبتنی‌بر نرم افزار (SDM)، محیط بی‌سیم قابل برنامه‌ریزی (PWE)، یادگیری تقویتی (RL)، عامل، امواج الکترومغناطیسی.

لیست مقالات مستخرج از پایان نامه

Submitted to Soft Computing journal:

A Reinforcement Learning-Based Configuring Approach in Next Generation Wireless Networks Using Software Defined Metasurface.

فهرست مطالب

فهرست شکل‌ها

فهرست جداول

فصل ۱: کلیات

۱-۱ مقدمه

ارتباط بی‌سیم[1] به انتقال اطلاعات بدون سیم و به وسیله امواج الکترومغناطیسی[2] گفته می‌شود [۱، ۲]. واژه بی‌سیم پس از اختراع تلگراف بی‌سیم و در مقابل مخابرات با سیم[3] ابداع شد. بی‌سیم‌ها انواع گوناگون دارند و در کاربردهای مختلف رسانه‌ای، صنعتی، نظامی، تفریحی و در باندهای فرکانسی و توان‌های ارسال و دریافت متفاوت در کاربردهایی مانند تلفن سلولی، سامانه موقعیت‌یاب جهانی، دستگاه‌های کنترل از راه دور، صفحه کلید بی‌سیم و تلویزیون ماهواره‌ای مورد استفاده قرار دارند.

مخابرات بی‌سیم به عنوان رشته‌ای علمی از دهه ۱۹۶۰ میلادی مورد مطالعه بوده، اما از اواسط دهه ۹۰ میلادی تحقیقات روی آن شدت یافته است. این متأثر از دلایل متعددی است. اولین دلیل، افزایش تقاضاهای مردمی برای ارتباطات بی‌سیم بود [۳]. تا انتهای دهه اول قرن بیست و یکم تقاضاها عمدتاً به سیستم‌های تلفن همراه مربوط می‌شد. چنانچه در سال ۲۰۰۲ میلادی تعداد تلفن‌های همراه در سطح جهان از تعداد تلفن‌های با خط ثابت فراتر رفت. انتظار می‌رود که در آینده کاربردهای انتقال بی‌سیم داده (اطلاعات) از اهمیت بیشتری برخوردار شود. دلیل دوم توجه به سامانه‌های بی‌سیم، پیشرفت چشم‌گیر در تکنولوژی مجتمع‌سازی در مقیاس بسیار بزرگ[4] بود که امکان پیاده‌سازی الگوریتم‌های پردازش سیگنال‌های پیچیده را در ابعاد کم و با توان مصرفی کم فراهم کرد. نهایتاً دلیل سوم، موفقیت استانداردهای مخابرات بی‌سیم دیجیتال نسل دوم و خصوصاً استاندارد دسترسی چندگانه مبتنی‌بر اختصاص کد[5] بوده است. اما از آن‌جایی که سامانه‌های نسل دوم شبکه تلفن همراه اساساً برای انتقال صوت طراحی شده بودند و ویژگی‌های اصلی آن‌ها مانند نرخ مخابره و تأخیر زمانی قابل قبولشان برای کاربردهای صوتی تنظیم شده بود، نسل سوم شبکه تلفن همراه ظهور و توسعه یافت [۳]. نسل چهارم شبکه تلفن همراه، نرخ انتقال داده را ۱۰ تا ۱۰۰ مگابیت بر ثانیه برای

[1] Wireless Communication
[2] Electromagnetic
[3] Wired Communication
[4] Very Large Scale Integration
[5] Code Division Multiple Access(CDMA)

هر کاربر[1] فراهم می‌کند (شبکه‌های امروزی نسل سوم، امکان انتقال داده تا سرعت ۲ مگابیت بر ثانیه را در برخی از نقاط جهان فراهم کرده‌اند). نسل چهارم تلفن‌های همراه کاربردهای زیادی در تجارت الکترونیکی و تجارت همراه دارند [۴]. شکل ۱-۱ کاربردهای مختلف ارتباط بی‌سیم را نشان می‌دهد.

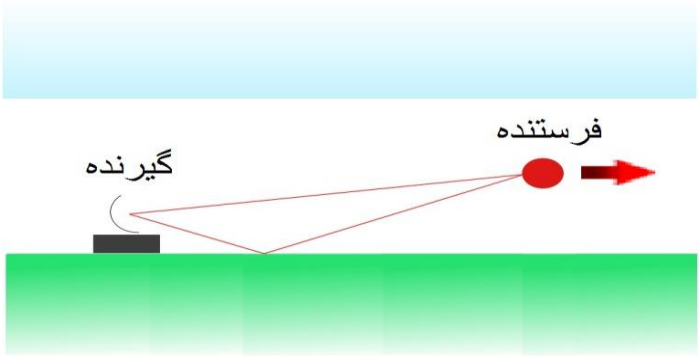
شکل ۱-۱) کاربردهای مختلف ارتباط بی‌سیم [۱]

بر خلاف مخابرات با سیم که هر جفت فرستنده و گیرنده به وسیله رابط‌های مجزا به هم متصل شده‌اند، در مخابرات بی‌سیم کاربران در هوا مخابره کرده و تداخل زیادی بین آن‌ها وجود دارد. این تداخل متشکل از تداخلی است که بین فرستنده‌هایی که با یک گیرنده خاص در ارتباط هستند، تداخلی که بین یک فرستنده خاص و چند گیرنده‌اش وجود دارد و تداخلی که بین جفت‌های مختلف از فرستنده و گیرنده وجود دارد [۳].

تفاوت‌های عمده مخابرات بی‌سیم و مخابرات با سیم وجود محوشدگی و تداخل است. این دو موضوع پژوهشگران را با چالش‌هایی روبرو کرده که در مخابرات با سیم وجود نداشته و بخش عمده‌ای از مطالعات به آن‌ها اختصاص داده شده است. پدیده محوشدن به تغییرات زمانی کیفیت کانال گفته می‌شود، که دلایل مختلفی دارد و در ادامه بیان می‌شود.

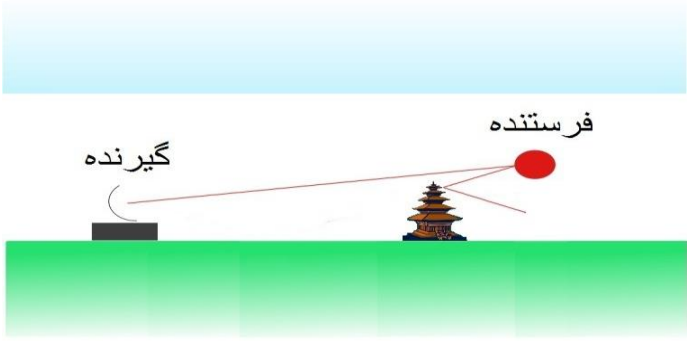
[1] User

حرکت فرستنده یا گیرنده یکی از دلایل محوشدگی است. حرکت فرستنده می‌تواند منجر به ضعیف شدن کیفیت مسیرها شده تا حدی که مسیر ارتباطی را عملا قطع کند. شکل ۱-۲ این موضوع را نشان می‌دهد.



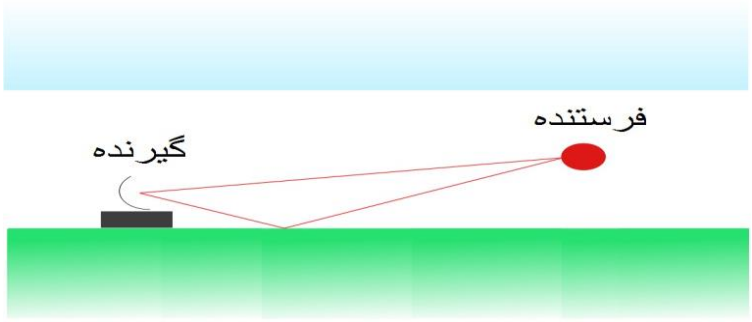
شکل ۱-۲) پدیده محوشدگی به دلیل حرکت فرستنده

قرار گرفتن یک مانع در مقابل یکی از مسیرها یکی دیگر از دلایل محوشدگی است. شکل ۱-۳ این موضوع را نشان می‌دهد.



شکل ۱-۳) پدیده محوشدگی به دلیل قرار گرفتن مانع در مسیر

همچنین وجود چند مسیر مختلف یکی دیگر از دلایل محوشدگی است، زیرا باعث می‌شود که یک سیگنال خاص با تاخیر و اختلاف فازهای متفاوت در گیرنده دریافت شده و این اختلاف فاز باعث بوجود آمدن تغییرات ناخواسته در دامنه سیگنال دریافتی شود. شکل ۱-۴ این موضوع را نشان می-دهد.



شکل۱-۴) پدیده محوشدگی به دلیل وجود چند مسیر مختلف

۱-۲ شبکه‌های بی‌سیم 5G و 6G

یک شبکه بی‌سیم یک شبکه کامپیوتری است که ارتباط بین نقاط (اجزاء) شبکه به روش بی‌سیم برقرار می‌شود و داده‌ها منتقل می‌شوند. با این‌که به نظر می‌رسد از نظر فنی عبارت شبکه بی‌سیم برای اشاره به هر نوع شبکه‌ای که بی‌سیم باشد به کار می‌رود، اما این اصطلاح بیشتر برای اشاره به شبکه‌های ارتباطی به کار می‌رود که در آن گره‌ها بدون استفاده از سیم به یکدیگر متصل می‌شوند. به عنوان مثال یک شبکه رایانه‌ای که نوعی از شبکه‌های ارتباطی است.

نمونه‌هایی از شبکه‌های بی‌سیم عبارت‌اند از: تلفن همراه، شبکه‌های محلی بی‌سیم[1]، شبکه‌های حسگر بی‌سیم، شبکه‌های ماهواره‌ای، شبکه‌های ارتباطی مایکروویو [۵].

شبکه‌های 5G و 6G نسل بعدی شبکه‌های اتصال به اینترنت هستند که نسبت به شبکه‌های قدیمی سرعت بالاتر و اتصال قابل اطمینان‌تری برای گوشی‌های هوشمند و سایر دستگاه‌ها ارائه می‌دهند و باعث افزایش چشم‌گیر قدرت تکنولوژی اینترنت اشیا می‌شوند [۶]. همچنین زیرساخت مورد نیاز برای انتقال مجموعه زیادی از داده‌ها را فراهم می‌کنند و برای دنیایی هوشمندتر با ارتباطات بیشتر تلاش می‌کنند. مشخصات این نسل‌ها سرعت انتقال داده بیشتر، کاهش تاخیر، صرفه جویی در انرژی است.

[1] Wireless Local Area Networks (WLAN)

برای دستیابی به چنین میزان داده و سرعت بالا، نسل‌های جدید سیستم‌های بی‌سیم باید در فرکانس‌های بالاتر کار کنند. اما استفاده از فرکانس‌های بالا مانند موج میلی‌متر و تراهرتز یکسری مشکلات دارند [۷-۹]. اولین مشکل این است که با افزایش فرکانس عمق نفوذ امواج کم می‌شود، یعنی با ایجاد یک مانع در محیط امواج قادر به عبور از آن نیستند و مسدود می‌شوند. مشکل دیگر محوشدگی است که دلایلش در بخش قبل بیان گردید. در ادامه با معرفی سطوح هوشمند، به ویژه فن‌آوری فرا سطوح نشان می‌دهیم که چگونه می‌توان این مشکل را رفع نماییم.

۱-۳ سطوح هوشمند با قابلیت پیکربندی مجدد

یک فرض رایج در تحلیل‌ها و مطالعات مربوط به شبکه‌های ارتباط بی‌سیم، رفتار غیر قابل کنترل محیط انتشار رادیویی برای بشر است. یعنی سیگنال رادیویی که از فرستنده منتشر می‌شود، در مسیر خود به موانع مختلف مانند ساختمان‌ها و درختان برخورد می‌کند و دچار پدیده‌های مختلفی مانند انعکاس، انکسار و پراکندگی می‌شود و نهایتا بعد از این‌که این پدیده‌های انتشاری تاثیر خود را روی سیگنال گذاشتند، سیگنال به گیرنده می‌رسد. این اثرات به شکل نوسانات شدید و یا افت در سیگنال ظاهر می‌شوند که به نام فیدینگ کانال شناخته می‌شوند. در سال‌های گذشته، محققین دنبال روش‌هایی برای شناسایی و مدل‌سازی این پدیده‌های انتشاری در کانال‌های بی‌سیم و طراحی الگوریتم‌ها، پردازش‌ها و سیستم‌های مناسب برای ایجاد ارتباط مطمئن روی این کانال‌ها بوده‌اند. یعنی افراد با فرض این‌که محیط در اختیار ما نیست، تلاش می‌کردند تا خود را با آن وفق دهند. این فرض امروزه نیز وجود دارد.

اما در سال‌های اخیر، مطالعات محققان نشان می‌دهد که با قرار دادن برخی سطوح هوشمند و قابل برنامه‌ریزی در مکان‌های مناسب در محیط انتشار، می‌توان کانال بی‌سیم را به صورت الکترونیکی کنترل کرد [۱۰]. به عنوان مثال اگر نمای یک ساختمان با توجه به مشخصات و جنس مواد خود، سیگنال را در یک جهت خاص منعکس کند، می‌توان با نصب تجهیزاتی روی این نما، خواص انعکاسی

آن را تغییر داد. این تغییر می‌تواند به صورت پویا[1] و متغیر با زمان انجام شود. به عبارت دیگر، از این پس، منعکس کننده‌ها در محیط ثابت[2] نیستند و می‌توانند توسط بشر کنترل شوند. این کنترل در راستای این است که کانال مخابراتی یک رفتار مطلوب داشته باشد و در نهایت عملکرد بهتری از سیستم‌های مخابرات بی‌سیم کنونی بدست آید.

سطوح هوشمند با قابلیت پیکربندی مجدد[3] که به نام سطوح انعکاسی هوشمند[4] نیز شناخته می‌شوند، سطوح الکترومغناطیسی هستند که بر روی دیوارها یا نمای ساختمان‌ها، سقف اتاق‌ها، سالن‌ها و یا سایرمکان‌ها در محیط نصب می‌شوند. این سطوح، در حقیقت منعکس کننده‌های هوشمند سیگنال‌های الکتریکی هستند که با جریان‌های ضعیف الکترونیکی کنترل می‌شوند. با کمک یک پردازش‌گر در پشت آن‌ها و با ارسال سیگنال‌های مناسب به این سطوح می‌توان آن‌ها را برنامه‌ریزی کرد. با این هدف که سیگنال‌های ارسالی به آن‌ها، در جهت‌های خاص انتشار یابند یا در جهت‌های دیگر منتشر نشوند. به بیان دیگر، هوشمندی این سطوح مربوط به همان پردازشگر مرکزی است. این پردازشگر با الگوریتم‌های مناسب، این سطوح را برنامه‌ریزی می‌کند تا مطابق هدف طراح سیستم عمل کنند.

۱-۴ فن‌آوری‌های پوشش برای محیط بی‌سیم قابل برنامه‌ریزی

رله‌ها [۱۱، ۱۲]، reflectarrays [۱۳-۱۶] و فرا سطوح[5] [۱۰، ۱۷-۲۰] محبوب‌ترین فن آوری‌های پوشش برای محیط بی‌سیم قابل برنامه‌ریزی هستند. رله‌ها حاوی مجموعه‌ای از آنتن‌های کم هزینه هستند که به طور منفعلانه، امواج الکترومغناطیسی را منعکس یا تضعیف می‌کنند و محیط را پیکربندی می‌کنند تا شبکه‌های بی‌سیم فعال در آن نزدیکی را بهبود بخشند. بدیهی است، این رویکرد

[1] Dynamic
[2] Static
[3] Reconfigurable Intelligent Surfaces (RIS)
[4] Intelligent Reflecting Surfaces (IRS)
[5] Metasurfaces

نمی‌تواند محیط بی‌سیم قابل برنامه‌ریزی را به صورت تطبیقی بازسازی کند. رویکرد دوم، reflectarrays شامل تعدادی از آنتن‌ها در یک آرایش شبکه ۲ بعدی هستند. علاوه بر این، برخی از عناصر فعال مانند پین دیودها برای تغییر فاز موج بازتاب شده از موج الکترومغناطیس را دارند. قابلیت‌های reflectarrays، هدایت و جذب موج است، این فن آوری در نقاط دور قابل دستیابی است. رویکرد دیگر استفاده از فرا سطوح است که شبیه به reflectarrays، اما با چگالی بیشتر هستند. با داشتن چگالی به اندازه کافی، فرا سطوح، هر نوع امواج الکترومغناطیسی را حتی در میدان نزدیک تولید می‌کنند [۲۱]. در این پایان‌نامه از رویکرد فرا سطوح مبتنی‌بر نرم افزار [۱۷] استفاده می‌شود. در زیر بخش ۱-۴-۱ جزئیات این رویکرد معرفی شده است.

۱-۴-۱ فرا سطوح

این سطوح، سطوح انعکاسی (آنتن‌های انعکاسی) هستند که از یک آرایه از المان‌های انعکاسی (به نام متا اتم) ساخته شده‌اند. به کل آرایه نیز فرا سطح می‌گویند. خواص بازتابی این سطوح توسط ابزارهای الکترونیکی مانند پین دیودها یا ورکتورها قابل کنترل است. به بیان دقیق‌تر، با نصب این سطوح در محیط و تغییر مشخصات انعکاسی المان‌های آن، می‌توان کاری کرد که برآیند اثر آن‌ها و اثرات انتشاری محیط، منجر به یک رفتار مطلوب و مد نظر در کانال بی‌سیم شوند [۲۲].

۱-۴-۱-۱ مزیت‌های فرا سطوح

برخی از مزیت‌هایی که با استفاده از سطوح هوشمند در محیط می‌توان انتظار داشت عبارت‌اند از [۲۳، ۲۴]:

- بهبود پوشش شبکه در نقاطی که پوشش مناسب ندارند (نقاط کور)
- بهبود نرخ انتقال اطلاعات

- کاهش مصرف توان در شبکه
- کاهش تداخل
- افزایش امنیت شبکه

۱-۵ هدف پایان‌نامه

در روش‌های موجود در گذشته [۱۰، ۱۷]، بیشتر مطالعات مربوط به طراحی کاشی‌های محیط می‌باشد و کنترلی روی رفتار امواج الکترومغناطیسی نیست. همچنین مواردی مانند ظهور موانع در محیط و همچنین عملکرد مطلوب، زمانی که تعداد کاربران افزایش می‌یابند وجود ندارد. هدف این پایان‌نامه، رفع این مشکلات و همچنین بالابردن میانگین توان دریافتی برای تمام کاربران موجود در محیط است.

۱-۶ چالش‌ها و راه‌حل‌های آن‌ها

یکی از چالش‌های مهم در طراحی محیط‌های بی‌سیم، کنترل روی رفتار امواج الکترومغناطیسی است. تاکنون تحقیقاتی برای حل این چالش با استفاده از مسائل بهینه‌سازی[1] [۱۷] و شبکه عصبی [۱۰] انجام شده است که در فصل دوم به آن‌ها می‌پردازیم. با این حال این روش‌ها محدودیت‌هایی دارند. در مسئله بهینه‌سازی، اگر در محیط یک مانع ایجاد شود، سیگنال بین فرستنده و گیرنده قطع می‌شود و محیط نمی‌تواند خودش را سریع تطبیق دهد. همچنین تکنیک استفاده از شبکه عصبی خود دارای مشکلاتی از قبیل محدودیت تعداد کاربران و نیز پیچیدگی‌های فاز آموزش است.

بنابراین، تطبیق‌پذیری عملکرد کاشی‌ها یک چالش بزرگ دیگر در یک محیط مبتنی‌بر SDM است، به ویژه هنگامی که تعداد و موقعیت کاربران می‌تواند دائما در حال تغییر باشد. به عنوان مثال، یک مانع می‌تواند به طور ناگهانی امواج الکترومغناطیسی هدایت شده را مسدود کند و بنابراین برخی از لینک‌های بی‌سیم با شکست مواجه می‌شوند. رویکردهای به کار گرفته شده برای کنترل ارتباط بی-

[1] Optimization

سیم در PWE از جمله حل مسئله بهینه‌سازی یا تکنیک مبتنی‌بر شبکه عصبی نمی‌توانند بطور دینامیکی عملکرد کاشی‌ها را به محض تغییر محیط بازگردانند. در این پایان‌نامه برای رفع این چالش از الگوریتم یادگیری تقویتی[1] استفاده شده است. با استفاده از این الگوریتم، سرور آموزش می‌بیند که با توجه به موقعیت هر کاربر و وضعیت محیط، عملکرد کاشی‌ها را انتخاب کند.

با استفاده از الگوریتم یادگیری تقویتی، کاشی‌ها، در حالی که امواج الکترومغناطیسی به تدریج با محیط در تعامل هستند، می‌توانند تصمیم بهتری بگیرند. بنابراین، هنگامی که تغییری در محیط رخ می‌دهد، به عنوان مثال یک مانع برخی از امواج الکترومغناطیسی را مسدود می‌کند، عامل مجازات زیادی دریافت می‌کند و بنابراین یک عمل جدید انتخاب می‌شود.

یک چالش دیگر این است که با افزایش تعداد کاربران در محیط، حداقل توان دریافتی برای آن‌ها، حداکثر شود و به یک حد قابل قبول برسد و ارتباط مناسب بین فرستنده و گیرنده‌ها فراهم شود. با استفاده از روش پیشنهادی الگوریتم یادگیری تقویتی خواهیم دید که میانگین توان دریافتی برای کاربران در مقایسه با کار گذشته که از الگوریتم ژنتیک (GA)[2] استفاده کرده است [۱۷]، افزایش می‌یابد. به این صورت که در مرجع [۱۷]، میانگین توان دریافتی برای ۱۲ گیرنده، ۲۵.۳۸ dBmW می‌باشد و روش پیشنهادی در این پایان‌نامه، میانگین توان دریافتی را ۱۲ درصد افزایش داده است.

۱-۷ نوآوری

در این پایان نامه یک روش پیشنهادی مبتنی‌بر یادگیری تقویتی را ارائه می‌کنیم که می‌تواند برای محیط‌های شامل چندین کاربر استفاده شود. علاوه بر این، هنگامی که محیط تغییر می‌کند، به عنوان مثال، یک مانع، مسیر برخی از امواج الکترومغناطیسی را مسدود می‌کند، پس از بررسی مجدد، امواج مسدود شده الکترومغناطیسی می‌توانند بلافاصله مسیری جدید را پیدا کنند.

