تسهیم طیف برای مقابله با تداخلگر با استفاده از یادگیری عمیق تقویتی چندعاملی

ندا کاظمی[1] و معصومه آذغانی[2]

[1] دانشجوی دکترا، گروه مخابرات، آزمایشگاه مخابرات بی سیم و پردازش سیگنال (WCSP) ، دانشکده مهندسی برق، دانشگاه صنعتی سهند، تبریز، n_kazemi97@sut.ac.ir

[2] **نویسنده مسئول**، دانشیار، گروه مخابرات، آزمایشگاه مخابرات بی سیم و پردازش سیگنال (WCSP) ، دانشکده مهندسی برق، دانشگاه صنعتی سهند، تبریز، mazghani@sut.ac.ir

دانشگاه صنعتی سهند

DOR: 20.1001.1.23223146.1401.9.2.1.9

نشریه سامانه‌های غیرخطی در مهندسی برق

دوره ۱۰ – شماره ۲

پاییز و زمستان ۱۴۰۲

صفحات ۵۵ الی ۷۳

ISSN: 2322-3146

http://journals.sut.ac.ir/jnsee

چکیده

در تسهیم طیف بین کاربران مختلف، مقابله با تداخلگر موجود در شبکه بسیار حیاتی است. اهمیت این کار به ویژه زمانی دوچندان می‌شود که تداخلگر، هوشمند بوده و قادر به یادگیری الگوهای ارتباطی بین کاربران باشد. در این مقاله، روشی برای تسهیم طیف با هدف مقابله با تداخلگر شبکه پیشنهاد می‌گردد. شیوه پیشنهادی بر پایه یادگیری تقویتی چندعاملی است. در مرحله اول، طیف‌های مورد حمله تداخلگر توسط ایستگاه پایه به دست می‌آیند. در مرحله بعد، یک چارچوب یادگیری تقویتی عمیق طراحی می‌شود که به هرکدام از کاربران طیف امن و مناسب را پیشنهاد می‌دهد. برای این منظور از همبستگی بین سیگنال راهنمای ارسالی و سیگنال دریافتی در هر بازه فرکانسی، میزان تاثیر تداخلگر در آن زیرباند حاصل می‌شود. سپس ضریب تخفیف الگوریتم یادگیری تقویتی بر پایه مقدار همبستگی به صورت وفقی تنظیم می‌گردد. به این ترتیب، زیرباندهایی که کمتر تحت تاثیر تداخلگر واقع شده اند، به کاربران اختصاص می‌یابند. روش پیشنهادی در سناریوهای مختلف مورد شبیه‌سازی و ارزیابی قرار گرفته و نتایج حاکی از آن است که نرخ مجموع شبکه با سرعت بالایی به نرخ مطلوب همگرا می‌شود و شیوه پیشنهادی مقاومت خوبی در برابر تداخلگر از خود نشان می‌دهد.

واژه‌های کلیدی

سیستم‌های بی‌سیم،

یادگیری تقویتی چند عاملی،

تسهیم طیف،

ارتباطات امن،

روش‌های ضد اختلال‌گر.

Sahand University of Technology

DOR: 20.1001.1.23223146.1401.9.2.1.9

Journal of Nonlinear Systems in Electrical Engineering

Vol.10, No.2

Autumn and Winter 2023

ISSN: 2322 – 3146

http://journals.sut.ac.ir/jnsee

Spectrum Sharing in Anti-jamming Using Multi-agent Reinforcement Learning

Neda Kazemi[1] and Masoumeh Azghani [2]

[1]Ph.D student, Laboratory of Wireless Communication and Signal Processing (WCSP), Faculty of Electrical Engineering, Sahand University of Technology, Tabriz, n_kazemi97@sut.ac.ir

[2] **Corresponding Author**, Laboratory of Wireless Communication and Signal Processing (WCSP), Faculty of Electrical Engineering, Sahand University of Technology, Tabriz, mazghani@sut.ac.ir

ABSTRACT

Keywords

Wireless systems,

Multi-agent reinforcement learning,

Spectrum sharing,

Secure communication,

Anti-jamming methods.