[1] Reinforcement Learning (RL)
[2] Genetic Algorithm

در این مدل یادگیری، امواج الکترومغناطیسی منتقل شده نقش عامل‌ها[1]، کاشی‌ها نقش حالت‌ها[2] را ایفا می‌کنند و عملکرد متفاوت کاشی‌ها به عنوان اقدامات[3] معتبر در نظر گرفته می‌شوند. با استفاده از الگوریتم یادگیری تقویتی، کاشی‌ها در حالی که امواج الکترومغناطیسی به تدریج با محیط در تعامل هستند، می‌توانند تصمیم بهتری بگیرند. بنابراین، هنگامی که تغییری در محیط رخ می‌دهد، به عنوان مثال یک مانع برخی از امواج الکترومغناطیسی را مسدود می‌کند، عامل مجازات زیادی دریافت می‌کند و بنابراین یک عمل جدید انتخاب می‌شود. از آن‌جا که معمولاً بیش از یک کاربر در محیط وجود دارد، در روش پیشنهادی قصد داریم محیط ارتباطات بی‌سیم را به عنوان یک مسئله یادگیری تقویتی چند عامله[4] مدل نماییم.

۱-۷ ساختار پایان‌نامه

در ادامه این تحقیق، در بخش دوم، مروری بر روش‌های پیشین انجام شده است. در بخش سوم، روش پیشنهادی مبتنی‌بر یادگیری تقویتی ارائه شده است. در بخش چهارم، نتایج آزمایش‌ها مورد بررسی قرار گرفته است و در بخش پایانی، جمع‌بندی و نتیجه‌گیری ارائه شده است.

[1] Agent
[2] State
[3] Action
[4] Multi Agent Reinforcement Learning (MALR)

فصل ۲: پیشینه تحقیق

۲-۱ مقدمه

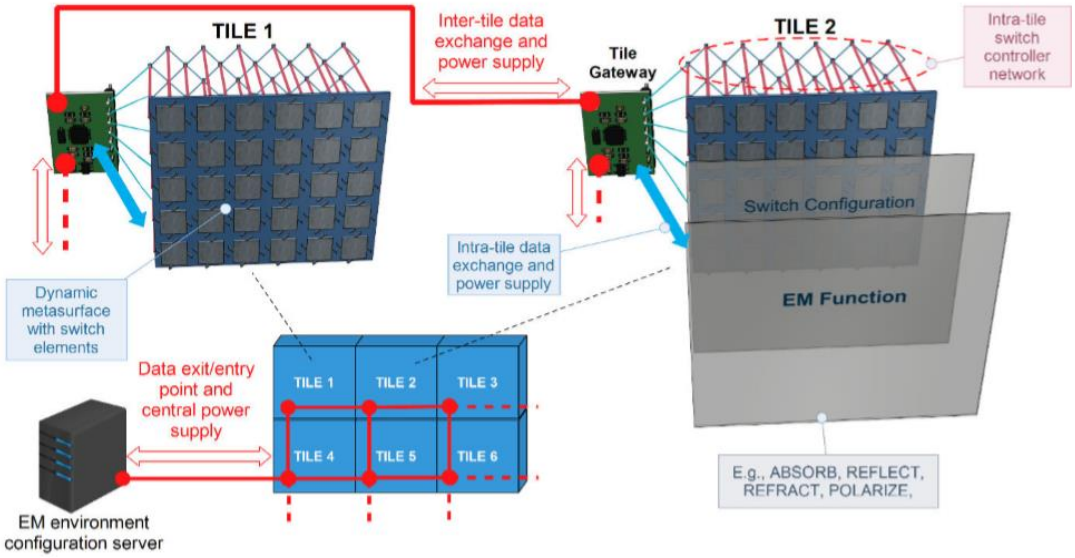
با توجه به پیشرفت روش‌های در نظر گرفته شده برای شبکه‌های بی سیم 6G آینده، محققان به مطالعه در مورد توانایی‌ها و محدودیت‌های استفاده از رویکردهای هوش مصنوعی،[1] در بهینه‌سازی شبکه‌های بی‌سیم مجهز به SDM، پرداخته‌اند [۲۵]. SDMها برای کنترل رفتار امواج الکترومغناطیسی و پشتیبانی از آن‌ها در زمان استفاده، به یکپارچه‌سازی ساختار خود با یک کنترل-کننده شبکه نیاز دارند [۲۶] و دستورات با فعال یا غیرفعال کردن سوئیچ‌های مرتبط در کنترل‌کننده-ها انجام می‌شوند. با توجه به این موارد، SDM برای عملکرد مطلوب خود، علاوه بر اطمینان از مقیاس پذیری[2]، کاهش سربار[3] و انرژی، به سطح هماهنگی پیچیده‌ای نیاز دارند. همچنین نه تنها تعداد زیادی پارامتر در محیط وجود دارد که باید بهینه شوند، بلکه داده‌های مورد نیاز برای بهینه‌سازی محیط بسیار زیاد است. از این‌رو برای بهینه‌سازی شبکه‌ها استفاده از رویکردهای هوش مصنوعی در محیط‌های بی‌سیم پیشنهاد شده است، که در بخش ۲-۳ و ۲-۴ به آن‌ها می‌پردازیم. برای این منظور ابتدا لازم است در بخش ۲-۲ ساختار کاشی‌های مبتنی‌بر SDM و همچنین شبکه بین آن‌ها را مورد بررسی قرار دهیم.

۲-۲ ساختار کاشی‌های مبتنی‌بر SDM

ایده ارتباطات مبتنی‌بر SDM این است که اشیاء مسطح را با واحدهایی از ابعاد کنترل شده نرم افزاری تحت عنوان کاشی‌های SDM تحت پوشش قرار دهند [۱۷]. همانطور که در شکل ۲-۱ مشاهده می-شود، کاشی‌ها از مجموعه‌ای از عناصر سوئیچ[4] کنترل شده توسط مجموعه‌ای از کنترل کننده‌های شبکه، و یک دروازه[5] ساخته شده است. با استفاده از دروازه، شبکه کنترل دستورات را از سرور[6]

[1] Artificial intelligence
[2] Scalability
[3] Overhead
[4] Switch
[5] Gateway
[6] Server

دریافت می‌کند تا وضعیت فعلی سوئیچ‌ها را تغییر دهد، بنابراین کاشی‌ها را برای به دست آوردن کارکرد مطلوب آماده می‌کند. علاوه بر این، این دروازه نقش یک پل را برای تأمین توان مورد نیاز کاشی‌ها دارد. مطابق شکل ۲-۱، کاشی‌ها شبکه‌ای با توپولوژی توری[1] را تشکیل می‌دهند که در آن حداقل یکی از دروازه‌ها متصل به سرور پیکربندی محیط، برای جمع‌آوری داده‌های حس شده محیطی، کشف وسایل ارتباطی در محیط و انتشار دستورهای مناسب در شبکه کاشی است. این دستورات حداقل نوع عمل مانند فرمان یا جذب و آدرس دروازه کاشی مورد نظر را دارند. شایان ذکر است که ترجمه یک عمل به یک عنصر سوئیچ کاشی، معمولاً با استفاده از یک جدول جستجوی تهیه شده در طی فرآیند طراحی / ساخت کاشی انجام می‌شود [۱۷].



شکل ۲-۱) اجزا یک کاشی SDM [۱۷]

[1] Grid

۲-۳ یادگیری ماشین (ML[۱])

روش‌های یادگیری ماشین مانند طبقه‌بندی[۲]، بهینه‌سازی، تخمین[۳] و تشخیص[۴]، گزینه‌های مناسب و امیدوارکننده‌ای برای هوشمند کردن SDMها هستند که می‌توانند راه حل‌های عملی برای مشکلات ناشی از ارتباطات بی‌سیم ارائه دهند.

الگوریتم‌های یادگیری ماشین به سه دسته اصلی تقسیم می‌شوند: یادگیری تحت نظارت[۵]، یادگیری بدون نظارت[۶] و یادگیری تقویتی. دو دسته متداول الگوریتم‌های یادگیری ماشین، یعنی الگوریتم‌های یادگیری تحت نظارت و بدون نظارت، به شرح زیر است:

• یادگیری تحت نظارت: در این روش، داده‌های ورودی (مجموعه آموزش) و خروجی (برچسب‌ها[۷]) برای یادگیری الگوریتم‌ها ضروری هستند. هدف سیستم یادگیر به دست آوردن فرضیه‌ای است که تابع یا رابطه بین ورودی یا خروجی را حدس بزند [۲۷].

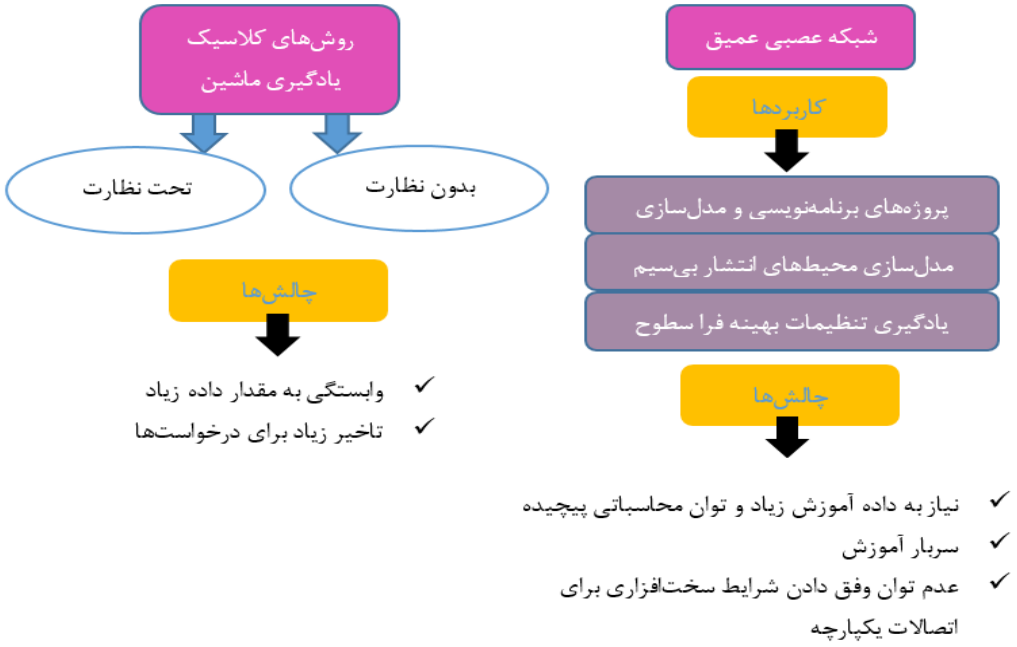
• یادگیری بدون نظارت: در این روش، یادگیری بر روی داده‌های بدون برچسب است و الگوریتم‌ها باید بتوانند ساختار الگوهای پنهان داده‌های ورودی را بیابند.

یک زیرمجموعه جداگانه از یادگیری ماشین که به طور گسترده مورد بحث قرار می‌گیرد، یادگیری عمیق (DL[۸]) است. یادگیری عمیق به عنوان یک الگوریتم ترکیبی شامل تکنیک‌های هر یک از سه الگوریتم فوق‌الذکر در نظر گرفته می‌شود، که نقاط قوت آن‌ها را به ارث می‌برد و نقاط ضعف آن‌ها را به حداقل می‌رساند.

---

[1] Machine learning
[2] Classification
[3] Estimation
[4] Detection
[5] Supervised learning
[6] Unsupervised learning
[7] Labels
[8] Deep learning

در سال‌های اخیر، الگوریتم‌های کلاسیک یادگیری ماشین، مانند درخت تصمیم[1]، مدل‌های مخلوط گاوسی[2]، k-نزدیکترین همسایه[3]، مدل‌های ماشین بردار پشتیبان[4]، مدل‌های برنامه‌ریزی منطقی مبتنی‌بر قاعده و استقرا ثابت کرده‌اند که در کمک و افزایش کارایی و طراحی سیستم‌های ارتباطی بی‌سیم سودمند هستند [۲۷]. قابلیت‌ها و چالش‌های یادگیری ماشین در شکل ۲-۲ خلاصه شده است.



شکل ۲-۲) قابلیت‌ها و چالش‌های یادگیری ماشین [۲۵]

۲-۴ یادگیری عمیق

در مقایسه با روش‌های یادگیری که مطرح شد، الگوریتم‌های یادگیری عمیق در چند سال گذشته محبوبیت قابل توجهی به ویژه در حوزه ارتباطات بی‌سیم به دست آورده‌اند. توانایی رویکردهای یادگیری عمیق برای یادگیری و نشان دادن یک ساختار همبستگی در داده‌های موجود با یا بدون دانش قبلی، یک مزیت نسبت به رویکردهای سنتی است. این کار از طریق یک معماری شبکه عصبی

[1] Decision Tree
[2] Gaussian mixture models
[3] K-nearest neighbour
[4] Support vector machine

با چندین لایه پنهان که قادر به کشف ساختارهای نهفته در داده‌های دارای برچسب و بدون برچسب است، حاصل می‌شود. از یادگیری عمیق در بین تمام لایه‌ها در ارتباطات بی‌سیم، از فیزیکی تا توسعه و برنامه ریزی شبکه، استفاده می‌شود [۲۸]. در لایه فیزیکی، روش‌های DL برای ساده‌سازی کارهای مختلف از جمله تخمین و رمزگشایی[۱] کانال، یکسان‌سازی و همگام‌سازی در سیستم‌های OFDM[۲] و محلی سازی[۳] استفاده شده است. همچنین قابلیت‌های یادگیری عمیق برای بهینه‌سازی منابع شبکه، مسیریابی[۴] و تنظیم مجدد آنتن مورد مطالعه قرار گرفته است. با این حال، این سوال مطرح می‌شود که آیا روش‌های مبتنی‌بر یادگیری عمیق توانایی مقابله با نیازهای سخت‌گیرانه ناشی از محیط رادیویی هوشمند پویا را دارند؟

با توجه به ماهیت پیچیده در SDM، به ویژه برای محیط‌های داخلی، روش‌های یادگیری عمیق برای تمرکز سیگنال از طریق یادگیری نگاشت بین موقعیت کاربر و تنظیمات بهینه فرا سطوح مناسب هستند [۲۹]. برای یادگیری بهینه‌سازی فرا سطوح، پس از برخورد امواج الکترومغناطیسی به آن‌ها با توجه به این‌که اطلاعات قبلی در مورد محیط انتشار در دسترس است، از روش‌های مختلف یادگیری عمیق استفاده شده است. مطالعات اخیر [۳۰] استفاده از روش‌های یادگیری عمیق را برای حمایت از توسعه روش‌های جدید پیشنهاد می‌کند که می‌توانند با محیط رادیویی هوشمند سازگاری بهتری داشته باشند. در [۳۱] نشان داده شده است که شبکه عصبی پیشرو[۵] می تواند به طور بالقوه محیط انتشار بی‌سیم اطراف SDMها را مدل‌سازی کند. تراکم[۶]، اطلاعات مکان فرستنده‌ها و گیرنده‌ها، سطح سر و صدا[۷]، ابعاد کاشی‌ها، به عنوان پارامترهای ورودی در مدل استفاده می‌شوند و عملکرد به صورت ترکیبی از توان سیگنال دریافت شده و زاویه ورود سیگنال‌ها مدل‌سازی می‌شود. هدف یادگیری یک مدل انتشار است، به طوری که هر کاشی SDM به طور مناسب تنظیم شود. بنابراین هوش کمکی

[1] Decoding
[2] Orthogonal Frequency-Division Multiplexing
[3] Localization
[4] Routing
[5] Feed-forward neural network
[6] Densities
[7] Noise levels

یادگیری عمیق می‌تواند محرک اصلی در تحقق عملکردهای پیش‌بینی شده توسط SDM مجهز به هوش مصنوعی باشد، اما محیط رادیویی هوشمند با شبکه‌های موجود تفاوت اساسی دارد و سطح بالایی از چابکی و انطباق‌پذیری شبکه را می‌طلبد که هر کدام در قسمت زیر شرح داده می‌شود.

قابلیت چابکی شبکه[1]: استفاده از روش‌های یادگیری عمیق به دلیل مشکلاتی که دارد می‌تواند چابکی شبکه را تحت تاثیر قرار بدهد، مانند نیاز به داده‌های آموزش زیاد، توان محاسباتی پیچیده، سربار آموزش و فقدان روش‌های مناسب برای بهینه‌سازی پارامترها و تعداد زیاد پارامترها در زمان مدل‌سازی محیط بی‌سیم. این مشکلات آن‌ها را برای SDM غیرقابل استفاده می‌کند. پیشرفت‌های اخیر در واحدهای پردازش گرافیکی (GPU)[2] همچنین امکاناتی را برای تسریع در محاسبه این مدل-ها، کاهش زمان و رفع پیچیدگی فرایند یادگیری فراهم کرده است. برای حفظ چابکی شبکه و برای سازگاری یادگیری عمیق با شبکه‌های بی‌سیم آینده، استدلال شده است که باید از استراتژی‌های یادگیری توزیعی[3] یا روی دستگاه[4] استفاده شود [۳۰]، زیرا محیط محاسبات توزیع شده برای آموزش و استنباط راه حل‌های مبتنی‌بر یادگیری عمیق می‌توانند به فرایندهای بسیار سریع، کم مصرف و کم هزینه برای طراحی شبکه SDM کمک کند. به عنوان مثال می‌توان از تکنیک‌های یادگیری فدرال[5] استفاده کرد که بر اساس یک اصل توزیع داده‌ها و وظایف محاسباتی بین فدرال منابع محلی است[6] و توسط یک سرور مرکزی اداره می‌شود. این روش می‌تواند کمک کند تا یک استراتژی آموزش توزیع شده کارآمد برای الگوریتم‌های یادگیری عمیق طراحی شود. به این صورت که هر منبع، داده‌ها را برای یادگیری یک مدل یادگیری عمیق محلی پردازش می‌کند، در حالی که یک سرور مرکزی با تلفیق مدل‌های محلی، یک مدل جهانی[7] را فرا می‌گیرد. برای سرعت بخشیدن به ایجاد مدل جهانی و

[1] Network Agility
[2] Graphics Processing Unit
[3] Distributed learning
[4] On-device
[5] Federated learning
[6] Local
[7] Global

کاهش سربار ارتباطات در بین منابع، پیشنهاد شده است که پارامترهای به‌روز شده فقط در یک سرور مرکزی به اشتراک گذاشته شود.

انطباق‌پذیری شبکه[1]: محیط‌های رادیویی هوشمند یک محیط تصادفی هستند و همچنین درخواست‌های خدمات، متشکل از انواع مختلف موقعیت‌های فرستنده‌ها و گیرنده‌ها را دارند. یک مسئله مهم این است که چگونه با پیچیدگی محیط‌های SDM کنار بیاییم و سازگار شویم. با توجه به محیط رادیویی هوشمند که داده‌های سنجش شده بسیار زیاد و محیط پویا است، تنها اتکا به تکنیک-های بهینه‌سازی مبتنی‌بر مدل یا داده محور امکان پذیر نیست. یادگیری انتقال[2] یک تکنیک جدید است که می‌تواند به ما کمک کند تا با امکان انتقال دانش به کار رفته در یک زمینه خاص، این چالش را برطرف کرده و در زمینه دیگری برای اجرای یک کار متفاوت استفاده شود [۲۹]. یکی از مزایای بارز استفاده از این روش کاهش داده‌های آموزش است. همچنین این یادگیری با تکنیک‌های یادگیری عمیق ترکیب می‌شود، که به عنوان یادگیری انتقال عمیق[3] شناخته می‌شود. دلیل استفاده از این رویکرد این است که دانش در یک زمینه می‌تواند در حوزه دیگری مورد بهره‌برداری قرار گیرد، بدون این‌که نیازی به بازسازی یا آموزش مجدد مدل از ابتدا باشد. این روش برای موفقیت در رابطه با همبستگی داده‌ها با مشکل تخصیص منابع استفاده شده است. کاربرد روش های یادگیری انتقال عمیق در [۳۰] بحث شده است.

با توجه به موارد گفته شده، در این فصل، به طور خاص، به سه کار مرتبط با روش‌های پیاده‌سازی محیط‌های بی‌سیم قابل برنامه‌ریزی مبتنی‌بر SDM[4] شامل بهینه‌سازی با الگوریتم ژنتیک و شبکه عصبی می‌پردازیم.

[1] Network Adaptability
[2] Transfer learning
[3] Deep transfer learning
[4] Software-Defined Metasurface

۲-۵ کارهای مرتبط

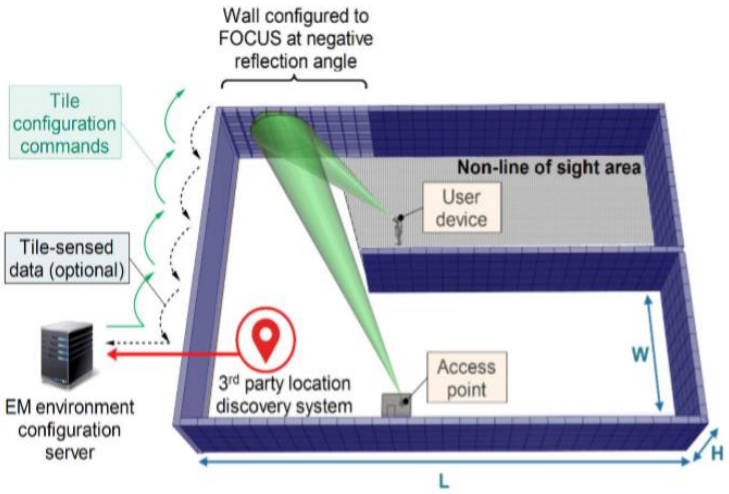
با توجه به توسعه تکنیک‌های کنترل رفتار امواج الکترومغناطیسی، محققان به محیط‌های بی‌سیم قابل برنامه‌ریزی مبتنی‌بر SDM، به خصوص برای تحقق نسل‌های بعدی شبکه‌های بی‌سیم جذب شده‌اند. بیشترین مطالعات مربوط به طراحی واحد کاشی PWE[1] است [۱۸، ۱۹، ۳۲-۳۴] و کمتر بر روی رویکردهای تطبیقی آن تمرکز شده است.

در مرجع [۱۷] یک روش ارتباطی نوین برای افزایش ظرفیت و امنیت با استفاده از محیط‌های بی-سیم قابل برنامه‌ریزی در سیستم‌های بی‌سیم نسل بعدی ارائه دادند و ایده استفاده از SDMها را برای اولین بار برای داشتن محیط بی‌سیم قابل برنامه‌ریزی، به ویژه برای نسل‌های بعدی شبکه‌های بی‌سیم معرفی کردند. نوآوری اصلی این روش جدید، مفهوم فرا سطوح است، که دارای ساختارهای مسطح هوشمند هستند که تاثیرات آن بر امواج الکترومغناطیسی برخوردی به طور کامل توسط ساختار آن‌ها تعریف می‌شود. ساخت محیط بی‌سیم به طور کامل از این طریق کنترل می‌شود و بهینه‌سازی عوامل انتشار اصلی بین دستگاه‌های بی‌سیم را ممکن می‌سازد. بنابراین، اثراتی مانند از دست دادن مسیر، محوشدگی چند مسیره و تداخل قابل کنترل است و امکان قابلیت‌های جدید در عملکرد و امنیت لایه فیزیکی را فراهم می‌آورد. به منظور برقراری ارتباط بین فرستنده و گیرنده، کاشی‌ها باید به طور تطبیقی انتخاب شوند و به طور بهینه کنترل شوند، تا به گیرنده‌های مورد نظر سیگنال کافی برسد. از آن‌جا که در سناریوهای ارتباطی دنیای واقعی، چندین کاربر می‌توانند در همان فضا حضور داشته باشند، لازم است که در مورد توزیع کاشی و الگوریتم کنترلی بحث شود.