In spectrum sharing among different users, anti-jamming process in the network is very critical. The importance of this work doubles especially when the jammer is intelligent and able to learn patterns of communications between users. In this article, a method for spectrum sharing is proposed with the aim of dealing with the jammer. The proposed method is based on multi-agent reinforcement learning. In the first stage, the jammed slots of spectrum are acquired by the base station. In the next stage, a deep reinforcement learning framework is designed that offers each user a safe and appropriate range. For this purpose, the influence of the jammer in that sub-band is obtained from the correlation between the transmitted pilot signal and the received signal in each frequency range. Then, the discount factor of the reinforcement learning algorithm is adjusted based on the correlation value. In this way, the sub-bands that are less affected by the jammer are allocated by the users. The proposed method has been simulated and evaluated in different scenarios and the results indicate that the total rate of the network converges to the desired rate at a high speed and the proposed method shows good resistance against the jammer.

۱- مقدمه

انفجار اطلاعات و رشد ترافیک داده و تجاری‌سازی شبکه‌های ارتباطی بی‌سیم نسل پنجم[۱](5th)، چالش‌های بزرگی در زمینه استفاده از طیف و امنیت داده‌ها تحمیل می‌کند. اقدامات ضد اختلالگر[۲] برای خنثی نمودن تهاجم اطلاعاتی، یک موضوع بسیار مهم در ارتباطات است. مخصوصا در ارتباطات بی‌سیم، اختلال‌گر خارجی، کیفیت لینک ارتباطی را به شدت تهدید می‌کند. در سال های اخیر، با توسعه هوش مصنوعی، مقابله هوشمند با حملات اختلال‌گرها بسیار مرسوم شده است [۱]، [۲]. در پژوهش‌های پیشین، تعاملات بین کاربران قانونی[۳] و تداخلگر[۴]ها به شکل تئوری بازی ها مدل‌سازی شده‌است [۳]–[۵]. بیشتر پژوهش‌های ضداختلال‌گر موجود، نیاز به دستیابی به اطلاعات قبلی از الگوی سیگنال ارسالی اختلال‌گر در فرمول‌بندی مربوط به استراتژی خود دارند. دو چالش مهم در این زمینه وجود دارد، چالش اول اینکه امکان دارد اطلاعات به دست‌آمده از تداخلگر برای تخمین الگوی آن کافی نباشد و کاربر برای اقدام ضد تداخلگر با کمبود اطلاعات مواجه شود اما چالش دوم زمانی به وجود می‌آید که اقدام ضد تداخلگر در یک محیط پویا اتفاق بیافتد و تداخلگر مدام الگوی خود را تغییر دهد. این درحالی است که با توسعه سریع هوش مصنوعی تداخلگرها به راحتی می‌توانند حملات هوشمند و پویا ایجاد کنند [۶]. برای غلبه بر این چالش‌ها، روش پیشنهادی در این مقاله برای عملیات ضد اختلال‌گری خود، نیازی به دستیابی به الگوی سیگنال اختلالگر ندارد. علاوه بر این، عملیات ضداختلال‌گری در فرستنده و قبل از ارسال اطلاعات انجام می‌شود. بسیاری از پژوهشگران الگوریتم‌های یادگیری ماشین را برای تصمیم‌گیری‌های ضداختلال‌گر آنلاین پیشنهاد کرده‌اند [۷]. یادگیری تقویتی (RL[۵]) ، یک روش مناسب برای تصمیم‌گیری در محیط‌های پویا است. در پژوهش‌های پیشین، یادگیری تقویتی به عنوان یک الگوریتم مناسب در اقدامات ضد تداخلگر به کار گرفته شده است [۸]–[۱۰]. یادگیری تقویتی شاخه ای از یادگیری ماشین[۶] است که در آن یک یا چند عامل با یک محیط معین تعامل می‌کنند و خط مشی[۷] بهینه را برای دستیابی به هدف از پیش تعریف شده می‌آموزند. تا به امروز، روش‌های یادگیری تقویتی بسیاری توسعه پیدا کرده‌اند [۱۱]–[۱۳]. این روش‌ها در مسایل مختلفی از جمله تخصیص منابع و طیف مورد استفاده قرار گرفته‌اند [۱۴]–[۱۷]. اما بسیاری از مسائل دنیای واقعی، فضای حالت بزرگ با پاداش تاخیری دارند. ساختار ابتدایی روش‌های یادگیری تقویتی به صورت کامل بهینه نبوده و فرآیند یادگیری آن ها بسیار کند است. یادگیری تقویتی عمیق روشی است که از شبکه های عصبی برای تخمین تابع ارزش[۸] در یادگیری تقویتی استفاده می کند [۱۸]. در یادگیری تقویتی تک‌عاملی[۹] (SARL)، یک عامل[۱۰]، یک اقدام[۱۱] مشخص را در یک وضعیت معین مطابق با خط مشی خود اتخاذ می‌کند و اقدام مربوطه بر روی محیط اعمال می شود. سپس عامل، حالت بعدی را بر مبنای پاداش مربوطه از محیط کسب می‌کند. برای محاسبه عمل بعدی، عامل مجموع