مطابق شکل ۲-۳ آن‌ها از حالتی شروع کردند که در آن یک جفت فرستنده و گیرنده در محیط وجود دارد. هنگامی که فرستنده سیگنال را ارسال می‌کند، کاشی‌ها می‌توانند سیگنال‌های منتقل شده را حس کنند [۳۵]. با توجه به محل دریافت گیرنده که توسط سیستم کشف مکان شناخته می-

[1] Programmable Wireless Environment

شود، این کاشی‌ها زاویه‌های تابش خود را برای ایجاد مسیر بین فرستنده و گیرنده تنظیم می‌کنند. بنابراین، سیگنال دریافت شده در گیرنده، مجموع تمام سیگنال‌های بازتابی از کاشی‌های مختلف است. در میان همه کاشی‌ها، تنها آن‌هایی که می‌توانند سیگنال‌های منتقل‌شده را حس کنند باید به درخواست‌های ارسال پاسخ دهند و زوایا را تنظیم کنند. از سوی دیگر، در یک مورد پیچیده‌تر که در آن چند کاربر در یک محیط وجود دارند، توزیع کاشی باید بهینه‌سازی شود.



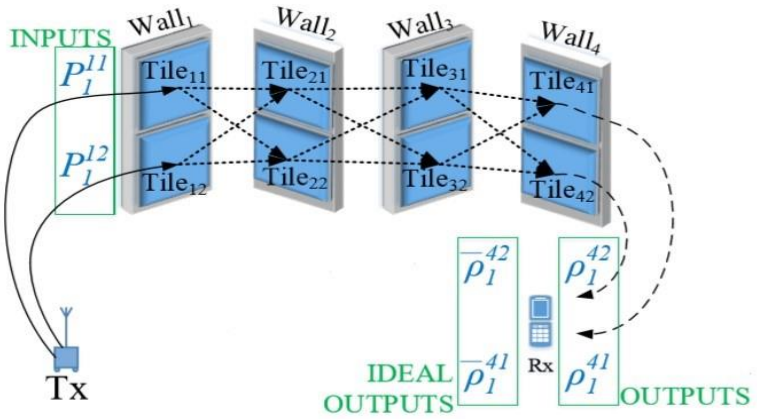
شکل ۲-۳) ارتباط بین فرستنده و گیرنده در یک محیط داخلی [۱۷]

آن‌ها با در نظر گرفتن ۱۲ گیرنده، پوشش را در یک منطقه دید غیر مستقیم[1](NLOS) فراهم کرده و با حل یک مسئله بهینه‌سازی با استفاده از تکنیک الگوریتم ژنتیک، حالت کاشی‌ها را به صورت مناسب تنظیم کردند. در این روش، هدف یافتن تنظیمات کاشی‌های مختلف به نحوی است که حداقل توان دریافتی بر روی ۱۲ گیرنده در منطقه NLOS حداکثر شود و همچنین حداکثر تاخیر توان دریافتی برای ۱۲ گیرنده حداقل شود. اگرچه این رویکرد منجر به پوشش کافی در منطقه NLOS می‌شود، اما به محض تغییر محیط نمی‌تواند وضعیت کاشی‌ها را به صورت تطبیقی تغییر دهد. به عنوان مثال، هنگامی که یک شی به طور ناگهانی ظهور می‌کند و برخی از لینک‌های بی‌سیم را مسدود می‌کند، باید یک مسئله بهینه‌سازی جدید حل شود که زمان‌بر می‌باشد.

[1] Non line of sight

در کار دیگر [۱۰]، رویکردی مبتنی‌بر الگوریتم‌های یادگیری ماشین، به ویـژه شـبکه‌هـای عصـبی ارائه شده است تا بتوان به صورت تطبیقی محیط‌های بی‌سیم قابل برنامـه‌ریـزی و کاشـی‌هـا را بـرای مجموعه‌ای از کاربران تنظیم کرد.

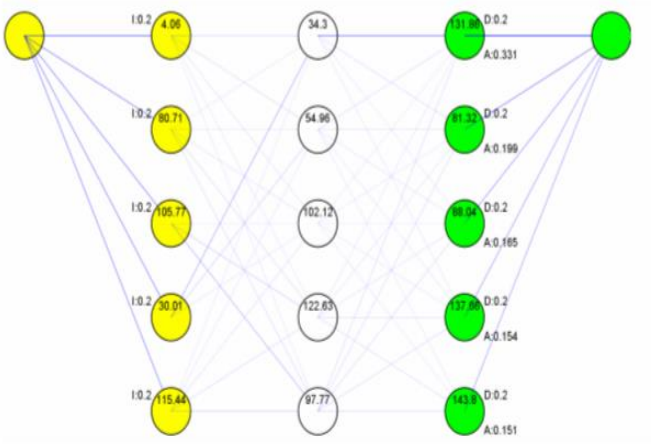
نویسندگان این مرجع مسئله انتشار بی‌سیم را به عنوان یک شبکه عصبی مطابق شکل ۲-۴ مـدل کردند که در آن دیوارها به عنوان لایه‌های شبکه، کاشی‌ها به عنوان نورون و برهم کنش متقابل آن‌ها به عنوان ارتباط مدل می‌شوند. در شبکه عصبی برای تولید یک مدل مورد نظر، فرآیند تعیین وزن‌هـا در هر لایه از اهمیت ویژه‌ای برخوردار اسـت. از میـان روش‌هـای مختلـف، روش برگشـت بـه عقـب[1] رایج‌ترین روش برای بهینه‌سازی وزن است. ایده برگشت به عقب عبارت است از تنظیم وزن نورون‌هـا براساس خطاهای لایه خروجی. پس از یک دوره آموزشی، شبکه اصول انتشار را یاد می‌گیرد و آن‌ها را پیکره‌بندی می‌کند تا ارتباط کاربران در نزدیکی آن‌ها تسهیل شود.



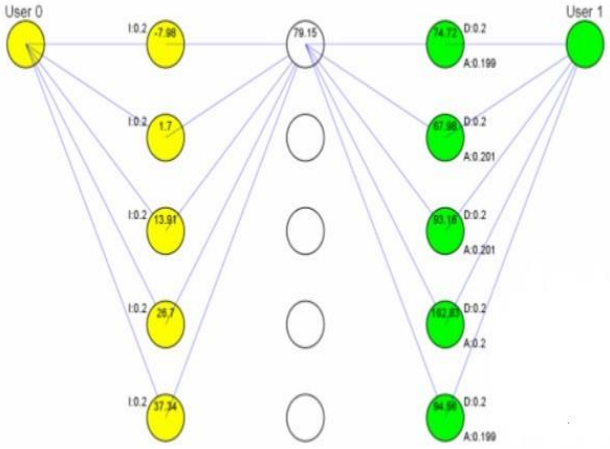
شکل ۲-۴) نمایش انتشار بی‌سیم به عنوان یک شبکه عصبی [۱۰]

قابل ذکر است که مطابق شکل ۲-۵ شبکه عصبی فاقد آموزش از تمام کاشی‌های لایه میانی اسـتفاده می‌کند و در صورتی که شبکه عصبی آموزش دیده یک راه‌حل پیدا می‌کند که در آن تنها یک کاشـی در لایه میانی فعال می‌شود. به عبارت دیگر، یکی از آن‌ها بخش قابل‌توجهی از کـل تـوان را مـدیریت می‌کند.

[1] Backpropagation



شکل ۲-۵: الف) شبکه عصبی فاقد آموزش [۱۰]

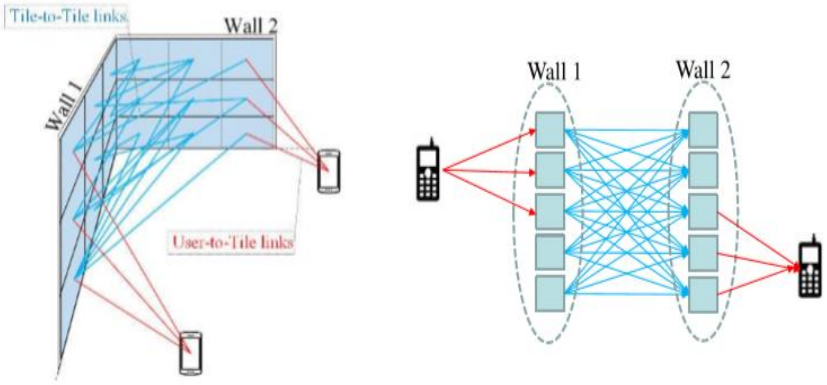


شکل ۲-۵: ب) شبکه عصبی آموزش دیده شده [۱۰]

در این تحقیق تنها یک کاربر در نظر گرفته شده است و مسئله با حضور چند کاربر مورد بحث قرار نگرفته است. همچنین تکنیک استفاده از شبکه عصبی خود دارای مشکلاتی مانند پیچیدگی‌های فاز آموزش می‌باشد.

در یک تحقیق دیگر [۲۰]، یک راه‌حل لایه به لایه شبکه برای ایجاد محیط‌های بی‌سیم قابل برنامه‌ریزی برای چندین کاربر ارائه شده است. اهداف دنبال شده شامل ترکیبی از کیفیت خدمات و بهینه‌سازی انتقال توان، کاهش استراق سمع و کاهش تداخل سیگنال است. در واقع برخلاف آثار فوق، این مرجع یک راه‌حل لایه به لایه شبکه برای تنظیم PWE برای چندین کاربر و اهداف چندگانه ارائه

داده است. علاوه بر این، یک الگوی گرافیکی از محیط‌های قابل برنامه‌ریزی ارائه شده است که به طور مؤثر نگرانی‌های فیزیکی و شبکه‌ای را از هم جدا می‌کند.



شکل ۲-۶) تصویر سه بعدی و دو بعدی از انواع ارتباطات (ارتباطات کاربری و ارتباطات بین کاشی‌ها) و نورون‌ها (کاشی و کاربران) در یک PWE [۲۰]

به عبارت دیگر مدل PWE مطابق شکل ۲-۶ به عنوان یک نمودار نشان داده شده است، که نورون‌ها کاشی‌های روی دیوارها هستند و اهداف ارتباطی به عنوان مشکل یافتن مسیر توصیف شده است. با این وجود، رویکرد پیشنهادی در این مرجع برای تغییر وضعیت حالت کاشی‌ها هنگام تغییر محیط بسیار زمان‌بر است.

۲-۶ نتیجه‌گیری

ما در این پایان نامه، یک محیط ارتباطی بی‌سیم شبیه به آن‌چه در مرجع [۱۷] طراحی شده است، در نظر می‌گیریم و با استفاده از الگوریتم یادگیری تقویتی، عملکرد کاشی‌ها را به نحو مناسب انتخاب می‌کنیم. از آن‌جا که معمولاً بیش از یک کاربر در محیط وجود دارد، ما محیط ارتباطات بی‌سیم را به عنوان یک مسئله یادگیری تقویتی چند عامله مدل می‌کنیم. سناریوهای مختلفی مانند ارائه پوشش دائمی برای هر کاربر، برقراری ارتباط در حضور موانع و در نهایت، فراهم کردن اتصال پس از ظهور

ناگهانی برخی موانع را در نظر می‌گیریم و با کارهای گذشته [۱۰، ۱۷] مقایسه می‌نماییم. برای این منظور، در فصل بعد، به بررسی الگوریتم پیشنهادی می‌پردازیم.

فصل ۳: روش پیشنهادی

۳-۱ مقدمه

در این فصل، ابتدا جزئیات مربوط به الگوریتم یادگیری تقویتی شرح داده می‌شود و سپس با توجه به این‌که معمول است بیش از یک کاربر با نقطه دسترسی ارتباط برقرار کند، باید الگوریتم تک عامل یادگیری تقویتی توضیح داده شده را در مورد چند عامل گسترش دهیم. دو رویکرد برای این منظور وجود دارد، اولین رویکرد استفاده از نظریه بازی‌ها می‌باشد. در بخش ۳-۴ ابتدا به شرح نظریه بازی‌ها پرداخته می‌شود و سپس در بخش ۳-۶ بیان می‌کنیم که چگونه می‌توان یک مسئله یادگیری تقویتی چند عامله را با نظریه بازی حل نماییم و در آخر در بخش ۳-۷ دومین رویکرد که روش پیشنهادی استفاده شده در این پایان‌نامه است، بیان می‌شود.

۳-۲ یادگیری تقویتی

فرآیند یادگیری موجودات زنده یکی از موضوعات تحقیقاتی جدید بشمار می‌آید. این تحقیقات بـه دو دسته کلی تقسیم می‌شوند. دسته نخست به شناخت اصول یادگیری موجودات زنده و مراحل آن می-پردازند و دسته دوم بدنبال ارائه یک متدولوژی برای قرار دادن این اصول در یک ماشـین مـی‌باشـند. یادگیری بصورت تغییرات ایجاد شده در کارایی یک سیستم بر اساس تجربه‌های گذشته تعریف مـی-شود. یک ویژگی مهم سیستم‌های یادگیر، توانایی بهبود کارایی خود با گذشت زمـان اسـت. بـه بیـان ریاضی می‌توان این‌طور عنوان کرد که هدف یک سیستم یادگیر، بهینه‌سازی وظیفه‌ای است که کاملا شناخته شده نیست. بنابراین یک رویکرد به این مسئله، کاهش اهداف سیستم یادگیر به یـک مسـئله بهینه‌سازی است که بر روی مجموعه‌ای از پارامترها تعریف می‌شود و هدف آن پیـدا کـردن مجموعـه پارامترهای بهینه می‌باشد.

در بسیاری از مسائل مطرح شده، اطلاعی از پاسخ‌های صحیح مسئله (که یادگیری با نظارت به آن-ها نیاز دارد) در دست نیست. به همین علت استفاده از یک روش یادگیری به نـام یـادگیری تقـویتی مورد توجه قرار گرفته است. یادگیری تقویتی نه زیر مجموعه شبکه‌های عصبی است و نه انتخابی به-جای آن‌ها محسوب می‌شود. بلکه رویکردی متعامد برای حل مسائل متفاوت و مشکل‌تر بشمار می‌رود.

یادگیری تقویتی به مسئله‌هایی می‌پردازد که در آن یک عامل مستقل، حالت‌هایی را درک کرده و مطابق با آن ادراک، اعمـــال بهینـــه‌ای را بـــرای رسیدن به اهدافش انجام می‌دهد. این مسئله بسیار جامع است و شامل مسائل یادگیری کنترل ربات‌هـای متحـرک، یـادگیری بهینه‌ســازی کارخانه‌ها و یادگیری بازی‌های صفحه‌ای[1] می‌شود. هرگاه که عامل، عملی را در محیطش انجام می‌دهد، یک نـاظر متناسب بـا حالـت و عمـل انجام شده به وی پاداش[2] می‌دهد یا وی را تنبیه[3] می‌کند (پاداش منفی). برای مثال، نـــاظر ممکن است در یک بازی برای برد پاداش مثبت و برای باخت پاداش منفی و برای اعمال دیگر پاداش صفر را در نظر بگیرد. کار عامل یادگیری از این پاداش‌ها (که گاهی تأخیر نیـــــز دارد) اســـت تـــا در اعمال بعدی بیشترین میزان تابع تجمعی پاداش را بگیرد. در این فصل بیشتر بر روی الگوریتمی به نام یادگیری Q[4] که پاداش‌های تأخیری را نیز در نظر می‌گیرد، تمرکز می‌کنیم. این الگوریتم حتی زمانی که عامل هیچ اطلاعاتی در مورد محیط ندارد درست کار می‌کند. الگوریتم‌های یادگیری تقویتی رابطه‌ی نزدیکی با الگوریتم‌های برنامه‌نویســـی پویـــا[5] دارند، که کاربرد بسیاری در مسائل بهینه‌سازی دارد. در بسیاری از حیوانات، یادگیری تقویتی، تنها شیوه‌ی یادگیری مورد استفاده است. همچنین یادگیری تقویتی، بخشی اساسی از رفتار انسان‌ها را تشکیل می‌دهد. هنگامی که دست ما در مواجهه با حرارت می‌سوزد، ما به سرعت یاد می‌گیریم که این کار را بار دیگر تکرار نکنیم. لـذت و درد مثال‌های خوبی از پاداش‌ها هستند که الگوهای رفتـاری مـا و بسـیاری از حیوانـات را تشـکیل

[1] Board Games
[2] Reward
[3] Punishment
[4] Q Learning
[5] Dynamic Programming

می‌دهند. در یادگیری تقویتی، هدف اصلی از یادگیری، انجام دادن کاری و یا رسیدن به هدفی است، بدون آن‌که عامل یادگیرنده، با اطلاعات مستقیم بیرونی تغذیه شود.

یادگیری تقویتی یک روش تعاملی است و با روش‌های یادگیری با نظارت تفاوت دارد. در روش‌های با نظارت مانند ماشین بردار پشتیبان[1] و درخت تصمیم[2] برچسب کلاس‌ها در فاز آموزش تعیین می‌شود ولی در این‌جا بحث یادگیری مطرح است. یعنی عامل باید خودش یاد بگیرد. یادگیری نه خوشه‌بندی[3] و نه دسته‌بندی[4] است و در کل روش متفاوتی است.

همه روش‌های یادگیری با ناظر مانند ماشین بردار پشتیبان و درخت تصمیم روش یادگیری مشابهی دارند. یک فاز آموزش دارند که به کمک داده‌های آموزشی مدل را می‌سازند و سپس یک فاز تست که با توجه به آن‌چه یاد گرفته شده، داده تست را دسته‌بندی می‌کنند. این روش برای کلیه مسائل و بازی‌هایی که دنباله تصمیم در آن‌ها مهم باشد مانند شطرنج، یادگیری فضای مارپیچ[5] توسط عامل و برنامه‌ریزی ربات‌ها مثلا ربات‌های تمیزکننده کارایی دارد. با این روش عامل یاد می‌گیرد که چگونه از وضعیتی که حرکت خود را آغاز کرده خود را به حالت هدف[6] برساند و در هر وضعیت بهترین عمل کدام است. هدف، پیدا کردن کوتاه‌ترین مسیر و بهترین مسیر برای رسیدن از مبدا به هدف است. این کار بصورت پویا انجام می‌شود یعنی نیازی نیست که فضای مسئله ثابت باشد بلکه فضای مسئله در حین اجرا می‌تواند تغییر کند. به عنوان مثال اگر سر راه ربات یک دیوار قرار گیرد باز هم می‌تواند مسیر خود را به سمت هدف پیدا کند. این نوع یادگیری در رباتیک کاربرد زیادی دارد چون ربات دقیقا یاد می‌گیرد که در هر وضعیت بهترین عمل کدام است [۳۶-۳۹]. یادگیری تقویتی، یک روش تعاملی و یادگیری با سعی و خطا است. یعنی عامل یادگیرنده دانش خود را بصورت سعی و خطا از تعامل با محیط می‌گیرد. عامل در محیط قرار گرفته یک عمل انجام می‌دهد، بازخوردی از

[1] Support Vector Machines (SVMs)
[2] Decision Tree
[3] Clustering
[4] Classification
[5] Maze
[6] Goal

محیط می‌گیرد و طبق آن پاسخ می‌فهمد عملی که انجام داده خوب بوده یا خیر. عمل در واقع عمل-هایی است که عامل مجاز است انجام دهد [۴۰]. یادگیری تقویتی غیر قطعی[1] و احتمالاتی است. عامل به ازای هر عمل یک پاداش می‌گیرد. هرچه پاداش بهتر باشد یعنی عملی که انجام شده بهتر بوده است [۴۱]. هدف نهایی عامل این است که بیشترین پاداش را جمع کند، بنابراین خانه هدف باید بیشترین پاداش را داشته باشد. مزیت اصلی یادگیری تقویتی نسبت به سایر روش‌های یادگیری عدم نیاز به هیچ‌گونه اطلاعاتی از محیط (بجز سیگنال تقویتی) است. یادگیری تقویتی یکی از گرایش‌های یادگیری ماشین است. این روش بر رفتارهایی تمرکز دارد که ماشین باید برای بیشینه کردن پاداشش انجام دهد. این مسئله، با توجه به گستردگی‌اش، در زمینه‌های گوناگونی مانند نظریه بازی‌ها، نظریه کنترل، تحقیق در عملیات، نظریه اطلاعات، سامانه چند عامله [۴۲، ۴۳]، هوش ازدحامی، آمار، الگوریتم ژنتیک [۴۴، ۴۵] و بهینه‌سازی بر مبنای شبیه‌سازی بررسی می‌شود. یادگیری تقویتی در اقتصاد و نظریه بازی‌ها بیشتر به بررسی تعادل‌های ایجاد شده تحت عقلانیت محدود می‌پردازد. یادگیری تقویتی با یادگیری با نظارت معمول دو تفاوت عمده دارد، نخست این‌که در آن زوج‌های صحیح ورودی و خروجی در کار نیست و رفتارهای ناکارآمد نیز از بیرون اصلاح نمی‌شوند و دیگر آن-که تمرکز زیادی روی کارایی زنده وجود دارد که نیازمند پیدا کردن یک تعادل مناسب بین اکتشاف[2] چیزهای جدید و بهره‌برداری[3] از دانش اندوخته شده دارد. در یادگیری تقویتی که در آن بازخوردهای فراهم شده برای عامل مجموعه صحیحی از اعمال جهت انجام دادن یک وظیفه هستند، بر خلاف یادگیری نظارت شده از پاداش‌ها و تنبیه‌ها به عنوان سیگنال‌هایی برای رفتار مثبت و منفی بهره برده می‌شود و هدف پیدا کردن مدل داده مناسبی است که پاداش کل را برای عامل بیشینه می‌کند.

یک مدل ابتدایی یادگیری تقویتی از موارد زیر تشکیل شده است:

[1] Nondeterministic
[2] Exploration
[3] Exploitation

- یک مجموعه از حالات مختلف مسئله.
- یک مجموعه از تصمیمات قابل اتخاذ.
- قوانینی برای گذار از حالات مختلف به یکدیگر.
- قوانینی برای میزان پاداش به‌ازای هر تغییر وضعیت.
- قوانینی برای توصیف آن‌چه که ماشین می‌تواند مشاهده کند.

معمولاً مقدار پاداش به آخرین گذار مربوط است. در بسیاری از کارها ماشین می‌تواند وضعیت فعلی مسئله را نیز به‌طور کامل (یا ناقص) مشاهده کند. گاهی نیز مجموعه فعالیت‌های ممکن محدود است (مثلاً این که ماشین نتواند بیشتر از مقدار پولی که دارد خرج کند).

در یادگیری تقویتی ماشین می‌تواند در گام‌های زمانی گسسته با محیط تعامل کند. در هر لحظه مفروض، ماشین یک مجموعه مشاهدات را دریافت می‌کند که معمولاً شامل پاداش نیز می‌شود و پس از آن یک واکنش از بین واکنش‌های ممکن از خود نشان می‌دهد که به محیط ارسال می‌شود. سپس وضعیت به حالت گذار می‌کند و پاداش نیز مشخص می‌شود که مربوط به گذار حاصل است؛ و این پاداش قبل از انجام حرکت بعدی ماشین به ماشین داده می‌شود تا بر اساس آن حرکت بعدی را انتخاب کند. هدف ماشین هم طبیعتاً این است که بیشترین پاداش ممکن را کسب کند. ماشین می‌تواند هر تصمیماتش را به صورت تابعی از روند تغییر بازی تا وضعیت حاضر یا حتی به صورت تصادفی انتخاب کند.

نکته مهمی که در این‌جا وجود دارد این است که یادگیری تقویتی برای مسائلی که در آن ها بیشترین سود در کوتاه مدت تضمین کننده بیشترین سود در دراز مدت نیست، بسیار مناسب است. مثلاً ما برای این‌که در بزرگسالی درآمد بیشتری داشته باشیم بهتر است که به دانشگاه برویم و درس بخوانیم در حالی که این امر در حال حاضر از نظر مالی برای ما بهینه نیست. دلیل وجود این برتری

در روش یادگیری تقویتی نیز این است که ماشین در هر مرحله لزوماً بهترین راه را انتخاب نمی‌کند و در نهایت هم سعی دارد مجموع پاداش دریافتی را حداکثر نماید.

دو عامل مهم هستند که باعث برتری این روش می‌شوند: استفاده از نمونه‌ها برای بهینه‌سازی کارایی و استفاده از تخمین توابع برای تعامل با محیط‌های پیچیده در هریک از حالات زیر:

- برای مسئله یک مدل وجود دارد اما راه‌حل تحلیلی‌ای وجود ندارد.
- فقط یک محیط شبیه‌سازی شده از مسئله در دسترس است.
- هنگامی که تنها راه برای بدست آوردن اطلاعات از محیط تعامل با آن باشد.

دو حالت اول را می‌توان تحت عنوان مسائل برنامه‌ریزی بررسی کرد. (با توجه به این که مدل‌هایی از مسئله موجود است)، در حالی که سومی را باید به عنوان یک مسئله یادگیری صرف در نظر گرفت. در هر حال در چارچوب یادگیری تقویتی هر سه مسئله به یک مسئله یادگیری ماشین تبدیل می‌شوند.

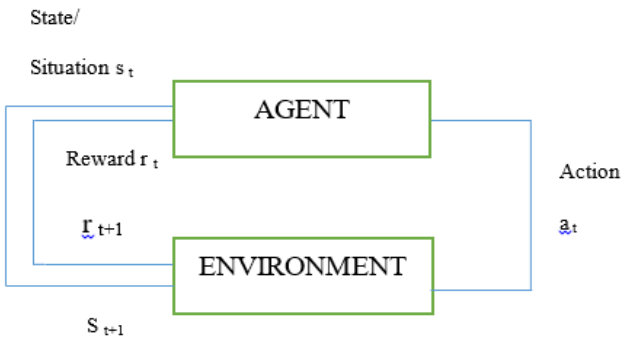
برخی از اصطلاحاتی که عناصر یک مسئله یادگیری تقویتی را تشریح می‌کنند در ادامه بیان شده است:

- محیط : جهان فیزیکی که عامل در آن عمل می‌کند.
- حالت : موقعیت کنونی عامل
- پاداش : بازخورد از محیط
- سیاست[1] : روشی برای نگاشت حالت عامل به عمل

[1] Policy

- ارزش : پاداش آینده که یک عامل با اقدام به یک عمـل در یـک حالـت خـاص بـه آن دست می‌یابد.

در یک مسئله یادگیری تقویتی استاندارد با اجزای اصلی عامل و محیط روبرو هستیم. عامـل مـی-تواند از طریق ورودی‌هایش تشخیص دهد که در چه حالتی قـرار دارد. عامـل در حالـت s_t عمـل a_t را انجام می‌دهد. این کار باعث می‌شود حالت محیط به s_{t+1} تغییر نماید. در اثر این تغییـر حالـت عامـل پاداش r_{t+1} را از محیط دریافت می‌نماید. شکل ۳-۱ این ارتباط را نشان می‌دهد.



شکل ۳-۱) ارتباط عامل و محیط

۳-۲-۱ کاربردهای یادگیری تقویتی

از آن‌جا که یادگیری تقویتی نیازمند حجم زیادی از داده‌ها است، بنابراین بیشتر در دامنه‌هایی مانند بازی‌ها و رباتیک کاربرد دارد که در آن‌ها داده‌های شبیه‌سازی شده به صورت آماده موجود هستند. در قسمت زیر برخی از کاربردهای یادگیری تقویتی بیان شده است.

- بازی‌ها
  - شطرنج

- تخته- نرد
- حل مسائل اقتصادی
  - بازارهای رقابتی (مثلا بازار برق)
  - مدل‌سازی و تصمیم‌گیری بهینه در بازار بورس
- تشخیص الگو
- آموزش شبکه‌های عصبی
- کنترل هوشمند

این روش حول محور به‌روز رسانی ارزش‌های Q است که ارزش انجام عمل a را در حالت s نشان می‌دهد. قاعده به‌روز رسانی ارزش، هسته الگوریتم یادگیری Q محسوب می‌شود. در بخش ۳-۳ این الگوریتم شرح داده می‌شود.

۳-۳ الگوریتم یادگیری Q

یادگیری Q نوعی از یادگیری تقویتی بدون مدل است که بر پایه برنامه‌ریزی پویای اتفاقی عمل می-کند. در یادگیری Q به جای انجام یک نگاشت از حالت‌ها به مقادیر حالت‌ها، نگاشتی از زوج state/action به مقادیری که Q-value نامیده می‌شوند، انجام می‌گردد. بنابراین یادگیری Q یک تکنیک یادگیری تقویتی است که با یادگیری یک تابع اقدام/مقدار، سیاست مشخصی را برای انجام حرکات مختلف در وضعیت‌های مختلف دنبال می‌کند. یکی از نقاط قوت این روش، توانایی یادگیری تابع مذکور بدون داشتن مدل معینی از محیط می‌باشد. در این‌جا مدل مسئله از یک عامل، وضعیت‌ها (S) و مجموعه‌ای از اقدامات (A) برای هر وضعیت تشکیل شده است. با انجام یک اقدام، عامل از یک وضعیت به وضعیت بعدی حرکت کرده و در هر وضعیت پاداشی به عامل می‌دهد. هدف عامل حداکثر کردن کل پاداش دریافتی خود است. این کار با یادگیری اقدام بهینه برای هر وضعیت انجام

می‌گردد. الگوریتم دارای تابعی است که ترکیب state/action را محاسبه می‌نماید. قبل از شروع یادگیری، Q مقدار ثابتی را که توسط طراح انتخاب شده برمی‌گرداند. سپس هر بار که به عامل پاداش داده می‌شود، مقادیر جدیدی برای هر ترکیب state/action محاسبه می‌گردد. هسته الگوریتم از یک به‌روز رسانی تکراری ساده تشکیل شده است. به این ترتیب که بر اساس اطلاعات جدید مقادیر قبلی اصلاح می‌شوند.

۳-۳-۱ تابع Q

به هر زوج <حالت، عمل> یک مقدار Q(s,a) نسبت داده می‌شود. این مقدار عبارت است از مجموع پاداش‌های دریافت شده وقتی که از حالت s شروع کرده و عمل a را انجام داده و به دنبال آن خط-مشی سیاست موجود را دنبال کرده باشیم. برای یادگیری تابع Q می‌توان از جدولی استفاده کرد که هر ورودی آن یک زوج <s,a> به همراه تقریبی است که یادگیر از مقدار واقعی Q بدست آورده است. مقادیر این جدول با مقدار اولیه تصادفی (معمولا صفر) پر می‌شوند. عامل به‌طور متناوب وضعیت فعلی s را تشخیص داده و عملی مثل a را انجام می‌دهد. سپس پاداش حاصله r(s,a) و همچنین حالت جدید ناشی از انجام عمل s'=d(s,a) را مشاهده می‌کند. از آن‌جایی که در محیط، یک حالت هدف در نظر گرفته می‌شود که با رسیدن عامل به آن حالت، حرکت عامل متوقف می‌شود، عمل یادگیری به‌صورت تکراری[1] انجام می‌شود. در هر تکرار عامل در یک محل تصادفی قرار داده می-شود و تا رسیدن به حالت هدف به تغییر مقادیر Q ادامه می‌دهد وقتی به هدف برسد، تکرار تمام می-شود و یک تکرار جدید شروع می‌شود. اگر مقادیر اولیه Q صفر در نظر گرفته شده باشند، در هر تکرار فقط یکی از مقادیر که به مقدار نهائی نزدیک‌تر هستند تغییر کرده و بقیه صفر باقی می‌مانند. با افزایش تکرارها این مقادیر غیر صفر به سایر مقادیر جدول گسترش پیدا کرده و در نهایت به مقادیر بهینه همگرا خواهند شد. در یادگیری Q از اختلاف بین مقدار حالت فعلی و حالت بعد از آن استفاده

[1] Episode

می‌شود. این امر را می‌توان به اختلاف بین حالت فعلی و چند حالت بعدی تعمیم داد. رابطه ۱-۳ [۴۶] تابع Q را نشان می‌دهد.

$$Q(s_t, a_t) \leftarrow \underbrace{Q(s_t, a_t)}_{\text{old value}} + \underbrace{\alpha_t(s_t, a_t)}_{\text{learning rate}} \times \left[\overbrace{\underbrace{R(s_t)}_{\text{reward}} + \underbrace{\gamma}_{\text{discount factor}} \underbrace{\max_{a_{t+1}} Q(s_{t+1}, a_{t+1})}_{\text{max future value}}}^{\text{learned value}} - \underbrace{Q(s_t, a_t)}_{\text{old value}} \right] \quad (۳-۱)$$

یک تکرار الگوریتم وقتی عامل به وضعیت نهایی یعنی هدف می‌رسد پایان می‌یابد. توجه کنید که برای همه وضعیت‌های نهایی، s و Q(s,a) مربوطه هیچ‌گاه به‌روز نمی‌شوند و مقدار اولیه خود را حفظ می‌کنند.

نرخ یادگیری تعیین می‌کند که تا چه میزان اطلاعات بدست آمده جدید بر اطلاعات قدیمی ترجیح داده شود. مقدار صفر باعث می‌شود عامل چیزی یاد نگیرد و مقدار یک باعث می‌شود عامل فقط اطلاعات جدید را ملاک قرار دهد.

نرخ یادگیری می‌گوید که عمل‌های آتی چقدر باید مورد توجه قرار گیرند. نرخ یادگیری زیاد و نزدیک به یک باعث کند شدن یادگیری عامل می‌شود. زیرا عامل به یادگیری‌هایی که از هر مرحله بدست آورده توجهی نکرده و فقط به یادگیری‌های جدید توجه می‌کند. بنابراین همگرایی دیرتر صورت می‌گیرد. این‌که در نهایت عامل می‌تواند یاد بگیرد به این خاطر است که این ضریب یادگیری طبق ماهیت الگوریتم یادگیری Q به صفر همگرا شده و کم کم کاهش پیدا می‌کند، بنابراین در آخر کار یادگیری تسریع می‌شود. نرخ یادگیری همچنین ارتباط بین حافظه قدیم و جدید را توصیف می‌کند. مقدار بالای آن نشان‌دهنده این است که اطلاعات جدید (حافظه جدید) بر حافظه قدیم ارجحیت دارد. نرخ یادگیری بالاتر در ابتدا منحنی یادگیری با شیب بیشتری ایجاد می‌کند زیرا هدف این پارامتر همین است. نتایج نهایی اغلب با نرخ یادگیری پایین‌تر بهتر است. احتمالا به این دلیل که هنگام استفاده از نرخ یادگیری بالاتر تجربیات قدیمی خیلی زود با تجربیات جدید جایگزین می‌شوند

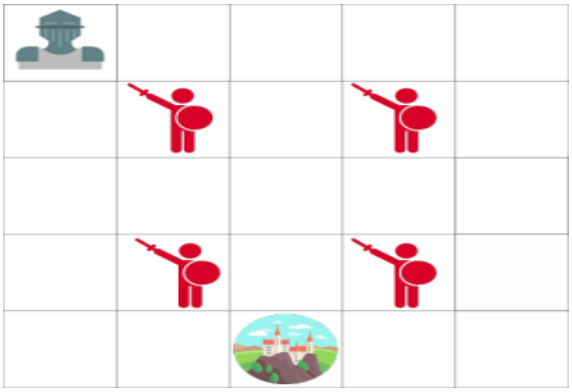
و یادگیری را مختل می‌کنند که به این معنی است که عامل یادگیرنده حافظه خوبی ندارد. برای نتایج نهایی خوب یک نرخ یادگیری نسبتا پایین بهتر است. نرخ یادگیری بالاتر برای شروع کار بهتر است. ولی در کل استفاده از نرخ‌های یادگیری پایین‌تر نتایج بهتری در بر دارد. به عنوان مثال نرخ یادگیری ۰,۸ نتایج کوتاه مدت بدی نداشته ولی نرخ یادگیری مثل ۰,۲ نتایج بلند مدت بهتری دارد. الگوریتم یادگیری Q نسبت به الگوریتم سارسا کندتر همگرا می‌شود ولی قابلیت این را دارد که هم‌زمان با تغییر سیاست‌ها به یادگیری ادامه دهد. در ساده‌ترین شکل یادگیری Q از جداول برای ذخیره داده استفاده می‌شود. این روش با پیچیده شدن سیستم مورد نظر، به سرعت کارایی خود را از دست می-دهد.

عامل تخفیف، اهمیت پاداش‌های آینده را تعیین می‌کند. مقدار صفر باعث می‌شود عامل ماهیت فرصت طلبانه یا کوتاه مدت گرفته و فقط پاداش‌های فعلی را مد نظر قرار می‌دهد. در حالی‌که مقدار یک، عامل را ترغیب می‌کند دوره زمانی طولانی‌تری برای پاداش تقلا کند. اگر این عامل، یک یا بیشتر از یک باشد مقادیر واگرا می‌شود.

۳-۳-۲ یک مثال از الگوریتم یادگیری Q (مسئله: شوالیه و پرنسس)

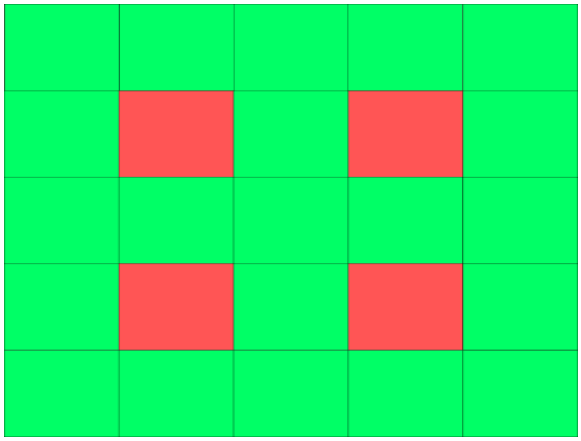
شوالیه‌ای باید پرنسس اسیر شده در یک قلعه که در شکل ۳-۲ نشان داده شده را نجات دهد. شوالیه می‌تواند در هر زمان به اندازه یک کاشی حرکت کند. دشمن توانایی حرکت ندارد، اما اگر شوالیه در خانه‌ای که دشمن وجود دارد قرار بگیرد، کشته می‌شود. هدف آن است که شوالیه از سریع‌ترین مسیر ممکن به قلعه برود. این مسیر را می‌توان با استفاده از سیستم امتیازدهی نقاط[1] ارزیابی کرد. شوالیه در هر گام ۱- امتیاز از دست می‌دهد. اگر شوالیه یک دشمن را لمس کند، ۱۰۰- امتیاز از دست می‌دهد و تکرار به پایان می‌رسد. اگر شوالیه در قلعه باشد، برنده شده و ۱۰۰+ امتیاز دریافت می‌کند.

[1] *Points Scoring*



شکل ۳-۲) مسئله شوالیه و پرنسس

پرسشی که در این وهله مطرح می‌شود این است که «چگونه می‌توان عاملی ساخت کـه چنـین کاری را انجام دهد؟». استراتژی اولی که می‌توان اتخاذ کرد در ادامه مطـرح شـده اسـت. فـرض می‌شود که شوالیه تلاش می‌کند به هر کاشی برود و آن کاشی را رنگی کند. مطابق شـکل ۳-۳ هر کاشی امن به رنگ سبز و هر کاشی ناامن قرمز رنگ می‌شود.

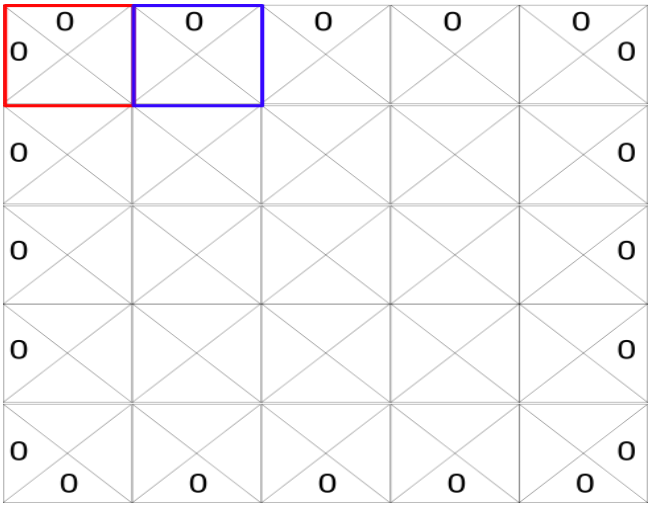


شکل ۳-۳) استراتژی اولیه مسئله شوالیه و پرنسس

می‌توان به عامل گفت که تنها به کاشی‌های سبز رنگ برود. اما مسئله این است که این کار واقعا مفید نیست. زیرا مشخص نیست که در صورت مجاور بودن کاشی‌ها کدام یک بهترین است. بنابراین، عامل در تلاش برای رسیدن به قلعه در یک حلقه بی‌نهایت قرار می‌گیرد. دومین استراتژی ممکن در ادامه بیان می‌شود.

۳-۳-۳ معرفی جدول Q[1]

دومین استراتژی ممکن، ساخت جدولی که در آن بیشینه پاداش آینده مورد انتظار برای هر عمل در هر حالت محاسبه شود. با بهره‌گیری از این جدول، مشخص می‌شود که بهترین عمل ممکن برای هر حالت چیست. هر حالت (کاشی) دارای چهار عمل ممکن است. این چهار عمل، پایین، بالا، چپ و راست رفتن هستند. شکل ۳-۴ شبکه تشکیل شده براساس حالت و عمل است.



شکل ۳-۴) شبکه تشکیل شده براساس حالت و عمل

به منظور انجام محاسبات، این شبکه به یک جدول تبدیل می‌شود. جدول ساخته شده، Q-table نامیده می‌شود. ستون‌ها چهار عمل (چپ، راست، بالا و پایین) هستند. سطرها حالت‌ها هستند. مقدار هر سلول، بیشینه پاداش آینده مورد انتظار برای حالت داده شده و عمل است. هر امتیاز Q-table، پاداش آینده مورد انتظار بیشینه‌ای است که شوالیه در صورت عمل کردن با بهترین سیاست داده شده، به آن دست می‌یابد. به لطف این جدول، شوالیه با نگاه کردن به بالاترین امتیاز در هر سطر برای هر حالت، می‌داند که بهترین عمل قابل انجام چیست. به نظر می‌رسد مسئله قلعه حل شده است. اما اکنون چطور باید ارزش را برای هر جز Q-Table محاسبه کرد؟

[1] Q-Table

۳-۳-۴ یادگیری تابع ارزش عمل[1]

تابع ارزش عمل دو ورودی حالت و عمل را دریافت می‌کند. سپس، پاداش آینده مورد انتظار برای آن حالت و عمل را باز می‌گرداند. رابطه ۲-۳ تابع ارزش عمل را نشان می‌دهد [۴۶].

(۳-۲)

$$\underbrace{Q^{\pi}(s_t, a_t)}_{\text{ارزش Q برای حالت با توجه به عمل}} = \underbrace{E}[\underbrace{R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots}_{\text{پاداش تجمعی تخفیف داده شده مورد انتظار}} | \underbrace{s_t, a_t}_{\text{حالت و عمل داده شده}}]$$

می‌توان تابع Q را به عنوان خواننده‌ای تصور کرد که جدول Q-table را برای یافتن سطر مرتبط به حالت و ستون تخصیص یافته به عمل شوالیه مرور می‌کند. سپس، ارزش Q را از سلول مطابق باز می‌گرداند، این «پاداش آینده مورد انتظار» است.

پیش از آن‌که شوالیه در محیط جست‌وجو کند، Q-table مقادیر دلخواه مشابهی (اغلب اوقات صفر) را ارائه می‌کند. با جست‌وجوی شوالیه در محیط، Q-table تخمین بهتری را با به‌روز رسانی بازگشتی Q(s,a) با استفاده از معادله بلمن[2] در اختیار قرار می‌دهد.

۳-۳-۵ شبه کد الگوریتم یادگیری Q

- مقداردهی اولیه Q-valueها (Q(s,a)) با ارزش‌های دلخواه برای همه جفت‌های حالت و عمل.
- برای زندگی یا تا هنگامی که یادگیری متوقف شود.

[1] Q-Function
[2] Bellman Equation

- عمل (a) را در حالت (s) کنونی محیط بر اساس تخمین کنونی Q-value (یعنی Q(s,a)) انتخاب کن.
- عمل (a) را انجام بده و حالت خروجی (s') و پاداش (r) را مشاهده کن.
- تابع Q(s,a) را با کمک معادله بلمن به‌روز رسانی کن.

حال هرکدام از موارد ذکر شده را شرح می‌دهیم.

گام ۱: مقداردهی اولیه ارزش‌های Q

جدول Q با m (ستون = m) تعداد اعمال و n (سطر = n) تعداد حالت‌ها ساخته شد. اکنون مقداردهی اولیه با «۰» انجام می‌شود. جدول ۳-۱ مقداردهی اولیه جدول با صفر است.

جدول ۳-۱) مقداردهی اولیه جدول با صفر

حالت‌ها	اعمال			
	0	0	0	0
	0	0	0	0
	0	0	0	0
	▬	▬	▬	
	0	0	0	0

گام ۲: برای زندگی (یا تا هنگامی که یادگیری متوقف شد)

گام‌های ۳ تا ۵، تا زمانی که به حداکثر تعداد تکرارها (مشخص شده توسط کاربر) برسد یا بصورت دستی توسط کاربر متوقف شود، تکرار خواهند شد.