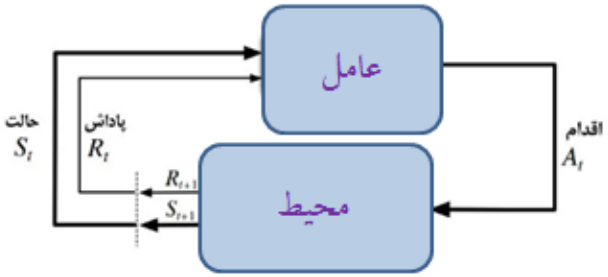
[1] 5th Generation
[2] Anti-jamming
[3] Legitimate users
[4] jammer
[5] Reinforcement learning
[6] Machine learning
[7] policy
[8] Value function
[9] Single agent reinforcement learning
[10] agent
[11] action

سیگنال‌های پاداش[1] تخفیف‌یافته را در یک افق محدود یا نامتناهی به حداکثر می‌رساند. در یادگیری تقویتی چندعاملی[2] (MARL) ، چند عامل با اشتراک‌گذاری پاداش‌های خود با یک تصمیم‌گیری جمعی، اقدام بعدی خود را تعیین می‌کنند [۱۹]. از آنجایی که پاداش آینده ناشناخته است، RL عمیق از شبکه های عصبی برای تخمین انتظارات به نام تابع ارزش برای آینده استفاده می کند. الگوریتم‌های یادگیری تقویتی مرسوم را فقط می‌توان در تداخلگرهای با نظم و مدل مشخص اعمال کرد و اگر حالت تداخلگر خارجی تغییر کرد، نیاز به آموزش مجدد برای رسیدن به همگرایی خواهند داشت [۱], [۲۰]. با توجه به اینکه تابع ارزش توسط پاداش و ضریب تخفیف محاسبه می شود، انتخاب صحیح آنها از اهمیت بالایی برخوردار است. در محیطی که عدم قطعیت وجود دارد، اگر تجربه به دست آوردن پاداش بدون در نظر گرفتن عدم قطعیت ارزیابی شود، می‌توان مجموعه را نسبت به عدم قطعیت هم آموزش داد. بنابراین، روش های یادگیری تقویتی برای تخمین توزیع پاداش تجمیعی معرفی شده است. به طور خاص، تمرکز این مقاله بر روی ضریب تخفیف است که مقدار آن در پژوهش‌های موجود اغلب ثابت است. اگر ضریب تخفیف در محیطی با تغییرات سریع نسبت به زمان، مقدار بزرگی انتخاب شود، تاثیر مقادیر پاداش آینده در پاداش فعلی زیاد خواهد بود و به همین دلیل تخمین مجموع پاداش ها در یک دوره زمانی طولانی‌تر به دلیل عدم اطمینان زیاد منجر به بروز موقعیت های غیرمنتظره می شود. به عنوان مثال، عوامل می‌توانند در آینده دور پاداش ها را بیش از حد برآورد کنند، یا ممکن است همگرایی تابع ارزش با مشکل روبرو شود [۲۱]. از سوی دیگر، اگر ضریب تخفیف خیلی کوچک باشد، تاثیر پاداش آینده در تابع پاداش بسیار کم می شود و تصمیم‌گیری بدون آینده‌نگری کافی انجام می‌گیرد. به عبارت دیگر، با ضریب تخفیف کوچک، پاداش تنها در آینده نزدیک در ارزیابی تابع ارزش در نظر گرفته می‌شود که می‌تواند منجر به اقدامات کوته‌بینانه یا تهاجمی گردد. علاوه براین، مقدار کوچک ضریب تخفیف سبب کاهش سرعت همگرایی یادگیری می‌شود [۲۲]. این مشاهدات ما را برمی‌انگیزد تا یک قانون تطبیقی برای به‌روزرسانی ضریب تخفیف پیشنهاد نماییم. برای این منظور ابتدا تعدادی سیگنال راهنما از ایستگاه پایه به کاربران ارسال می‌گردد. کاربران در مرحله بعدی سیگنال دریافتی را به ایستگاه پایه فیدبک می‌نمایند. در ایستگاه پایه همبستگی بین سیگنال ارسالی به کاربران و سیگنال دریافتی از هر کاربر محاسبه می‌گردد و زیربانده‌ی که میزان همبستگی بالاتری دارد، احتمال تخصیص بالاتری خواهد داشت. به این ترتیب، ضریب تخفیف تطبیق‌یافته در روش پیشنهادی منجر به دسترسی به طیف فرکانسی امن‌تر برای خنثی نمودن اثر تداخلگر می‌شود. با توجه به اینکه زمانی که فرآیند تصمیم‌گیری ابعاد بالایی دارد و محیط دارای عدم قطعیت است و یافتن ضریب تخفیف ثابت بهینه، یک کار پرچالش است، وجود ضریب تخفیف تطبیق‌یافته می‌تواند به حل مساله کمک شایانی کند. ضریب تخفیف تطبیق‌یافته پیشنهادی، عامل را قادر می‌سازد تا برآورد درست‌تری از مبلغ پاداش آینده داشته باشد. برای ارزیابی عملکرد روش پیشنهادی، مجموع نرخ ارسال داده، توان تداخلگر و توان نویز و تعداد تکرار برای رسیدن به پاداش مطلوب مدنظر قرار گرفته است و عملکرد روش پیشنهادی بهتر از حالت با ضریب تخفیف ثابت است. نتایج حاصل نشان می‌دهد که روش پیشنهادی با ضریب تخفیف تطبیق‌یافته در مقایسه با سایر روش‌ها در محیط‌های پویا بهترعمل می‌کند.

[1] reward
[2] Multi agent reinforcement learning

۲- یادگیری تقویتی چندعاملی

یادگیری تقویتی چندعاملی نوع مهمی از یادگیری ماشین است که در آن چند عامل[1] در همکاری یا رقابت با هم می‌آموزند که چگونه در محیط با انجام اقدامات و بررسی نتایج آن‌ها رفتار کنند. در هوش مصنوعی[2]، عامل هوشمند، موجودیت خودکاری است که محیط اطراف خود را می‌بیند و با استفاده از محرک‌ها اقداماتی را در محیط انجام می‌دهد و فعالیت‌های خود را در مسیر کسب اهداف هدایت می‌کند. عامل‌های هوشمند ممکن است از یادگیری یا دانش برای کسب اهداف خود بهره بگیرند. این عامل‌ها ممکن است بسیار ساده یا خیلی پیچیده باشند. در روش یادگیری تقویتی، یک عامل در حال یادگیری محیط از طریق پاداش اقدام عامل‌ها است، به این صورت که عامل حالت[3] S_0 را از محیط دریافت می‌کند، بر اساس حالت S_0 ، عمل[4] A را انجام می‌دهد و محیط به حالت جدید S_1 انتقال پیدا می‌کند و محیط پاداش[5] R_1 را به عامل می‌دهد. این حلقه یادگیری تقویتی دارای یک توالی از حالت، عمل و پاداش است. هدف عامل آن است که پاداش انباره‌ای (تجمعی[6]) مورد انتظار را بیشینه کند. طرز کار یادگیری تقویتی در شکل ۱ نشان داده شده است.



شکل ۱- طرز کار یادگیری تقویتی

اما چالش اصلی در مبحث پاداش این است که پاداشی که در گام‌های زمانی ابتدایی برای اقدامات در نظر گرفته شده‌اند، احتمال وقوع بیشتری دارند، زیرا از پاداش‌های بلند مدت آینده قابل پیش‌بینی‌تر هستند. بنابرین یک ضریب تخفیف با عنوان β تعریف می‌شود. این مقدار باید بین ۰ و ۱ باشد. هر چه β بزرگ‌تر شود، تخفیف کمتر است. این یعنی عامل یادگیرنده به پاداش‌های بلند مدت اهمیت بیشتری می‌دهد. از سوی دیگر، هرچه β کوچک‌تر باشد، تخفیف بیشتر است. این یعنی عامل توجه بیشتری به پاداش‌های کوتاه مدت می‌کند. پاداش مورد انتظار انباره‌ای تخفیف داده شده به صورت جمع پاداش‌ها در گام‌های زمانی مختلف محاسبه می‌شود. هر پاداش با β به توان گام زمانی، تخفیف داده می‌شود و به پاداش تخفیف داده شده سایر زمان‌ها افزوده می‌گردد. در سناریو پیاده‌سازی یادگیری تقویتی، یک نقطه آغازین و یک نقطه پایانی وجود دارد. این منجر به ایجاد یک اپیزود یعنی لیست حالت‌ها، اعمال، پاداش‌ها و حالت‌های جدید می‌شود. در یادگیری به روش مونت کارلو، هنگامی که اپیزود به پایان می‌رسد ، عامل به پاداش تجمعی کل نگاه می‌کند تا از نحوه عملکرد خود آگاه شود. در روش مونت کارلو، پاداش‌ها تنها در پایان بازی دریافت می‌شوند. بنابراین، بازی جدید با یک