گام ۳: انتخاب یک عمل

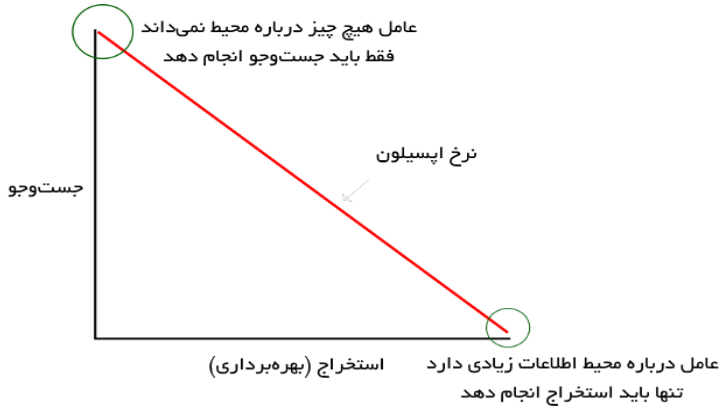
یک عمل a در وضعیت کنونی s برمبنای ارزش کنونی Q-value تخمین زده شده، انتخاب کن. اما، اگر همه Q-valueها برابر با صفر باشند، شوالیه کدام عمل را در ابتدا انجام خواهد داد؟ توازن بین جست‌وجو و استخراج، اکنون حائز اهمیت می‌شود. ایده آن است که در آغاز، از استراتژی حریصانه اپسیلون[1] استفاده شود.

یک نرخ جست‌وجوی اپسیلون تعیین می‌شود که در آغاز برابر با یک قرار داده می‌شود. این نرخ گام‌هایی است که شوالیه به طور تصادفی انجام می‌دهد. در آغاز، این مقدار باید بیشترین ارزش ممکن باشد، زیرا شوالیه هیچ چیز درباره ارزش‌های موجود در Q-table نمی‌داند. این یعنی نیاز به جست‌وجوهای زیاد و انتخاب تصادفی اعمال توسط او است.

یک عدد تصادفی تولید می‌شود. اگر این عدد بزرگ‌تر از اپسیلون بود، استخراج انجام می‌شود (این یعنی از آن‌چه در حال حاضر مشخص است برای انتخاب بهترین عمل در هر گام استفاده می‌شود). در غیر این صورت، جست‌وجو صورت می‌پذیرد.

ایده آن است که باید یک اپسیلون بزرگ در ابتدای راه آموزش Q-function وجود داشته باشد. سپس، به تدریج و پس از آن‌که اطمینان عامل در تخمین Q-valueها افزایش یافت، مقدار آن کاهش پیدا کند. شکل ۳-۶ رابطه بین اپسیلون و جست‌وجو را نشان می‌دهد.

[1] Epsilon Greedy Strategy



شکل ۳-۵) رابطه بین اپسیلون و جست‌وجو

گام ۴-۵: انجام عمل a و مشاهده حالت خروجی s و پاداش r.

اکنون باید تابع Q(s,a) به‌روز رسانی شود. عمل a که در گام ۳ انتخاب شده بود، انجام می‌شود و اجرای این عمل حالت s و پاداش r را در خروجی می‌دهد. سپس، برای به‌روز رسانی Q(s,a)، از معادله بلمن۳-۳ استفاده می‌شود.

$$NewQ(s,a) = Q(s,a) + \alpha[R(s,a) + \gamma max Q'(s',a') - Q(s,a)] \qquad (۳-۳)$$

$NewQ(s,a)$: Q value جدید برای آن حالت و عمل
$Q(s,a)$: کنونی Q Value
α: نرخ یادگیری
$R(s,a)$: پاداش برای انجام آن عمل در آن حالت
γ: نرخ تخفیف
$max Q'(s',a')$: حداقل پاداش آینده پیش‌بینی شده با توجه به s' جدید و همه اعمال ممکن در حالت جدید

۳-۳-۶ جمع‌بندی الگوریتم یادگیری Q

یادگیری Q یک الگوریتم یادگیری تقویتی ارزش محور است، که برای پیدا کردن سیاست انتخاب عمل بهینه از تابع Q استفاده می‌کند. این الگوریتم، عمل مناسب قابل انجام را بر مبنای تابع اقدام/مقدار ارزیابی می‌کند، که مقدار یا ارزش قرار گرفتن در حالت قطعی و انجام یک عمل مشخص در آن حالت را بدست می‌آورد. هدف این الگوریتم بیشینه‌سازی تابع ارزش Q است. (پاداش آینده مورد انتظار با توجه به یک حالت و عمل داده شده).

- جدول Q به عامل در پیدا کردن بهترین عمل برای حالت کمک می‌کند.
- بیشینه‌سازی پاداش مورد انتظار با انتخاب همه بهترین اعمال ممکن انجام می‌شود.
- Q سرنام quality یک عمل مشخص در یک حالت مشخص است.
- تابع Q(state, action)، پاداش آینده مورد انتظار برای یک عمل در حالت تعیین شده را باز می‌گرداند.
- این تابع با استفاده از Q-learning قابل تخمین زدن است و به طور تکرارشونده Q(s,a) را با استفاده از معادله بلمن به‌روز رسانی می‌کند.
- پیش از آن‌که محیط جست‌وجو شود، جدول Q مقادیر ثابت دلخواه را بدست می‌دهد، اما با آغاز جست‌وجوی عامل در محیط، Q تخمین‌های بهتر و بهتری را ارائه می‌کند.

همان‌طور که قبلاً نیز گفته شد، معمول است که بیش از یک کاربر با نقطه دسترسی ارتباط برقرار می‌کند. برای چنین سناریویی، باید الگوریتم تک عامل یادگیری تقویتی توضیح داده شده را در مورد چند عامل گسترش دهیم [۴۷، ۴۸].

دو رویکرد برای این منظور وجود دارد، اولین رویکرد استفاده از نظریه بازی‌ها می‌باشد، که در بخش بعد این رویکرد بیان می‌شود و سپس در بخش ۳-۶ نشان می‌دهیم که چگونه می‌توان یک مسئله یادگیری تقویتی چند عامله را با نظریه بازی حل نماییم و در بخش ۳-۷ دومین رویکرد که روش پیشنهادی استفاده شده در این پایان‌نامه است، بیان می‌شود.

۳-۴ نظریه بازی[1]

یک رویکرد رایج برای یک مسئله چند عامل، یادگیری Nash-Q است [۴۹، ۵۰]. که در آن می‌توان تصور کرد که عوامل برای بدست آوردن حداکثر بازدهی[2] بازی می‌کنند. در این روش به جای تعیین حداکثر مقدار Q برای هر عامل، از ویژگی تعادل نش[3] استفاده می‌شود تا مجموعه بهینه از اقدامات را برای هر حالت انجام دهد [۵۰، ۵۱].

نظریه بازی‌ها، شاخه‌ای ازریاضیات کاربردی است که در علوم اجتماعی و به ویژه در اقتصاد شهری، زیست شناسی، مهندسی،علوم سیاسی، روابط بین الملل، علوم کامپیوتر و فلسفه مورد استفاده قرارگرفته است. نظریه بازی‌ها، در تلاش است با کمک گرفتن از ریاضیات، رفتار را در شرایط راهبردی[4] یا بازی‌ها، که در آن‌ها موفقیت فرد یا بنگاه در انتخاب کردن، وابسته به انتخاب دیگران می‌باشد، تحلیل کند. نظریه بازی‌ها، علم استراتژی است؛ از این رو در اقتصاد شهری،کاربردهای زیادی دارد. این نظریه تلاش می‌کند تا اعمالی که بازیکن‌ها در گستره ای از بازی‌ها باید انجام دهند تا بهترین نتیجه را کسب کنند به شکل ریاضی و منطقی تعیین کند. رقابت شرکت‌ها در کسب منافع بیشتر در سطح شهرها، نمونه‌ای از بازی‌های اقتصادی را نشان می‌دهد. تمام این بازی‌ها یک ویژگی مشترک دارند و آن نوعی وابستگی درونی است؛ به این معنا که نتیجه بازی برای هر یک از شرکت کننده‌ها به انتخاب‌های (استراتژی‌های) همه افراد بستگی دارد [۵۰، ۵۱].

نظریه بازی‌ها برای اولین بار توسط ریاضی‌دان آقای جان نش به عموم معرفی شد [۵۲، ۵۳]. اما نظریه بازی‌ها کمی با بازی‌هایی که می‌شناسیم فرق دارد. نظریه بازی تلاش می‌کند تا رفتار ریاضی حاکم بر یک موقعیت راهبردی را مدل‌سازی کند. این موقعیت، زمانی پدید می‌آید که موفقیت یک

[1] Game Theory
[2] Pay Off
[3] Nash Equilibrium
[4] Strategic

فرد وابسته به راهبردهایی است که دیگران انتخاب می‌کنند. هدف نهایی این دانش، یافتن راهبرد بهینه برای بازیکنان است.

امروزه، نظریه بازی‌ها، علومی است که به تحلیل رفتار منطقی متقابل انسان‌ها، حیوانات و رایانه‌ها می‌پردازد. نظریه بازی‌ها در ابتدا برای درک مجموعه بزرگی از رفتارهای اقتصادی به عنوان مثال نوسانات شاخص سهام در بورس اوراق بهادار و افت‌وخیز بهای کالاها در بازار مصرف‌کنندگان ایجاد شد. تحلیل پدیده‌های گوناگون اقتصادی و تجاری نظیر پیروزی در یک مزایده، داد و ستد، معامله و شرکت در یک مناقصه، از دیگر مواردی است که نظریه بازی در آن نقش ایفا می‌کند [۵۲، ۵۳].

۳-۴-۱ منظور از بازی چیست؟

در این نظریه: بازی، یک تعامل یا رقابت بین چندین بازیکن است که تصمیم یا تغییر حالت یکی از بازیکنان روی بقیه تاثیر می‌گذارد. یک بازی شامل مجموعه‌ای از بازیکنان، مجموعه‌ای از حرکت‌ها یا راهبردها و نتیجه مشخصی برای هر ترکیب از راهبردها می‌باشد. پیروزی در هر بازی تنها تابع احتمالات نیست، بلکه اصول و قوانین ویژه خود را دارد و هر بازیکن در طی بازی سعی می‌کند با به کارگیری آن اصول خود را به برد نزدیک کند. شما از این سیستم می‌توانید در تمام شرایطی که چند نفر در آن درگیر هستند و نیاز به تصمیم‌گیری است، استفاده کنید. مانند بازی شطرنج، تصمیمات و برنامه‌های یک شرکت، که با نوع فعالیت‌های رقبای آن‌ها تغییر می‌کند. مثلا اگر یکی از رقبا قیمت-هایش را کاهش دهد، دیگران نیز ممکن است مجبور به همین کار شوند. به طور مثال شما در پاساژی لباس فروشی دارید. حالا دو مغازه کنار شما که آن هم لباس فروشی با کیفیت مانند شما می‌باشد، شرایط خرید خاصی مانند تخفیف برای مشتریان خود وضع کند. با این شرایط تخفیف، شما مشتریان خود را به تدریج به آن مغازه از دست خواهید داد و برای مقابله با این شرایط باید استراتژی خود را پیاده کنید. حالا تصور کنید در آن پاساژ، بیشتر از ۲ لباس فروشی وجود داشته باشد و این باعث می-شود که تغییر در استراتژی هر کدام از آن‌ها در کاسبی بقیه تاثیر گذاشته و باعث تغییر استراتژی بقیه

مغازه‌ها خواهد شد. هر کدام شاید نقشه خود را داشته باشند و این تصمیمات روی بقیه مغازه‌دارها تاثیر خواهد گذاشت و به همین صورت این چرخه بازی ادامه پیدا می‌کند. بنابراین نظریه بازی‌ها، رفتار تعداد ۲ یا بیشتر شرکت‌کننده را برای رسیدن به پاداش یا دوری از مجازات، بررسی و مدل‌سازی می‌کند. رقابت شهرها برای دست‌یابی به منابع مالی، سرمایه انسانی، جذب شرکت‌های بین المللی و رقابت شرکت‌های تجاری در بازار بورس کالا نمونه‌هایی از بازی‌ها هستند.

برای تعریف فضای بازی، مشخص کردن عناصر زیر لازم و کافی است:

۱- بازیکنان: طرف‌های بازی که هر کدام حداقل دو راهبرد در اختیار دارند.

۲- راهبرد در اختیار هر بازیکن: زنجیره‌ای مرتب از اقداماتی است که بازیکن می‌تواند در قدم‌های مختلف بازی برگزیند.

۳- ترتیب بازی: این که در هر قدمی از بازی، چه بازیکنی حرکت می‌کند.

۴- ساختار اطلاعاتی: در هر لحظه از بازی هر بازیکن می‌تواند چه اطلاعاتی را از حرکت‌ها و ترجیحات طرف مقابلش بداند.

۵- خروجی‌های بازی: وقتی بازی به انتها می‌رسد چه نتایجی به بار می‌آید.

۲-۴-۳ انواع بازی

انواع بازی را می‌توان به شکل زیر طبقه‌بندی کرد:

۱- بازی با مجموع صفر: در این بازی سود یک بازیکن معادل زیان بازیکن دیگر است.

۲- بازی با مجموع غیر صفر: در این بازی تصمیمات یک بازیکن ممکن است به نفع همه بازیکنان تمام شود.

۳- بازی تعاونی: در این نوع بازی امکان سازش و تبانی با دیگران وجود دارد.

۴- بازی غیر تعاونی: در این نوع بازی امکان سازش و تبانی بین شرکت کنندگان وجود ندارد.

در بازی‌های موسوم به «بازی با حاصل جمع صفر»، منافع بازیکن‌ها به طور کامل با یک‌دیگر تعارض پیدا می‌کند، به گونه‌ای که بهره‌ای که یک فرد کسب می‌کند، همواره معادل ضرر فرد دیگر است. شایان ذکر است که بازی‌هایی که در آن‌ها همه بازیکنان ممکن است به منفعت دست یابند (یعنی بازی‌های با حاصل جمع مثبت) یا ضرر کنند (یعنی بازی‌های با حاصل جمع منفی) معمول‌تر هستند. در این بازی‌ها، درجات مختلفی از تعارض وجود دارد. تمرکز این نظریه در سال‌های اول شکل گیری‌اش، بر بازی‌های دارای تعارض خالص (بازی‌های با حاصل جمع صفر) بود. سایر بازی‌ها، همکارانه در نظرگرفته می‌شدند؛ به این معنا که شرکت کنندگان اعمال خود را به اتفاق یکدیگر انتخاب کرده و انجام می‌دهند. اما در تحقیقات اخیر بر بازی‌هایی تمرکز شده که نه دارای حاصل جمع صفر هستند و نه به طور خالص همکارانه هستند. در این بازی‌ها بازیکنان اقدامات خود را به صورت جداگانه انتخاب می‌کنند، اما روابط آن‌ها با یکدیگر حاوی رقابت و همکاری است. بازیکنان باید در تصمیمی که اتخاذ می‌کنند، هم به تعارض و هم به احتمال همکاری توجه داشته باشند. اصل و جوهر هر بازی، وابستگی درونی میان استراتژی‌های بازیکنان است.

دو نوع مختلف از ارتباط و وابستگی درونی وجود دارد: وابستگی پیاپی و وابستگی هم‌زمان. در نوع پیاپی، بازیکن‌ها به ترتیب عمل کرده و هریک از اقدامات قبلی دیگران آگاهند. در نوع هم‌زمان، بازیکنان در یک زمان عمل می‌کنند و هریک، از اقدامات دیگری بی‌اطلاع است. نگاه به جلو و استدلال عقب گرد، یک اصل عمومی برای بازیکنان شرکت‌کننده در بازی با حرکت‌های پیاپی است. هر بازیکن باید در نظر بگیرد که دیگران چگونه به حرکت فعلی او واکنش نشان خواهند داد و در مقابل، خودش دست به چه انتخابی خواهد زد. هر بازیکن پیش‌بینی‌اش به کجا خواهد انجامید و از این اطلاعات برای محاسبه بهترین تصمیمی که باید اتخاذ کند، استفاده می‌کند. هر بازیکن زمانی که راجع به چگونگی

واکنش دیگران فکر می‌کند، باید خود را به‌جای آن‌ها قرار دهد و همانند آن‌ها فکر کند. او نباید استدلال خود را به آن‌ها تحمیل کند.

در بازی‌های هم‌زمان اگرچه بازیکن‌ها به صورت هم‌زمان و بدون اطلاع از حرکات فعلی یک‌دیگر عمل می‌کنند، اما همه آن‌ها باید از این نکته آگاه باشند که بازیکنان دیگر عاقل هستند. طبق این تفکر، هر فرد «فکر می‌کند که بازیکن دیگر فکر می‌کند که او فکر می‌کند...»؛ بنابراین، هر بازیکن باید خود را به صورت مجازی به‌جای تمامی افراد دیگر قرار دهد و سعی کند پیامد هر حرکت را محاسبه نماید. بهترین اقدام برای او با توجه به این‌گونه محاسبه‌های کلی بدست می‌آید.

پژوهش‌ها در این زمینه، اغلب بر مجموعه‌ای از راهبردهای شناخته شده به عنوان تعادل در بازی‌ها استوار است که از قواعد عقلانی به نتیجه می‌رسند. مشهورترین تعادل‌ها، تعادل نش است.

۳-۵ تعادل نش

براساس نظریه تعادل نش اگر فرض کنیم در هر بازی با استراتژی مختلط، بازیکنان به طریق منطقی و معقول راهبردهای خود را انتخاب کنند و به دنبال حداکثر سود در بازی هستند، حداقل یک راهبرد برای بدست آوردن بهترین نتیجه برای هر بازیکن قابل انتخاب است و چنان‌چه بازیکن راهکار دیگری به غیر از آن را انتخاب کند، نتیجه بهتری بدست نخواهد آورد. در واقع تعادل نش به حالت پایداری در یک بازی اطلاق می‌گردد که با فرض ثابت بودن راهبرد سایر بازیکنان، یک بازیکن با تغییر بازی خود نتواند به شرایط بهتری دست یابد [۵۴].

۳-۶ حل یک مسئله یادگیری تقویتی چند عامله با استفاده از نظریه بازی‌ها

برای حل یک مسئله یادگیری تقویتی چند عامله با استفاده از نظریه بازی‌ها یک بازی بین عامل‌ها طراحی می‌شود. به این صورت که عامل‌ها به عنوان بازیکن‌ها هستند و برای هر یک از آن‌ها یکسری عمل تعریف می‌شود که می‌توانند انجام دهند. عوامل دارای اهداف فردی و توانایی تصمیم‌گیری مستقل هستند، اما در تصمیمات یکدیگر نیز موثر هستند [۵۵]. اگر فرض کنیم که دو بازیکن در محیط بازی می‌کنند، عمل انتخاب شده توسط بازیکن اول، یک ردیف در ماتریس را انتخاب می‌کند، در حالی که بازیکن دوم، ستون را تعیین می‌کند. در واقع بازیکنان یک و دو نیز به ترتیب به عنوان ردیف و ستون ماتریس معرفی می‌شوند. با استفاده از ماتریس‌های چند بعدی، بازی‌هایی با تعداد بازیکن بیشتر می‌توانند نمایش داده شوند، که هر ورودی در ماتریس شامل بازپرداخت هر یک از عوامل برای اقدامات مربوط به ترکیب آن است [۵۶]. در این روش به جای تعیین حداکثر مقدار Q برای هر عامل، از ویژگی تعادل نش استفاده می‌شود تا مجموعه بهینه از اقدامات را برای هر حالت انجام دهد.

با این وجود، این رویکرد نمی‌تواند برای مشکل ما مؤثر باشد زیرا علاوه بر تعداد زیاد کاربران، تعداد اقدامات معتبر برای هر عامل می‌تواند زیاد باشد. به عنوان مثال، اگر تعداد عامل‌ها و اقدامات معتبر آنها به ترتیب ۱۰ و ۱۲ باشد، برای دست‌یابی به تعادل نش برای استراتژی مختلط باید مجموعه‌ای از ۱۲۰ معادله حل شود. بنابراین، تعیین تعادل نش برای چنین مشکلی با این مقدار محاسبات، که باید هر بار برای انتقال از یک حالت به حالت دیگر تکرار شود، بسیار زمان‌بر است. علاوه بر این، هر زمان که کاربر جدید بخواهد با نقطه دسترسی ارتباط برقرار کند، تعادل نش برای همه کاربران با هم

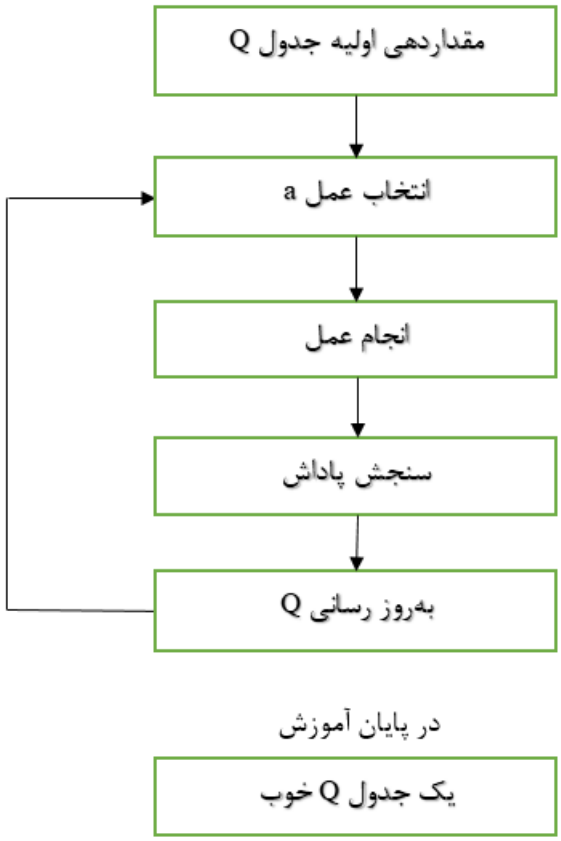
محاسبه می‌شود که می‌تواند منجر به تغییر در توالی حالت کاربران قبلاً متصل شده و در نتیجه افزایش تأخیر برای آن‌ها شود.

۳-۷ رویکرد پیشنهادی

از آن‌جا که کاهش تاخیر در برقراری ارتباط بین فرستنده و گیرنده تا حد امکان مهم است، باید برای مسئله خود رویکرد دیگری در نظر بگیریم. ما می‌توانیم مدل را با توجه به این که کاربران معمولاً هم-زمان درخواست اتصال نمی‌کنند، ساده کنیم. به عبارت دیگر، ما می‌توانیم این مسئله را به عنوان یک مدل تک عاملی در نظر بگیریم که برای هر کاربر و محیط تکرار می‌شود و بعد از هر تکرار اصلاح می-شود. شایان ذکر است که این تغییر مورد نیاز است زیرا هر کاشی فقط توسط یک عامل قابل استفاده است. بدین ترتیب، در پایان هر قسمت برای هر عامل، برخی از حالات برای سایر عوامل نامعتبر هستند. در این مدل، همه عوامل دارای یک جدول Q مشترک هستند و آن را به‌روز می‌کنند [۳۸]. بنابراین، می‌توانیم معادله بلمن را طبق رابطه ۴-۳ بازنویسی کنیم. با اضافه کردن یک زیرنویس i برای نشان دادن عامل i و تعیین مقدار Q به عنوان تابعی از عملکرد همه عوامل [۵۰]:

$$Q(s_t^i, a_t^i) \leftarrow Q(s_t^i, a_t^i) + \alpha[r_{t+1}^i(s_t^i, a_t^i) + \gamma \max Q(s_{t+1}^i, a_{t+1}^i) - Q(s_t^i, a_t^i)] \quad (۳-۴)$$

شکل ۳-۶ معماری روش پیشنهادی را نشان می‌دهد.