[1] Agent
[2] Artificial intelligence
[3] State
[4] Action
[5] Reward
[6] Cumulative reward

دانش افزوده آغاز می‌شود و عامل در هر تکرار تصمیمات بهتری اتخاذ می‌کند. در یادگیری تقویتی سیاست‌محور، قصد بهینه‌سازی تابع سیاست[1] P است. سیاست چیزی است که رفتار عامل را در یک زمان داده شده، تعیین می‌کند. عامل یک تابع سیاست را می‌آموزد. این امر به او کمک می‌کند تا هر حالت را به بهترین عمل ممکن نگاشت کند. یادگیری تقویتی عمیق از شبکه‌های عصبی عمیق برای حل مسائل یادگیری تقویتی استفاده می‌کند، از این رو در نام آن از کلمه عمیق استفاده شده است. با در نظر گرفتن Q-Learning که یادگیری تقویتی کلاسیک محسوب می‌شود و Deep Q-Learning(DQN)[2] می‌توان تفاوت آن‌ها با یکدیگر را دید. در رویکرد اول، از الگوریتم‌های سنتی برای ساخت جدول Q استفاده می‌شود تا به عامل در یافتن اقدامی که باید در هر حالت انجام شود کمک کند. در دومین رویکرد، از شبکه عصبی برای تخمین پاداش استفاده می‌شود. به عبارت دیگر، به جای استفاده از یک جدول اقدام–حالت که تعداد عناصر محدودی دارد، از یک شبکه عصبی آموزش دیده برای تخمین پاداش در هر حالت استفاده می گردد. نحوه عملکرد یادگیری تقویتی در مرحله آموزش در الگوریتم ۱ نشان داده شده است.

الگوریتم۱. یادگیری تقویتی در مرحله آموزش

- **تولید مقادیر اولیه:** β
- **برای هر اپیزود**

به ازای هر زمان t

a. تعیین وضعیت S_t

b. شیوه $\{\varepsilon - greedy$ **شیوه** $\}$

- با احتمال ε **یک اقدام تصادفی انتخاب گردد**

-در غیر این صورت (با احتمال $1-\varepsilon$ **) اقدام متناظر با بیشینه پاداش تجمعی انتخاب گردد:**

$$A_t = \arg\max_a Q_P(S_t, a) = \arg\max_a E_P\{\sum_{k=0}^{\infty} \beta^k R_{k+t+1} \mid S_t, a, P\}$$

c. اجرای اقدام A_t **و مشاهده وضعیت بعدی** S_{t+1} **و محاسبه پاداش** R_{t+1}

d. ذخیره $(S_t, A_t, R_{t+1}, S_{t+1})$ **در حافظه** D

e. $\{$**انتخاب تصادفی mini-batch**$\}$ **چند دسته** (S_j, A_j, R_j, S_{j+1}) **به عنوان نمونه از** D **انتخاب می‌شود**

f. **محاسبه**

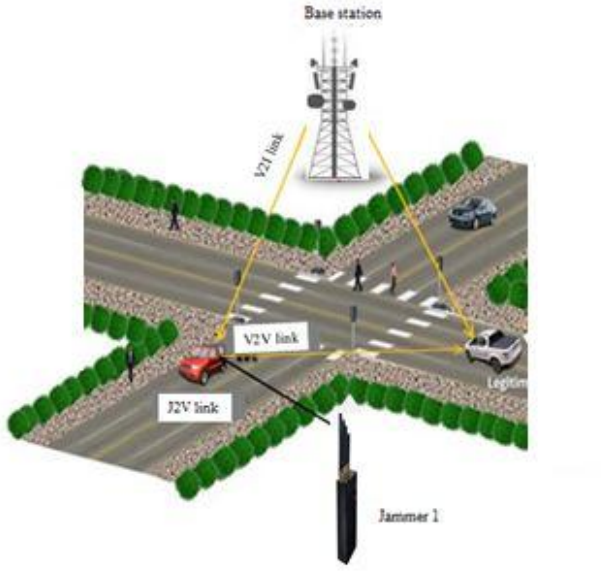
$$y_j = \begin{cases} r_j & S_j = \text{terminal} \\ r_j + \beta \max_a Q_P(S_j, a) & S_j = \text{non-terminal} \end{cases}$$