شکل ۳-۶) معماری روش پیشنهادی

۳-۸ مدل سیستم

در این بخش، مدل یادگیری تقویتی یک سیستم ارتباطی مبتنی‌بر SDM را ارائه می‌دهیم که در فرکانس‌های موج میلی‌متر، برای نسل‌های بعدی شبکه‌های بی‌سیم کار می‌کند. با استفاده از یادگیری تقویتی، عامل می‌تواند تصمیم بهتری بگیرد در حالی که به تدریج با محیط در تعامل است [۵۷، ۵۸]. ما از روش یادگیری Q، که یکی از روش‌های یادگیری تقویتی مبتنی‌بر ارزش است، برای ایجاد ارتباطات بی‌سیم مناسب بین کاربران و نقطه دسترسی[1](AP) با کنترل وضعیت کاشی‌ها در محیط تحت پوشش SDMها، استفاده می‌کنیم. یادگیری Q به دنبال یک سیاست عملیاتی مطلوب است،

[1] Access point

یعنی دنباله ای از اقدامات که پاداش تخفیف یافته مورد انتظار را به حداکثر می‌رساند [۴۶]. این سیاست بهینه، انتخاب اقدام عامل را با توجه به وضعیت آن تعیین می‌کند. به عنوان یک سیاست انتخاب عمل، عامل با احتمال ε-۱ عملی را با بالاترین مقدار Q انتخاب می‌کند و با احتمال ε تصادفی عمل می‌کند. به عبارت دیگر، باید بین اکتشاف و بهره‌برداری ارتباط برقرار شود. اکتشاف سعی می‌کند کارهایی را انجام دهد که قبل از اجرای آن، هنگام اجرای عامل در یک وضعیت خاص انجام نشده باشد. در ابتدای فرایند یادگیری، عامل هیچ تجربه‌ای از محیط ندارد و بنابراین باید پاداش و مجازات-هایی از محیط کسب کند. بنابراین، در این شرایط، اکتشاف غالب است و مقدار ε نزدیک به یک است. البته، هرچه عامل بیشتر با محیط ارتباط برقرار کند و تجربیاتی را بدست آورد، ارزش آن کاهش می-یابد و بنابراین، بهره‌برداری غالب می‌شود.

۳-۸-۱ عناصر مدل یادگیری تقویتی برای سیستم ارتباطی مبتنی‌بر SDM

مدل یادگیری تقویتی برای سیستم ارتباطی مبتنی‌بر SDM دارای عناصر زیر است:

- عامل: موج (سیگنال) الکترومغناطیسی منتقل شده بین کاربر و نقطه دسترسی است.

- حالت: کاشی‌های موجود در مدل، نقش حالت‌ها را در یک مسئله یادگیری تقویتی ایفا می‌کنند.

در آغاز هر قسمت، با توجه به موقعیت هر کاربر، سیگنال آن توسط یک کاشی مناسب (به عنوان مثال نزدیکترین کاشی) به عنوان اولین حالت دریافت می‌شود. پس از آن، عامل (سیگنال) با توجه به سیاست مطلوب، به حالت‌های بعدی (کاشی) تا رسیدن به هدف حرکت می‌کند. لازم به ذکر است که هر کاشی می‌تواند امواج الکترومغناطیسی را با زوایای مشخص دریافت کند. بنابراین وقتی عامل در حالت s_t می‌باشد، برخی از حالت‌ها می‌توانند به عنوان حالت بعدی $s_t + 1$ برای عامل معتبر باشند. علاوه بر این، هر حالت در طول ارتباط تنها توسط یک عامل قابل استفاده است. بنابراین، اقداماتی که

منجر به انتقال از یک حالت به حالت‌های قبلی استفاده شده می‌شوند، معتبر نیستند. شایان ذکر است که مقادیر Q برای حالت‌های مختلف در سرور ذخیره می‌شوند.

عمل: برای هر کاشی، موج می‌تواند با برخی از زوایای از پیش مشخص شده منتقل شود. بنابراین، برای هر حالت، عمل می‌تواند با توجه به یکی از این زاویه‌های معتبر، هدایت امواج (عوامل) یا جذب امواج باشد. سرور مطابق با مقادیر Q یکی از این زوایای معتبر را انتخاب می‌کند.

پاداش: برای هر هاپ[1] یعنی انتقال از یک کاشی به کاشی دیگر، پاداش منفی (مجازات) را در نظر می‌گیریم (r_h). این مجازات باعث می‌شود عامل با داشتن هاپ کمتر به هدف برسد و بنابراین مانع از نوسان عامل می‌شود. بعلاوه، اگر عامل با انتخاب یک عمل معتبر به هدف برسد، پاداش بزرگی را در نظر می‌گیریم (r_g). از طرف دیگر، اگر مسیر عامل با مانع مسدود شود، عامل مجازات بزرگی دریافت می‌کند (r_o).

۳-۹ جمع‌بندی

در روش پیشنهادی هدف ایجاد ارتباطات بی‌سیم مناسب بین کاربران و نقطه دسترسی است، به نحوی که در برقراری ارتباط تاخیر کاهش یابد. همچنین در صورتی که محیط تغییر کند، به عنوان مثال یک مانع در مسیر امواج قرار بگیرد، محیط سریع بتواند خودش را با شرایط جدید تطبیق بدهد و مسیر جدید را برای امواج مسدود شده، انتخاب نماید.

[1] hop

فصل ۴: نتایج و آزمایش‌ها

۴-۱ مقدمه

در این پایان‌نامه، رویکردی مبتنی‌بر آموزش یادگیری تقویتی پیشنهاد شده است تا بتواند محیط ارتباطات بی‌سیم داخلی را کنترل و تنظیم کند. در این مدل یادگیری، امواج الکترومغناطیسی منتقل شده نقش عامل‌ها، کاشی‌ها نقش حالت‌ها را ایفا می‌کنند و عملکرد متفاوت کاشی‌ها به عنوان اقدامات معتبر در نظر گرفته می‌شوند. با استفاده از الگوریتم یادگیری تقویتی، کاشی‌ها در حالی که امواج الکترومغناطیسی به تدریج با محیط در تعامل هستند، می‌توانند تصمیم بهتری بگیرند. بنابراین، هنگامی که تغییری در محیط رخ می‌دهد، به عنوان مثال یک مانع برخی از امواج الکترومغناطیسی را مسدود می‌کند، عامل مجازات زیادی دریافت می‌کند و بنابراین یک عمل جدید انتخاب می‌شود. از آن‌جا که معمولاً بیش از یک کاربر وجود دارد، ما محیط ارتباطات بی‌سیم را به عنوان یک مشکل یادگیری تقویتی چند عامله مدل می‌کنیم.

در این فصل، عملکرد روش پیشنهادی بررسی شده است. همچنین روش پیشنهادی ارائه شده با برخی از روش‌های دیگر در این زمینه مقایسه شده و نتایج گزارش شده است. ما سناریوهای مختلفی را برای ارزیابی اثربخشی رویکرد خود، مانند ارائه پوشش دائمی برای هر کاربر زمانی که تعداد کاربران افزایش می‌یابد، ارتباط در حضور موانع، فراهم کردن اتصال پس از ظهور ناگهانی برخی موانع و حذف تداخل چند مسیره در نظر می‌گیریم. نتایج شبیه‌سازی نشان می‌دهد که رویکرد یادگیری تقویتی، میانگین توان دریافتی را تا ۱۲٪ در مقایسه با کار گذشته که از الگوریتم ژنتیک برای ایجاد محیط قابل برنامه‌ریزی استفاده کرده [۱۷]، بهبود داده است. در کار گذشته، میانگین توان دریافتی برای ۱۲ گیرنده، ۲۵.۳۸ dBmW می‌باشد.

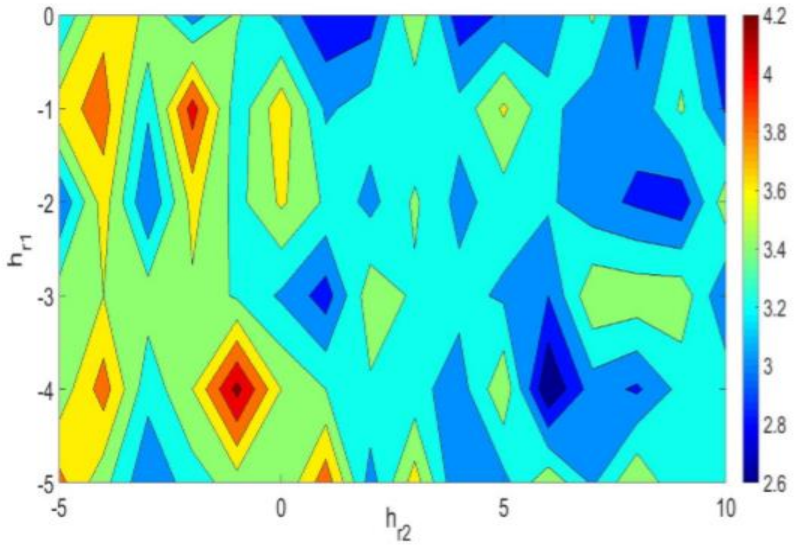
۴-۲ محیط آزمایش

در این بخش الگوریتم یادگیری تقویتی را برای ایجاد پیوندهای ارتباطی بی‌سیم در محیط تحت پوشش SDM ارزیابی می‌کنیم. ما یک فضای داخلی، شبیه به آن‌چه در [۱۰، ۱۷] استفاده می‌شود، با ابعاد ۳×۱۰×۱۵ متر در نظر می‌گیریم. دیواری به طول ۱۲ متر و ضخامت ۱ متر در وسط اتاق وجود دارد که دو بخش جداگانه، خطوط بینایی[1](LOS) و مناطق بدون دید ایجاد می‌کند. تمام دیوارها با کاشی های ۱×۱ متر پوشانده شده‌اند. بنابراین، با توجه به ابعاد اتاق، ۱۹۳ کاشی وجود دارد. یک AP که در۶۰ گیگاهرتز با قدرت انتقال ۱۰۰ dBmW کار می‌کند، در ارتفاع ۲ متری جایی در منطقه LOS قرار دارد. در مدل ما، کاشی‌ها می‌توانند امواج الکترومغناطیسی دریافت شده را از جهت-های ایجاد شده با ترکیبی از زوایای $\{۶۰^\circ ، ۴۵^\circ ، ۳۰^\circ ، ۱۵^\circ ، ۰^\circ ، -۱۵^\circ ، -۳۰^\circ ، -۴۵^\circ ، -۶۰^\circ\}$ تشکیل دهند یا جذب کنند. فرض می‌کنیم وقتی کاشی موج را هدایت می‌کند، هیچ توانی از بین نمی‌رود و از طرف دیگر، تمام توان موج جذب می‌شود. برای ارزیابی عملکرد سیستم مبتنی‌بر یادگیری تقویتی پیشنهادی، یک شبیه‌ساز ردیابی سه بعدی توسط پایتون را توسعه می‌دهیم. ما پاداش را به صورت یک رابطه در نظر می‌گیریم که وابسته به وضعیت‌های بعدی در مناطق LOS و NLOS می‌باشد. اگر فرض کنیم امواج الکترومغناطیسی از کاربر در منطقه NLOS به AP در منطقه LOS منتقل می‌شوند، برای r_{h2} ارزش بیشتری را برای تشویق امواج الکترومغناطیسی برای ورود به منطقه LOS برای رسیدن به AP در اسرع وقت در نظر می‌گیریم. به عبارت دیگر اگر برای حالت بعدی با $x > ۵$ باشد، پاداش r_{h1} می‌شود و در غیر این صورت پاداش r_{h2} می‌شود.

برای بدست آوردن مقادیر بهینه برای پاداش در مدل‌سازی یادگیری تقویتی، میانگین تعداد کاشی‌های فعال شده را در یک سناریو با سه کاربر مستقر در منطقه NLOS برای مقادیر مختلف r_{h1} و r_{h2} محاسبه می‌کنیم.

[1] line of sight

مطابق شکل ۴-۱، حداقل تعداد متوسط کاشی‌های فعال شده برای ۴- = r_{h1} و ۶ = r_{h2} بدست می‌آید.



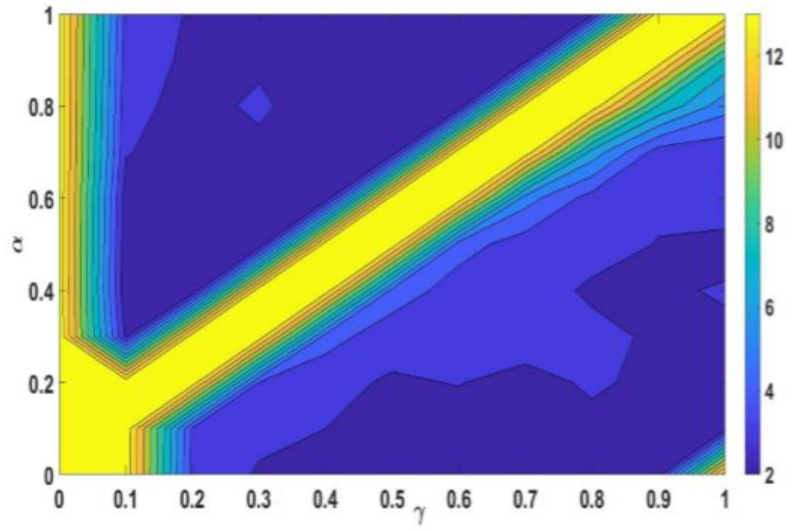
شکل ۴-۱) حداقل تعداد متوسط کاشی‌های فعال شده برای r_{h1} و r_{h2}

همچنین اگر عامل با انتخاب یک عمل معتبر به هدف برسد، پاداش بزرگ (r_g) و اگر مسیر عامل با مانع مسدود شود، مجازات بزرگ (r_o) را در نظر می‌گیریم.

علاوه بر این، مطابق با یک فرآیند مشابه، بهینه‌های γ و α برای معادله بلمن با توجه به رابطه ۴-۱ نیز باید تعیین شوند.

$$Q(s_t^i, a_t^i) \leftarrow Q(s_t^i, a_t^i) + \alpha[r_{t+1}^i(s_t^i, a_t^i) + \gamma \max Q(s_{t+1}^i, a_{t+1}^i) - Q(s_t^i, a_t^i)] \quad (۴-۱)$$

همانطور که در شکل ۴-۲ مشاهده می‌شود، برای پارامترهای γ و α انتخاب‌های زیادی وجود دارد که برای محاسبه آن‌ها میانگین تعداد کاشی‌ها را حداقل می‌کنیم و ۰.۹ = γ و ۰.۲ = α را در نظر می‌گیریم.



شکل ۴-۲) پارامترهای γ و α

بعد از انجام مراحل بهینه‌سازی، پارامترهای مربوط به الگوریتم یادگیری تقویتی و PWE که در شبیه-سازی‌ها استفاده می‌شوند، به ترتیب مطابق جدول ۴-۱ و جدول ۴-۲ تعیین می‌شوند.

جدول ۴-۱) پارامترهای مربوط به الگوریتم یادگیری تقویتی مورد استفاده در شبیه‌سازی

پارامتر	مقدار
r_{h1}	-۴
r_{h2}	۶
r_g	۱۰۰
r_o	-۱۰۰
γ	۰.۹
α	۰.۲

جدول ۴-۲) پارامترهای مربوط به PWE مورد استفاده در شبیه‌سازی

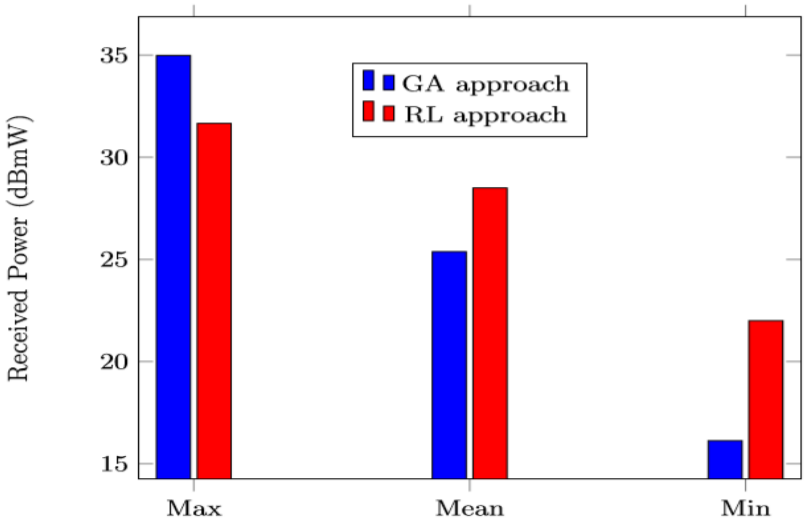
پارامتر	مقدار
ارتفاع سقف	۳ m
ابعاد کاشی	۱ m × ۱ m
فرکانس	۶۰GH_z
توان	۱۰۰dBmW

۴-۳ خلاصه آزمایش‌ها

در این بخش چهار سناریو مختلف را در نظر می‌گیریم و روش پیشنهادی خود را در مقایسه با آثار مرتبط ارزیابی می‌کنیم.

۴-۳-۱ سناریو ۱: تامین پوشش منطقه NLOS

یکی از رویکردهای تأمین پوشش در منطقه NLOS، قرار دادن تعداد زیادی گیرنده در آن منطقه و تنظیم مناسب وضعیت هر کاشی است. این تنظیم به طور معمول با استفاده از الگوریتم‌های بهینه-سازی مانند GA انجام می‌شود. برای مثال، در مرجع [۱۷]، ۱۲ گیرنده به‌طور یکنواخت در ناحیه NLOS توزیع می‌شوند و سپس با استفاده از GA، نتایج بهینه کاشی‌ها جستجو می‌شوند تا حداقل توان دریافت شده بر روی این گیرنده‌ها به حداکثر برسد. در این زیر بخش، ما از مدل پیشنهادی مبتنی‌بر یادگیری تقویتی استفاده می‌کنیم تا ارتباط‌های بی‌سیم مناسب بین این ۱۲ گیرنده و نقطه دسترسی که در موقعیت {۲، ۳، ۲.۲۵} متر قرار دارد، برقرار شود. بنابراین، ما الگوریتم چند عامله یادگیری تقویتی پیشنهادی را برای تنظیم درست محیط استفاده می‌کنیم.



شکل ۴-۳) حداکثر ، میانگین و حداقل توان دریافتی توسط ۱۲ گیرنده قرار داده شده در منطقه NLOS با استفاده از رویکردهای GA و RL

شکل ۴-۳، حداکثر، میانگین و حداقل کل توان دریافتی این گیرنده‌ها را در مقایسه با نتایج بدست آمده در مرجع [۱۷] نشان می‌دهد. مطابق با این روش، رویکرد مبتنی‌بر یادگیری تقویتی میانگین و حداقل مقادیر توان‌های دریافتی را به ترتیب، پس از حدود ۵۰۰ تکرار، به ترتیب ۱۲٪ و ۳۵٪ بهبود می‌بخشد.

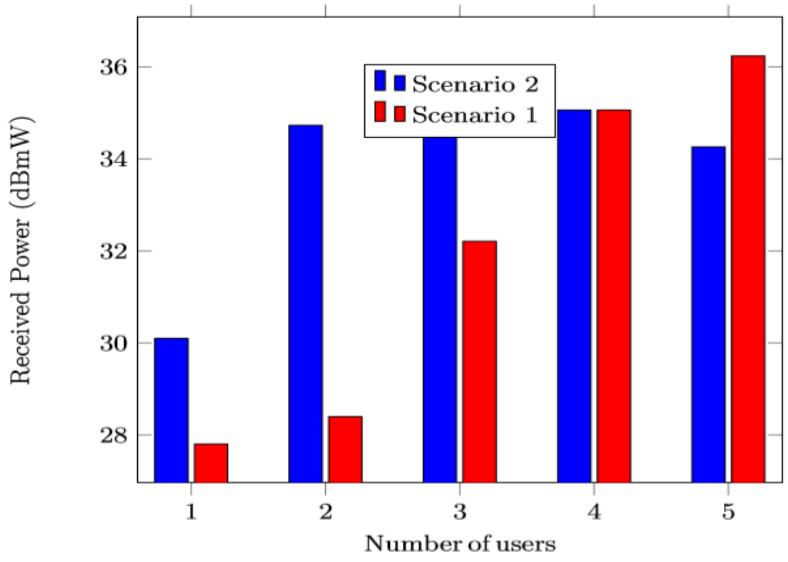
علاوه بر این، تعداد متفاوتی از کاربران مستقر در منطقه NLOS را در نظر می‌گیریم و توانایی پوشش این سناریو را بر اساس یادگیری پیشین یعنی جدول Q حاصل به دست آمده مربوط به ۱۲ گیرنده که قبلاً ذکر شده را بررسی می‌کنیم. جدول ۴-۳ حداقل، حداکثر و میانگین توان دریافتی توسط کاربران را نشان می‌دهد. میانگین مقادیر توان‌های دریافت شده بیشتر از ۲۶.۴۳ dBmW است، که بدیهی است برای برقراری ارتباط مناسب به اندازه کافی بزرگ هستند.

جدول ۴-۳) حداقل ، حداکثر و میانگین کل توان دریافتی توسط کاربران به صورت تصادفی در سناریو۱، مطابق با رویکرد پیشنهادی RL

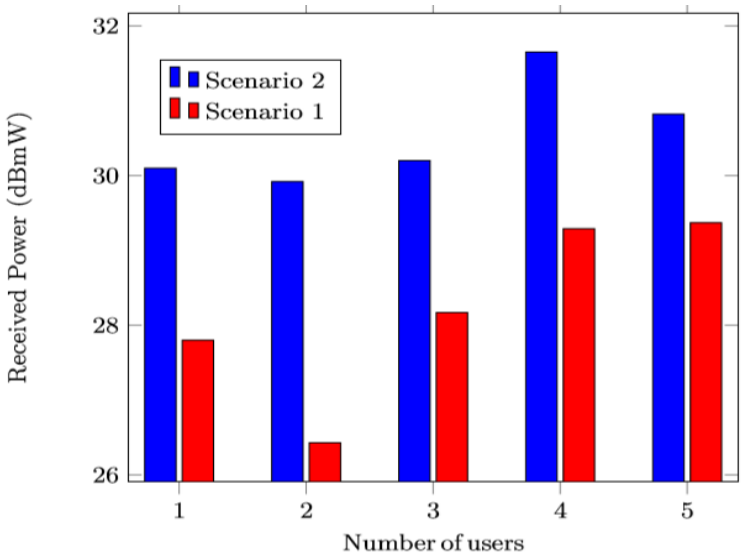
تعداد کاربران	حداکثر	میانگین	حداقل
۱	۲۷.۸	۲۷.۸	۲۷.۸
۲	۲۸.۴	۲۶.۴۳	۲۴.۴۴
۳	۳۲.۲۱	۲۸.۱۷	۲۳.۴۵
۴	۳۵.۰۶	۲۹.۲۹	۲۱.۷۶
۵	۳۶.۲۴	۲۹.۳۷	۱۹.۳۷

۴-۳-۲ سناریو ۲: ارزیابی عملکرد با تعداد کاربران متغیر

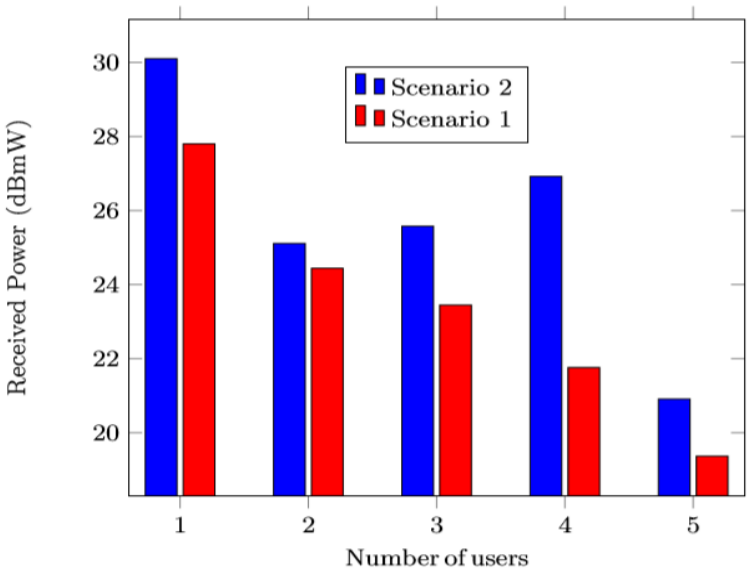
در سناریوی دوم، هیچ آموخته‌ای از قبل وجود ندارد و امواج الکترومغناطیسی منتقل شده باید مسیرهای مناسبی را به مقصد خود پیدا کنند. برای چنین سناریویی، ما عملکرد سیستم را برای تعداد متفاوتی از کاربران ارزیابی می‌کنیم. بر این اساس، تعداد کاربران را افزایش می‌دهیم و آن‌ها را بصورت تصادفی در منطقه NLOS قرار می‌دهیم. مطابق شکل ۴-۴ برای هر مورد، ماکزیمم، میانگین و حداقل توان‌های دریافت شده را محاسبه می‌کنیم و نتایج را با نتایج بدست آمده در اولین سناریو (جدول ۴-۳) مقایسه می‌کنیم.



شکل ۴-۴: الف) حداکثر توان دریافتی برای تعداد متفاوتی از کاربران محاسبه شده توسط سناریو ۲ در مقایسه با موارد به دست آمده در سناریو ۱



شکل ۴-۴: ب) میانگین توان دریافتی برای تعداد متفاوتی از کاربران محاسبه شده توسط سناریو ۲ در مقایسه با موارد به دست آمده در سناریو ۱



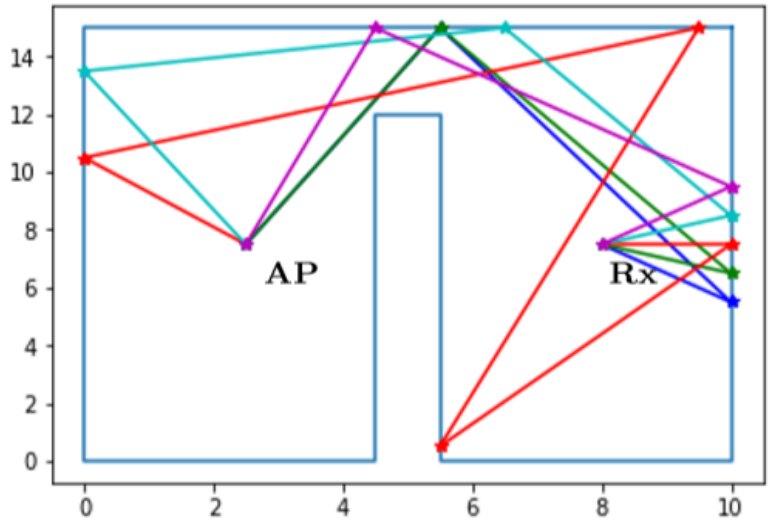
شکل ۴-۴: ج) حداقل توان دریافتی برای تعداد متفاوتی از کاربران محاسبه شده توسط سناریو ۲ در مقایسه با موارد به دست آمده در سناریو ۱

مطابق این شکل‌ها، هنگامی که الگوریتم یادگیری تقویتی برحسب موقعیت خود کاربران اجرا می‌شود (سناریو ۲)، توان دریافتی بیشتر است نسبت به حالتی که وضعیت محیط برحسب گیرنده‌های از پیش تعیین شده، محاسبه می‌شود (سناریو ۱).

علاوه بر این، ما نتایج حاصل از روش ارائه شده بر اساس شبکه عصبی را در مرجع [۱۰] با نتایج بدست آمده در سناریوی ۲ برای کاربر واحد و همچنین انتشار معمولی[1] در یک محیط غیر قابل برنامه‌ریزی مقایسه می‌کنیم. شایان ذکر است که در این مرجع، کاربر می‌تواند چندین سیگنال منعکس شده از کاشی‌های مختلف را دریافت کند. بنابراین برای انجام یک مقایسه عادلانه، ما با فرض قرار دادن کاربر در همان مکان، مطابق شکل ۴-۵، رویکرد خود را اصلاح می‌کنیم و مجموع توان‌های دریافتی توسط این کاربران را به عنوان توان کلی گزارش می‌کنیم.

ما کاربران را در موقعیت (۰.۵، ۷.۵، ۸) در نظر گرفتیم. علاوه بر این، فرستنده در فرکانس ۲.۴ گیگا هرتز با توان انتقال دهنده ۳۰- dBmW کار می‌کند.

[1] Regular



شکل ۴-۵) مسیر امواج الکترومغناطیسی بین یک جفت گیرنده و فرستنده مطابق سناریو ۲

جدول ۴-۴ نشان می‌دهد که رویکرد پیشنهادی یادگیری تقویتی تقریباً سیگنال گزارش شده در [۱۰] را برای چنین مورد خاص فراهم می‌کند و روش یادگیری تقویتی هم‌پای شبکه عصبی پیش رفته است. همچنین زمانی که محیط غیر قابل برنامه‌ریزی و انتشار معمولی است، توان بسیار کمی دریافت می‌شود.

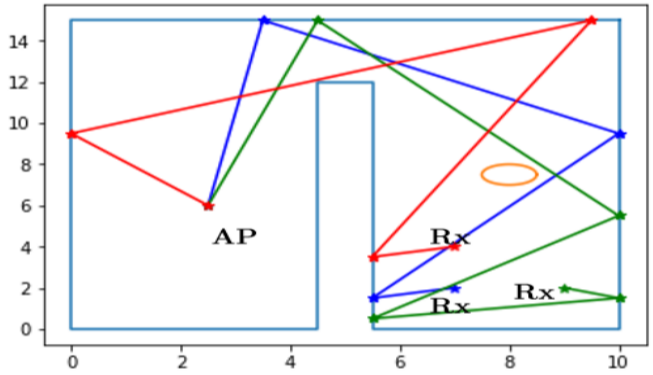
جدول ۴-۴) کل توان دریافت شده توسط یک کاربر واحد برای سه روش مختلف

رویکرد	توان دریافت شده
انتشار معمولی	-۷۰.۷
شبکه عصبی [۱۰]	-۴۹.۸۷
یادگیری تقویتی	-۵۱.۷

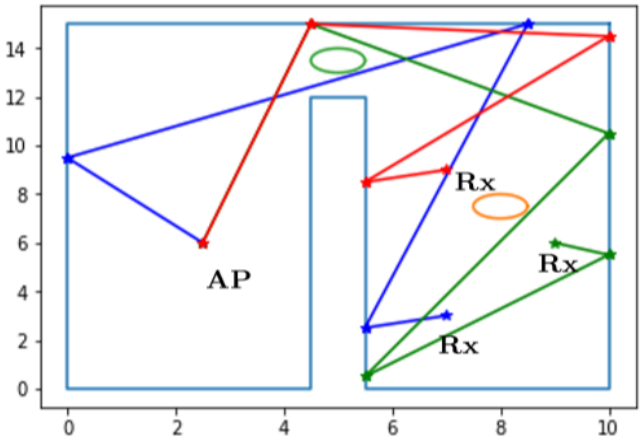
۴-۳-۳ سناریو ۳: ارزیابی عملکرد در صورت وجود موانع

در این زیر بخش ما یک مسئله واقع بینانه‌تر را در نظر می‌گیریم که در آن برخی موانع می‌توانند مسیر امواج الکترومغناطیسی را مسدود کنند. در حالت اول مطابق شکل ۴-۶، عملکرد سیستم

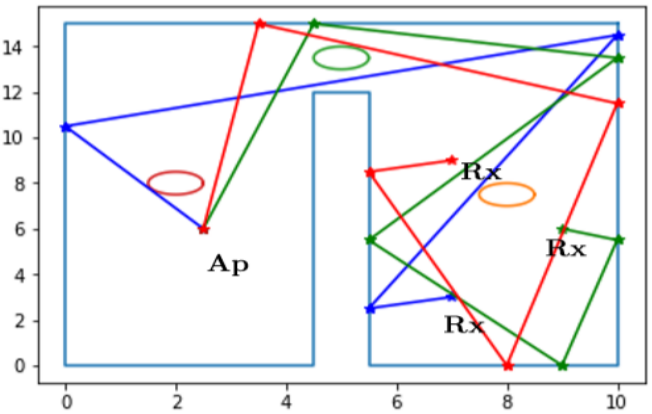
پیشنهادی را با فرض یک، دو و سه مانع موجود در محیط داخلی ارزیابی می‌کنیم. شایان ذکر است، برای هر مورد، شبیه‌سازی را ده بار تکرار کردیم.



شکل ۴-۶: الف) نمونه‌ای از ارتباطات بی‌سیم با حضور یک مانع

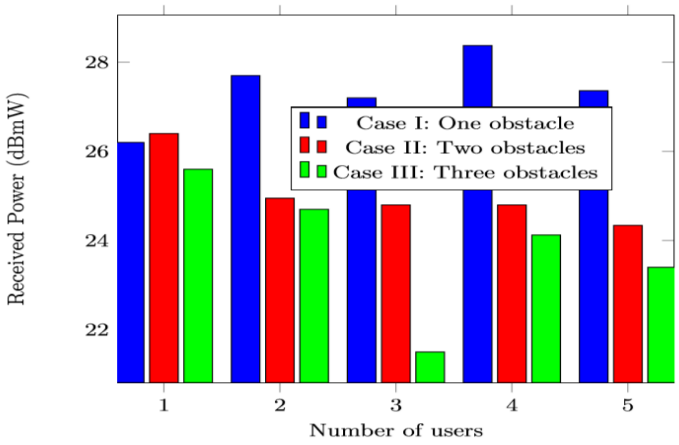


شکل ۴-۶: ب) نمونه‌ای از ارتباطات بی‌سیم با حضور دو مانع



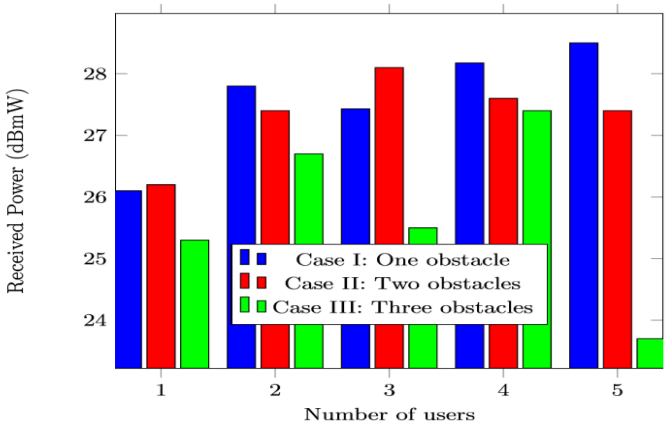
شکل ۴-۶: ج) نمونه‌ای از ارتباطات بی‌سیم با حضور سه مانع

شکل ۴-۷ میانگین توان دریافتی همه موارد برای تعداد متفاوت کاربران را نشان می‌دهد. مطابق این روش، با افزایش تعداد موانع در PWE، میانگین توان دریافتی به‌طور کلی کاهش می‌یابد. با این وجود، حداقل توان دریافتی در حدود ۲۱ dBmW است که در یک محیط داخلی یک مقدار قابل قبول است.



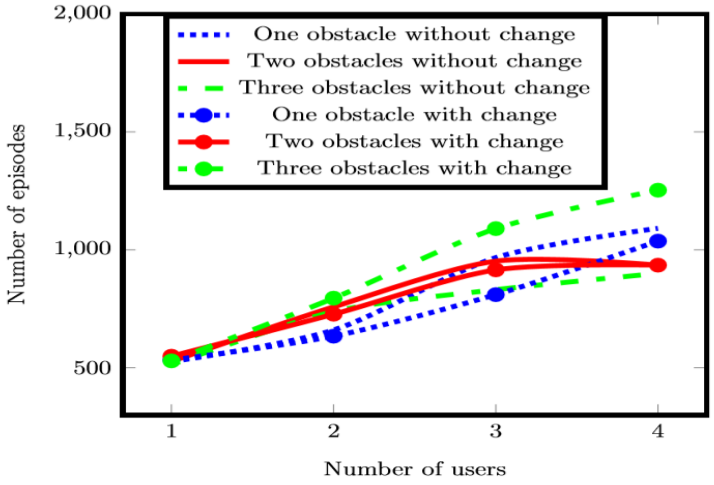
شکل ۴-۷) مقایسه توان توان‌های دریافت شده با وجود یک ، دو و سه مانع

پس از آن، ما توانایی برقراری ارتباطات بی‌سیم بین کاربران و نقطه دسترسی را وقتی ناگهان برخی موانع پس از ۵۰۰ تکرار ظاهر می‌شوند، بررسی می‌کنیم. باز هم، ما سه مورد نشان داده شده در شکل ۴-۶ را با توجه به اینکه هیچ‌گونه مانعی در ابتدای ارتباط وجود ندارد، برای ارزیابی تأثیر تغییرات محیط در نظر می‌گیریم. شکل ۴-۸ میانگین توان دریافتی همه موارد برای تعداد متفاوت کاربران را نشان می‌دهد.



شکل ۴-۸) مقایسه میانگین توان‌های دریافت شده با حضور یک ، دو و سه موانع پس از تغییر محیط

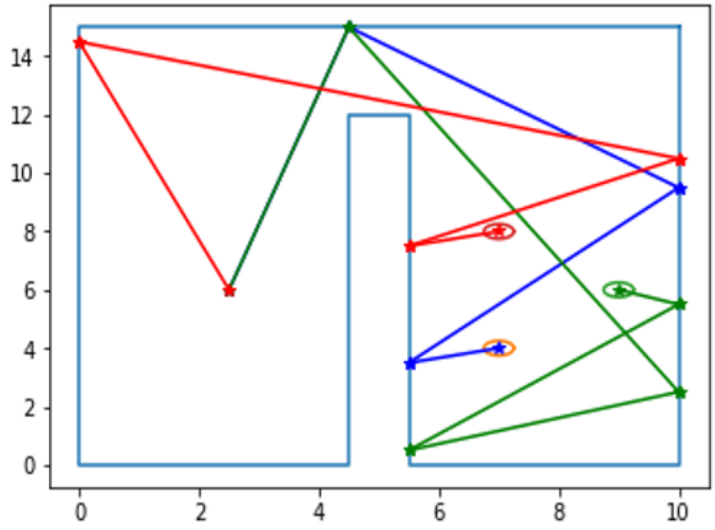
همان‌طور که انتظار داشتیم، برای شرایط جدید رویکرد مبتنی‌بر یادگیری تقویتی، PWE را به درستی برای اتصال نقطه دسترسی و گیرنده‌ها مشخص می‌کند. علاوه بر این، ما تعداد تکرارهای مورد نیاز برای همگرایی الگوریتم خود را برای هر دو مورد، یعنی با و بدون تغییر محیط، مقایسه می‌کنیم، همان‌طور که در شکل ۴-۹ نشان داده شده است. بدیهی است، وقتی محیط به طور ناگهانی تغییر می‌کند، تعداد تکرارهای مورد نیاز افزایش می‌یابد.



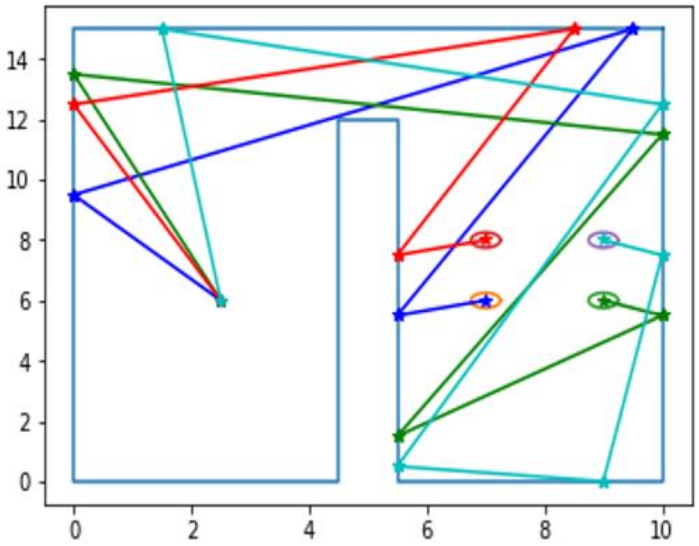
شکل ۴-۹) مقایسه تعداد تکرارهای لازم در الگوریتم با حضور یک ، دو و سه موانع با و بدون تغییر محیط

۴-۳-۴ سناریو ۴: ارزیابی عملکرد با حذف تداخل چند مسیره

در یک ارتباط بی‌سیم معمولی، سیگنال نه تنها از طریق مسیر مستقیم به گیرنده نیز می‌رسد، بلکه نتیجه بازتاب‌ها از اشیا مختلف است. بنابراین، سیگنال کلی در گیرنده برآیند انواع سیگنال‌های دریافتی است. از آن‌جا که آن‌ها طول مسیر متفاوتی را طی می‌کنند و بنابراین با فازهای مختلفی می-رسند، قدرت کلی سیگنال برابر مجموع توان سیگنال‌ها نمی‌باشد و در اکثر موارد باعث خنثی شدن یکدیگر می‌گردند. در این زیر بخش با استفاده از نتایج به دست آمده در سناریو ۳، می‌توانیم بر پدیده چند مسیره‌گی غلبه کنیم. به عبارت دیگر، برای جلوگیری از عبور سایر سیگنال‌ها از روی گیرنده، مانعی را در محل هر کاربر در نظر می‌گیریم. لذا الگوریتم ارائه شده، وضعیتی را انتخاب می‌کند که از محل گیرنده سیگنال‌های دیگری عبور نکنند. شکل ۴-۱۰ برای این سناریو دو نمونه را نشان می‌دهد.

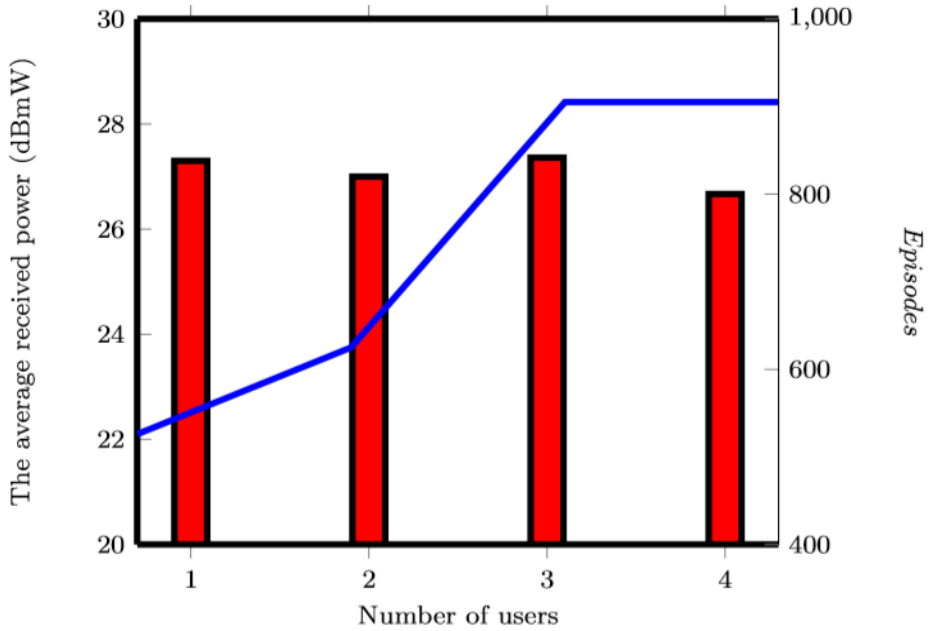


شکل ۴-۱۰: الف) مثال برای سناریو ۴ با سه کاربر



شکل ۴-۱۰: ب) مثال برای سناریو ۴ با چهار کاربر

علاوه بر این، شکل ۴-۱۱ میانگین توان دریافت شده و همچنین تعداد تکرارهای مورد نیاز برای تعداد متفاوت کاربران برای این سناریو را نشان می‌دهد.