g. **آموزش شبکه DQN بر مبنای روش بهینه سازی** RMSProp **و با تابع هزینه** $(y_j - Q_p(S_j, A_j))^2$

[1] Policy
[2] Deep Q-learning

۳- مدل سیستم

شبکه کاربران و ایستگاه پایه در حضور تداخل گر در شکل ۲ نشان داده شده‌است.



شکل ۲- مدل سیستم

در این شکل، ما یک ارتباط بی‌سیم را در نظر می‌گیریم که شامل یک ایستگاه پایه، چند کاربر و یک اختلال‌گر است. هدف کاربران ارسال داده در محدوده فرکانسی کمتر درگیر شده با اختلال‌گر است.کاربرها در بازه‌های فرکانسی مشخص، داده خود را به ایستگاه پایه ارسال می‌کنند. تصمیم‌گیری در مورد تخصیص طیف مورد استفاده هر کاربر توسط ایستگاه پایه انجام می‌شود. ایستگاه پایه به عنوان عامل در یادگیری تقویتی می‌تواند اطلاعات مربوط به توان ارسالی و شرایط کانال کاربر را بررسی کرده و تصمیم خود را برای اقدام هرکاربر مبنی بر استفاده یا عدم استفاده از طیف مورد نظر به کاربران ارسال کند. الگوی رفتاری اختلال‌گر به وسیله اطلاعات به دست آمده از محیط بین جفت فرستنده و گیرنده به دست می‌آید. برای اعمال الگوریتم یادگیری تقویتی در این سناریو هر بازه فرکانسی یک حالت[1] است و هر اقدام[2] تخصیص یا عدم تخصیص فرکانس به یک کاربربه صورت باینری خواهد بود. لازم به ذکر است که هر بازه فرکانسی تنها می‌تواند به یک کاربر تخصیص پیدا کند. هدف این است که زیرباندهای فرکانسی را به نحوی بین کاربران توزیع کنیم که اثر تداخلگر را تا جای ممکن خنثی نماییم.

[1] state
[2] action

٤- روش پیشنهادی

در این بخش روشی برای جلوگیری از شنود اطلاعات توسط شنودگر پیشنهاد شده است که در آن نحوه تسهیم طیف به صورتی است که باندهای فرکانسی کمتر تخریب شده، برای ارسال داده کاربران شناسایی می‌شوند و به این ترتیب با مدیریت تسهیم فرکانس، با تخریب اطلاعات توسط تداخلگر مقابله می‌گردد. برای این منظور تکنیک انتخاب باند فرکانسی امن مبتنی بر یادگیری عمیق تقویتی پیشنهاد شده است. تابع پاداش در یادگیری تقویتی به صورت مجموع نرخ ارسال داده در یک باند فرکانسی مشخص تعریف می‌شود:

$$R(f_i)=\sum_{k=1}^{K}\rho_k C_{V2B}^{k}(f_i) \quad (۱)$$

در این رابطه، ρ_k نشانگر باینری برای دسترسی یا عدم دسترسی کاربر k ام به زیر باند f_i است. اگر کاربر k ام به زیر باند f_i دسترسی داشته باشد، این ضریب برابر یک خواهد بود و در غیر این صورت مقدار آن صفر لحاظ می‌شود. $C_{V2B}^{k}(f_i)$ ظرفیت k امین لینک $V2B$ در زیرباند f_i به صورت زیر محاسبه می‌شود:

$$C_{V2B}^{k}[f_i]=W\log(1+\gamma_{V2B}^{k}[f_i]) \quad (۲)$$

که در آن، W پهنای باند و $\gamma_{V2B}^{k}[f_i]$ نسبت سیگنال به تداخل و نویز[1] ($SINR$) برای k امین لینک $V2B$ به صورت زیر نوشته می‌شود:

$$\gamma_{V2B}^{k}[f_i]=\frac{p_{V2B}[f_i]g^{k}{}_{V2B}}{I_k+I_J+\sigma^2} \quad (۳)$$