شکل ۴-۱۱) میانگین توان دریافتی (نمودار نوار قرمز) و همچنین تعداد تکرارها (منحنی آبی) برای سناریو ۴

۴-۴ پیچیدگی زمانی آزمایش

در نهایت، زمان اجرای هر تکرار را پیش‌بینی می‌کنیم. فرض کنید R نرخ داده ارتباط بی‌سیم، L طول بسته‌های ارسال شده برای به‌روزرسانی جدول Q قبل از مرحله انتقال داده‌ها، d_p کل تأخیر انتشار بین کاربر و AP و d_{proce} پردازش تاخیر در هر کاشی است. فرض کنید سرعت لینک‌های بی‌سیم در حدود ۱۰۰ مگابیت در ثانیه است و ما از بسته‌های با ۱۰۰۰ بیت برای به‌روز رسانی جدول Q به عنوان مرحله اولیه ارتباط استفاده می‌کنیم. اگر تأخیر انتشار را نادیده بگیریم (که فرض قابل قبول برای یک ارتباط داخلی است) و ۱ میکروثانیه برای تأخیر پردازش در هر کاشی در نظر بگیریم، زمان لازم برای رسیدن یک بسته آزمایشی به مقصد از رابطه (۴-۲) محاسبه می‌شود. اگر میانگین تعداد اپیزودهای مورد نیاز برای همگرا شدن الگوریتم RL پیشنهادی، $N_{episode}$ و میانگین تعداد کاشی‌هایی که برای انتقال امواج الکترومغناطیسی برای هر کاربر فعال شده‌اند ، n_{tile} باشد، کل زمان مورد نیاز یک ارتباط طبق این رابطه حدود ۵۵ میلی‌ثانیه خواهد بود. لازم به ذکر است که مطابق شکل ۴-۱، تعداد متوسط کاشی‌هایی که برای پیشبرد امواج الکترومغناطیس فعال شده‌اند را ۵ و تعداد تکرارهای مورد نیاز برای همگرایی الگوریتم را ۱۰۰۰ در نظر می‌گیریم.

$$T = \left[\frac{(n_{tile} + 1)L}{R} + d_p + n_{tile} d_{proce}\right] N_{episode} \quad (۴-۲)$$

۴-۵ نتیجه‌گیری

در این فصل، عملکرد روش پیشنهادی بررسی شد، تا بتواند محیط ارتباطات بی‌سیم داخلی را کنترل و تنظیم کند. همچنین روش پیشنهادی ارائه شده با برخی از روش‌های دیگر در این زمینه مقایسه و نتایج گزارش شد. سناریوهای مختلفی برای ارزیابی اثربخشی رویکرد خود، مانند ارائه پوشش دائمی برای هر کاربر زمانی که تعداد کاربران افزایش می‌یابد، ارتباط در حضور موانع، فراهم کردن اتصال پس

از ظهور ناگهانی برخی موانع و حذف تداخل چند مسیره در نظر گرفتیم. نتایج شبیه‌سازی نشان داد که رویکرد یادگیری تقویتی، میانگین توان دریافتی را تا ۱۲٪ در مقایسه با کار گذشته که از الگوریتم ژنتیک برای ایجاد محیط قابل برنامه‌ریزی استفاده کرد [۱۷]، بهبود داده است. در کار گذشته، میانگین توان دریافتی برای ۱۲ گیرنده، ۲۵.۳۸ dBmW می‌باشد.

همچنین در پایان این فصل، خلاصه‌ای از آزمایشات و نتایج بدست آمده در جدول ۴-۵ ارائه شده است.

جدول ۴-۵) خلاصه آزمایش‌ها و نتایج بدست آمده

سناریوها	اهداف	نتایج
سناریو ۱: تامین پوشش منطقه NLOS	مقایسه روش پیشنهادی RL با رویکرد GA	عملکرد مناسب روش RL
سناریو ۱: توانایی پوشش توسط جدول Q پیشین برای تعداد متفاوت کاربران	محاسبه مقادیر توان دریافتی توسط کاربران مطابق با روش پیشنهادی RL	میانگین مقادیر توان‌های دریافت شده بیشتر از ۲۶.۴۳ dBmW است، که بدیهی است برای برقراری ارتباط مناسب است.
سناریو ۲: ارزیابی عملکرد با تعداد کاربران متغیر	مقایسه با سناریو۱ (توانایی پوشش توسط جدول Q پیشین برای تعداد متفاوت کاربران)	توان دریافتی بیشتر نسبت به سناریو ۱
سناریو ۲: محاسبه مجموع توان‌های دریافت شده برای ۵ کاربر در یک مکان	مقایسه نتایج حاصل از روش شبکه عصبی [۱۰] با نتایج بدست آمده در سناریوی ۲ مطابق روش RL و همچنین انتشار معمولی	روش RL تقریباً توان گزارش شده در [۱۰] را فراهم کرده است.
سناریو ۳: ارزیابی عملکرد در صورت وجود موانع	مسدود کردن مسیر امواج الکترومغناطیسی و محاسبه میانگین توان دریافتی برای تعداد متفاوت کاربران	عدم برخورد امواج با موانع مطابق با روش RL و همچنین کاهش توان دریافتی با افزایش تعداد موانع
سناریو ۳: ظهور موانع ناگهانی	توانایی برقراری ارتباطات بی‌سیم بین کاربران و نقطه دسترسی پس از ۵۰۰ تکرار و محاسبه میانگین توان دریافتی برای تعداد متفاوت کاربران	برقراری ارتباط مناسب مطابق با روش RL و همچنین کاهش توان دریافتی با افزایش تعداد موانع
سناریو ۴: ارزیابی عملکرد با حذف تداخل چند مسیره	مقابله با پدیده محوشدگی و محاسبه میانگین توان دریافت شده	عبور امواج تنها از یک مسیر

فصل ۵: نتیجه‌گیری و جمع‌بندی

۵-۱ مقدمه

در این فصل، الگوریتم ارائه شده را جمع‌بندی می‌کنیم و همچنین پیشنهادی برای کارهای آتی ارائه می‌دهیم.

۵-۲ جمع‌بندی و نتیجه‌گیری

در این پایان‌نامه ما یک رویکرد مبتنی‌بر آموزش یادگیری تقویتی پیشنهاد کردیم تا بتوانیم محیط ارتباطی بی‌سیم داخلی را پیکربندی کنیم. رویکرد پیشنهادی می‌تواند برای یک مسئله چند کاربر و همچنین یک محیط در حال تغییر اعمال شود. برای ارزیابی عملکرد روش، چهار سناریوی مختلف یعنی تأمین پوشش دائمی برای منطقه NLOS، فراهم آوردن ارتباط خاص برای تعداد متفاوتی از کاربران، فراهم آوردن اتصال در صورت وجود موانع و بطور خاص، فراهم آوردن اتصال پس از ظهور ناگهانی برخی موانع را در نظر گرفتیم. ارزیابی نشان می‌دهد که رویکرد مبتنی‌بر یادگیری تقویتی می‌تواند بطور صحیح لینک‌های بی‌سیم را برای کلیه سناریوهای فوق با سطح سیگنال کافی فراهم کند.

۵-۳ پیشنهاد برای کارهای آتی

در یادگیری تقویتی محدودیت‌هایی وجود دارد که با وجود این محدودیت‌ها برخی از مسائل که خیلی از مسائل دنیای واقعی را شامل می‌شوند، با این تکنیک قابل حل نیستند و مجبور هستیم از تکنیک-هایی استفاده کنیم که آن محدودیت‌ها را رفع کنند. در یادگیری تقویتی جدول Q می‌سازیم که مبتنی‌بر هر عمل و وضعیت یک ارزش Q تعریف می‌شود که نشان‌دهنده این است که اگر عامل در یک وضعیت یک عملی را انجام دهد در مسیرش چه مقدار پاداش ممکن است بگیرد. برای جدول Q باید تعداد اعمال و حالت‌ها محدود باشند. با افزایش تعداد اعمال و حالت‌ها، دیگر به‌ازای هر حالت و عمل نمی‌توان مقدار ارزش Q را در جایی ذخیره کرد.

به عنوان کار آینده، ما قصد داریم که به جای الگوریتم یادگیری تقویتی معمولی، از یادگیری تقویتی عمیق[1] استفاده کنیم تا رویکرد خود را برای ایجاد محیط‌های پیچیده‌تر گسترش دهیم [۵۹-۶۲]. برای این منظور می‌توانیم شناسه‌های کاشی‌های فعال شده و کاشی فعلی و همچنین مکان نقطه دسترسی را به عنوان ورودی شبکه عصبی در نظر بگیریم. از طرف دیگر، خروجی شبکه می‌تواند زاویه بازگرداندن (عملکرد مناسب) را برای پیشبرد امواج الکترومغناطیسی به کاشی بعدی تعیین کند.

همچنین الگوریتم‌های یادگیری تقویتی، از زمان همگرایی کند رنج می‌برند، به عنوان کار آینده دیگر، قصد داریم از یادگیری ماشین کوانتومی (QML[2]) استفاده نماییم که از مفاهیم کوانتوم برای تسریع در یادگیری استفاده می‌کند [۶۳]. این روش برای هوشمند کردن SDM مورد توجه قرار می-گیرد، زیرا نوید بسیاری برای کاربردهای مختلف در شبکه‌های ارتباطی می‌دهد [۶۳]. همچنین QML مزایای مقیاس پذیری، آموزش سریع، سرعت استنتاج و قابلیت اطمینان را در مقایسه با روش‌های معمول یادگیری ماشین ارائه می دهد. قابل ذکر است، تقاطع بین ارتباطات کوانتومی و شبکه‌های عصبی عمیق یک منطقه جالب است که محققان شروع به کشف آن کرده‌اند و دارای دامنه قابل توجهی برای باز کردن پتانسیل کامل SDM است.

[1] Deep reinforcement learning
[2] Quantum machine learning

مراجع

[1] Groth, David. (2005), Network+ Study Guide, Fourth Edition, Toby Skandier, Sybex, Inc, ISBN 0-7821-4406-3

[2] Karney, James. (2000), A+ Certification Training Kit, Second Edition, Redmond, Washington: Microsoft Press, ISBN 0-7356-1109-2

[3] Tse D, Viswanath P. (2005) "Fundamentals of Wireless Communication", Cambridge University Press, pp. 1-4

[4] Dekleva S, Shim J. P, Varshney U, Knoerzer G. (2007) "Evolution and emerging issues in mobile wireless networks", Communications of the ACM, ACM, Vol. 50, No. 6, p. 38-43, ISSN 0001-0782

[5] Miao G, Zander J, Sung K. W and Slimane B. (2016) "Fundamentals of Mobile Data Networks", Cambridge University Press, ISBN 1-107-14321-7.

[6] Letaief K. B, Chen W, Shi Y, Zhang J and Zhang Y.J. A. (2019) "The roadmap to 6g: Ai empowered wireless networks", IEEE Communications Magazine, vol. 57, no. 8, pp. 84–90.

[7] Agiwal M, Roy A and Saxena N. (2016) "Next generation 5g wireless networks: A comprehensive survey", IEEE Communications Surveys & Tutorials, vol. 18, no. 3, pp. 1617–1655.

[8] Niu Y, Li Y, Jin D, Su L and Vasilakos A. V. (2015) "A survey of millimeter wave communications (mmwave) for 5g: opportunities and challenges", Wireless networks, vol. 21, no. 8, pp. 2657–2676.

[9] Tariq F, Khandaker M, Wong K.K, Imran M, Bennis M and Debbah M. (2019) "A speculative study on 6G," arXiv: 1902.06700.

[10] Liaskos C, Tsioliaridou A, Nie S, Pitsillides A, Ioannidis S and Akyildiz I. (2019) "An interpretable neural network for configuring programmable wireless environments", in 2019 IEEE 20th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC). IEEE, pp. 1–5.

[11] Welkie A, Shangguan L, Gummeson J, Hu W and Jamieson K. (2017) “Programmable radio environments for smart spaces”, in Proceedings of the 16th ACM Workshop on Hot Topics in Networks, pp. 36–42.

[12] Chandra R and Winstein K. (2017) “Programmable radio environments for smart spaceshotnets-xvi dialogue”, in ACM Workshop on Hot Topics in Networks, Palo Alto.

[13] Sheikh M. A and Khan S. A. (2017) “Performance optimization of unit-cell reflectarray antenna for future 5g communications”, International Journal of Computer Applications, vol. 975, p. 8887.

[14] Subrt L and Pechac P. (2012) “Intelligent walls as autonomous parts of smart indoor environments”, IET communications, vol. 6, no. 8, pp. 1004–1010.

[15] Tan X, Sun Z, Jornet J. M and Pados D. (2016) “Increasing indoor spectrum sharing capacity using smart reflect-array”, in 2016 IEEE International Conference on Communications (ICC). IEEE, pp. 1–6.

[16] Wu Q and Zhang R. (2019) “Beamforming optimization for intelligent reflecting surface with discrete phase shifts”, in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 7830–7833.

[17] Liaskos C, Nie S, Tsioliaridou A, Pitsillides A, Ioannidis S and Akyildiz I. (2019) “A novel communication paradigm for high capacity and security via programmable indoor wireless environments in next generation wireless systems”, Ad Hoc Networks, vol. 87, pp. 1–16.

[18] Taghvaee H, Cabellos-Aparicio A, Georgiou J and Abadal S. (2020) “Error analysis of programmable metasurfaces for beam steering”, IEEE Journal on Emerging and Selected Topics in Circuits and Systems.

[19] Tang W, Li X, Dai J. Y, Jin S, Zeng Y, Cheng Q and Cui T. J. (2019) “Wireless communications with programmable metasurface: Transceiver

design and experimental results", China Communications, vol. 16, no. 5, pp. 46–61.

[20] Liaskos C, Tsioliaridou A, Nie S, Pitsillides A, Ioannidis S and Akyildiz I. F. (2019) "On the network-layer modeling and configuration of programmable wireless environments", IEEE/ACM Transactions on Networking, vol. 27, no. 4, pp. 1696–1713.

[21] Fairweather G, Karageorghis A and Martin P. A. (2003) "The method of fundamental solutions for scattering and radiation problems", Engineering Analysis with Boundary Elements, vol. 27, no. 7, pp. 759–769.

[22] Huang C, Zappone A, Alexandropoulos G. C, Debbah M and Yuen C. (2019) "Reconfigurable intelligent surfaces for energy efficiency in wireless communication," IEEE Trans. Wireless Commun., vol. 18, no. 8, pp. 4157–4170, Aug.

[23] Zhang L, Chen X. Q, Liu S, Zhang Q, Zhao J, Dai J. Y, Bai G. D, Wan X, Cheng Q, Castaldi G et al. (2018) "Space-time-coding digital metasurfaces," Nat. Commun., vol. 9, no. 1-11, Oct.

[24] Song M, Baryshnikova K, Markvart A, Belov P, Nenasheva E, Simovski C and Kapitanova P. (2019) "Smart table based on a metasurface for wireless power transfer," Phys. Rev. Applied, vol. 11, p. 054046.

[25] Mohjazi, Lina, Zoha A, Bariah L, Muhaidat S, Paschalis C. Sofotasios, Imran M.A and Octavia A. Dobre. (2020) "An Outlook on the Interplay of AI and Software-Defined Metasurfaces." arXiv preprint arXiv: 2004.00365.

[26] Liaskos, Christos, Nie S, Tsioliaridou A, Pitsillides A, Ioannidis S and Akyildiz I. (2018) "A new wireless communication paradigm through software-controlled metasurfaces."IEEE Communications Magazine 56, no. 9: 162-169.

[27] Simeone, Osvaldo. (2018) "A very brief introduction to machine learning with applications to communication systems." IEEE Transactions on Cognitive Communications and Networking 4, no: 648-664.

[28] Zhang C, Patras P and Haddadi H. (2019) “Deep learning in mobile and wireless networking: A survey,” IEEE Commun. Surveys Tuts, vol. 21, no. 3, pp. 2224–2287.

[29] Huang C, Alexandropoulos G. C, Yuen C and Debbah M. (2019) “Indoor signal focusing improvement via deep learning configured intelligent metasurfaces,” in Proc. IEEE Int. Workshop. Signal Process. Advances in Wireless Commun. Cannes, France.

[30] Zappone A, Di Renzo M and Debbah M. (2019) “Wireless networks design in the era of deep learning: Model-based, AI-based, or both?” arXiv preprint arXiv: 1902.02647.

[31] Liaskos C, Tsioliaridou A, Nie S, Pitsillides A, Ioannidis S and Akyildiz I. (2019) “An interpretable neural network for configuring programmable wireless environments,” arXiv:1905.02495.

[32] Abadal S, Liaskos C, Tsioliaridou A, Ioannidis S, Pitsillides A, Sol´e-Pareta J, Alarco´n E and Cabellos-Aparicio A. (2017) “Computing and communications for the software-defined metamaterial paradigm: A context analysis”, IEEE access, vol. 5, pp. 6225–6235.

[33] Subrt L, Pechac P. (2012) “Intelligent walls as autonomous parts of smart indoor environments”, IET communications, vol. 6, no. 8, pp. 1004–1010.

[34] Di Renzo M, Debbah M, Phan-Huy D.-T, Zappone A, Alouini M.-S, Yuen C, Sciancalepore V, Alexandropoulos G. C, Hoydis J, Gacanin H, et al. (2019) “Smart radio environments empowered by reconfigurable ai meta-surfaces: an idea whose time has come”, EURASIP Journal on Wireless Communications and Networking, vol, no. 1, pp. 1–20.

[35] Tsioliaridou A, Liaskos C, Pitsillides A and Ioannidis S. (2017) “A novel protocol for network-controlled metasurfaces”, in ACM NANOCOM’17, NanoCom ’17, (New York, NY, USA), pp. 3:1–3:6, ACM.

[36] Kober, Jens, Andrew Bagnell J and Peters J 11. (2013) "Reinforcement learning in robotics: A survey." The International Journal of Robotics Research 32, no.1238-1274.

[37] Matarić, Maja J. (1997) "Reinforcement learning in the multi-robot domain." In Robot colonies, Springer, Boston, MA, pp. 73-83.

[38] Johannink, Tobias, Bahl S, Nair A, Luo J, Kumar A, Loskyll M, Aparicio Ojea J, Solowjow E and Levine S. (2019) "Residual reinforcement learning for robot control." In 2019 International Conference on Robotics and Automation (ICRA), IEEE, pp. 6023-6029.

[39] Riedmiller, Martin, Gabel T, Hafner R and Lange S. (2009) "Reinforcement learning for robot soccer." Autonomous Robots 27, no. 1: 55-73.

[40] Sutton, Richard S. (1992) "Introduction: The challenge of reinforcement learning." In Reinforcement Learning, Springer, Boston, MA, pp. 1-3.

[41] Mahadevan, Sridhar. (1996) "Average reward reinforcement learning: Foundations, algorithms, and empirical results." Machine learning 22, no. 1-3: 159-195.

[42] Lauer, Martin and Riedmiller M. (2000) "An algorithm for distributed reinforcement learning in cooperative multi-agent systems." In In Proceedings of the Seventeenth International Conference on Machine Learning.

[43] Sun, Ron, and Peterson T. (1999) "Multi-agent reinforcement learning: weighting and partitioning." Neural networks 12, no. 4-5. 727-753.

[44] Pettinger, James E and Richard M. Everson. (2002) "Controlling genetic algorithms with reinforcement learning." In Proceedings of the 4th Annual Conference on Genetic and Evolutionary Computation, pp. 692-692.

[45] Moriarty, David E, Alan C. Schultz and John J. Grefenstette. (1999) "Evolutionary algorithms for reinforcement learning." Journal of Artificial Intelligence Research 11. 241-276.

[46] Watkins C. J and Dayan P. (1992) “Q-learning”, Machine learning, vol. 8, no. 3-4, pp. 279– 292.

[47] Busoniu, Lucian, Babuska R and De Schutter B. (2008) "A comprehensive survey of multiagent reinforcement learning." IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 38, no. 2: 156-172.

[48] Buşoniu, Lucian, Babuška R and De Schutter B. (2010) "Multi-agent reinforcement learning: An overview." In Innovations in multi-agent systems and applications-1, Springer, Berlin, Heidelberg, pp. 183-221.

[49] Bu L ¸soniu, Babu˘ska R and De Schutter B. (2010) “Multi-agent reinforcement learning: An overview”, in Innovations in multi-agent systems and applications-1. Springer, pp. 183–221.

[50] Shoham Y, Powers R and Grenager T. (2003) “Multi-agent reinforcement learning: a critical survey”, Web manuscript.

[51] Tuyls11, Karl and Nowé A. (2005) "Evolutionary game theory and multi-agent reinforcement learning".

[52] Myerson, Roger B. (1999) "Nash equilibrium and the history of economic theory." Journal of Economic Literature 37, no. 3: 1067-1082.

[53] Kuhn, Harold W, Harsanyi J. C, Selten R, Weibul J. W and van Damme E. E. C. (1996) "The work of John Nash in game theory." journal of economic theory 69, no. 1: 153-185.

[54] Holt, Charles A and Roth Alvin E. (2004) "The Nash equilibrium: A perspective." Proceedings of the National Academy of Sciences 101, no. 12: 3999-4002.

[55] Roughgarden, Tim. (2010) "Algorithmic game theory." Communications of the ACM 53, no. 7: 78-86.

[56] Nowé, Ann, Vrancx P and De Hauwere Y-M. (2012) "Game theory and multi-agent reinforcement learning", In Reinforcement Learning, Springer, Berlin, Heidelberg, pp. 441-470

[57] Kaelbling L. P, Littman M. L and Moore A. W. (1996) “Reinforcement learning: A survey”, Journal of artificial intelligence research, vol. 4, pp. 237–285.

[58] Zhang C and Zheng Z. (2019) “Task migration for mobile edge computing using deep reinforcement learning”, Future Generation Computer Systems, vol. 96, pp. 111–118.

[59] Mnih, Volodymyr, Kavukcuoglu K, Silver D, Rusu A. A, Veness J, Bellemare M. G, Graves A, et al. (2015) "Human-level control through deep reinforcement learning." nature 518, no. 7540, 529-533.

[60] Henderson, Peter, Islam R, Bachman P, Pineau J, Precup D and Meger D. (2017) "Deep reinforcement learning that matters." arXiv preprint arXiv: 1709.06560.

[61] Li, Yuxi. (2017) "Deep reinforcement learning: An overview." arXiv preprint arXiv: 1701.07274.

[62] Arulkumaran, Kie, Deisenroth M.P, Brundage M and Bharath A. (2017) "Deep reinforcement learning: A brief survey." IEEE Signal Processing Magazine 34, no. 6: 26-38.

[63] Oneto L, Ridella S and Anguita D. (2017) “Quantum computing and supervised machine learning: Training, model selection, and error estimation,” in Quantum Inspired Computational Intelligence. Elsevier, pp. 33–83.

Abstract

With the increasing use of smart wireless devices and the advent of the IoT idea, wireless communication systems are faced higher data rates as well as better quality services. To achieve such a high data, the wireless networks should be work at high frequencies, such as millimeter and THz bands. On the other hand, with increasing frequency, the waves are not able to pass through obstacles such as walls and are lost in the environment that cause less penetration depth. Multi-path fading also limits the use of such high frequencies of electromagnetic waves due to the interaction of signals from several different paths in the air from the transmitter to the receiver. To control this, the researchers came up with the idea of covering objects in the environment with Software-Defined Metasurface to turn publishing environments into a programmable environments. In this method, all the walls are covered with pieces of the same size from the Metasurface, called tiles. The main challenge is that the environment can change during the communication, for example a sudden obstacle blocks the signal path, this management must be adaptive. This paper, the use of Reinforcement Learning (RL) algorithm to create such an environment is presented. Since an agent in the RL algorithm gradually interacts with the environment, it can be better decision, especially when the environment changes. The results of the proposed method are compared with the previous work, which used a genetic algorithm to create a programmable environment. In the previous work, the average of the received power for 12 receivers is 25.38 dBmW and the reinforcement learning method in this paper, increases the average of the received power by 12%.

Keywords: Software-Defined Metasurface (SDM), Programmable Wireless Environment (PWE), Reinforcement Learning (RL), Agent, Electromagnetic waves.

Shahrood University of Technology

Faculty of Computer Engineering

M.Sc. Thesis in Artificial Intelligence Engineering

Configuring wireless networks including Software Define Material using reinforcement learning

By: Fatemeh Aliannezhadi

Supervisor:
Dr. Esmaeel Tahanian

Advisor:
Dr. Mansoor Fateh
Dr. Mohsen Rezvani

October 2020