در رابطه فوق، $p_{V2B}[f_i]$ توان یکسان مورد استفاده کاربران به عنوان فرستنده در لینک $V2B$، $g^{k}{}_{V2B}[f_i]$ ضریب کانال لینک $V2B$ مورد نظر در فرکانس f_i، σ^2 توان نویز در زیر باند f_i می‌باشند، I_k تداخل ناشی از ارتباطات مربوط به لینک‌های دیگر کاربران با ایستگاه پایه و I_J تداخل ناشی از ارتباطات مربوط به لینک‌تداخلگر با کاربر در همین فرکانس است که به صورت زیر تعریف می‌شوند:

[1] Signal to interference and noise ratio

$$I_k = \sum_{k' \neq k} \rho_{k'} p_{V2B}[f_i] g^k{}_{V2B}[f_i] \quad (۴)$$

$$I_J = \int_{f_i - \frac{b_u}{2}}^{f_i + \frac{b_u}{2}} g_{J2V}[f_i] J(f - f_i) df \quad (۵)$$

$J(f)$ چگالی طیفی توان تداخلگر و b_u پهنای باند سیگنال باند پایه تداخلگر را نشان می‌دهد. سه نوع تداخلگر نقطه‌ای[۱]، تداخلگر جاروب[۲] و تداخلگر رگباری[۳] در این پژوهش در نظر گرفته شده‌است[۲۳]. تداخلگر نقطه ای شکلی از تداخلگر است که در آن یک تداخلگر تمام توان خود را بر روی یک فرکانس متمرکز می‌کند. تداخلگر جاروب توانایی انتقال قدرت کامل از یک فرکانس به فرکانس دیگر را دارد. این حرکت فراگیر چندین فرکانس را به‌سرعت و نه در یک زمان مسدود می‌کند. تداخلگر رگباری، چند فرکانس را به طور همزمان توسط یک تداخلگر واحد مورد حمله قرار می‌دهد. داده ارسالی از تداخلگر نقطه‌ای، جاروب و رگباری به ترتیب در روابط (۶-۸) تعریف شده است.

$$J(f) = a(f_0)\delta(f - f_0) \quad (۶)$$

$$J(f) = \sum_{m=1}^{M} a(f_n)\delta(f - f_n)e^{-j2\pi f_n t_m} \quad (۷)$$

$$J(f) = \sum_{n=1}^{N}\sum_{m=1}^{M} a(f_n)\delta(f - f_n)e^{-j2\pi f_n t_m} \quad (۸)$$

که در آن $a(f_n)$ دامنه داده ارسالی توسط تداخلگر در فرکانس f_n و $\delta(f_n)$ تابع ضربه نشان‌دهنده فرکانس مورد حمله f_n است. $e^{-j2\pi f_n t_m}$ نشانگر تاخیر زمانی داده ارسالی توسط تداخلگر در فرکانس مورد نظر است. ضرایب کانال بین کاربرها و ایستگاه پایه بر مبنای محوشدگی مقیاس بزرگ (شامل تلفات مسیر[۴] و محوشدگی سایه[۵]) و محوشدگی مقیاس کوچک برحسب زمان باید به‌روز رسانی شوند. محوشدگی مقیاس کوچک به صورت یک پارامتر وابسته به زمان با توزیع نمایی در نظر گرفته شده است. ضرایب کانال لینک $V2B$ به طریق زیر محاسبه می‌شود:

$$g_{V2B}[f_i] = \alpha_{V2B} h_{V2B}[f_i] \quad (۹)$$

پارامتر $h_{V2B}[f_i]$، محوشدگی مقیاس بزرگ کانال در فرکانس f_i و کانال لینک $V2B$ است و پارامتر α_{V2B}، محوشدگی مقیاس کوچک کانال در لینک مورد نظر و مستقل از فرکانس است.

در سیستم مورد نظر، برخی از زیر باندهای طیف فرکانسی موجود هستند که کاربرها خواستار ارسال داده در آن ها هستند. این باندها کم و بیش مورد حمله اختلال‌گر هستند. در روش پیشنهادی ابتدا یک داده راهنما از ایستگاه پایه به کاربرها ارسال می‌شود. داده راهنمای ارسالی در فرکانسهای مختلف و در گام زمانی t برابر است با:

$$X_t = (X_t^{f_1}, X_t^{f_2}, ..., X_t^{f_T}) \quad (۱۰)$$

[1] Spot jammer
[2] Sweep jammer
[3] Barrage jammer
[4] Path loss fading
[5] Shadow fading

داده ارسالی توسط تداخلگر به صورت زیر تعریف می‌شود:

$$J_t = (J_t^{f_1}, J_t^{f_2}, ..., J_t^{f_T}) \quad (۱۱)$$

در این صورت داده دریافتی توسط کاربر برابر خواهد بود با:

$$Z_t = h_{B-U} X_t + h_{j-U} J_t + n_t \quad (۱۲)$$

که در آن n_t نویز سفید گوسی با میانگین صفر است. کاربر مورد نظر پس از دریافت داده راهنما همان داده را به ایستگاه پایه برمی‌گرداند. داده‌های دریافتی توسط ایستگاه پایه $Y_t = (Y_t^{f_1}, Y_t^{f_2}, ..., Y_t^{f_T})$ است که درآن علاوه بر داده ارسالی از BS (X_t)، عامل دیگر مربوط به تداخلگر نیز حضور دارد:

$$Y_t = h_{B-U}^T Z_t = h_{B-U}^T (h_{B-U} X_t + h_{j-U} J_t + n_t) = X_t + h_{B-U}^T h_{j-U} J_t + h_{B-U}^T n_t \quad (۱۳)$$

در ایستگاه پایه، همبستگی میان X_t و Y_t برای باندهای مختلف فرکانسی برای هر کاربر محاسبه می‌شود:

$$\beta^k = [\rho_{X_t^{f_1}, Y_t^{f_1}}, \rho_{X_t^{f_2}, Y_t^{f_2}}, ..., \rho_{X_t^{f_T}, Y_t^{f_T}}] \quad (۱۴)$$

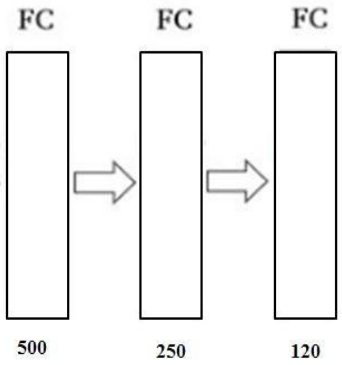
در این رابطه $\rho_{X_t^{f_i}, Y_t^{f_i}}$ نشان‌دهنده همبستگی میان $X_t^{f_i}$ و $Y_t^{f_i}$ در فرکانس f_i است. واضح است که هرچه میزان همبستگی نزدیک به عدد ۱ باشد نشان می‌دهد که داده راهنما کمتر تحت تاثیر عوامل اختلال‌گر قرار گرفته است و فرکانس مورد نظر صلاحیت بیشتری برای انتخاب توسط کاربر مورد نظر دارد و باید تاثیر بیشتری در پاداش درنظر گرفته شده در الگوریتم یادگیری تقویتی عمیق داشته باشد. ضرایب همبستگی محاسبه شده به عنوان ضریب تخفیف در الگوریتم یادگیری تقویتی مورد بهره‌برداری قرار می‌گیرند. بنابراین، مقدار پاداش تجمعی به ازای شروع اقدام از حالت S به ازای اقدام مشترک اولیه A تحت خط مشی P برابر خواهد بود با:

$$Q(t) = E\{\sum_{k=0}^{\infty} \beta^k(i) \rho_k R(f_i) \mid S_0 = S, A_0 = A, p\} \quad (۱۵)$$

به این ترتیب، پاداش‌های آینده با ضریب تخفیف متغیر محاسبه شده به هم مربوط می‌شوند. روند انتخاب اقدام‌ها با طرح $\varepsilon - greedy$ انجام شده است [۲۴]. طرح $\varepsilon - greedy$ برای بهینه‌سازی تابع هزینه در یک اکتشاف اتخاذ می‌شود. به عبارت دیگر، با استفاده از $\varepsilon - greedy$، اقدام با احتمال $1-\varepsilon$ با حداکثر پاداش انتخاب شده است، در حالی که یک اقدام تصادفی با احتمال ε انتخاب می‌شود.در پایان مرحله آموزش، جدولی از مقدار پاداش‌های تجمعی به ازای سیاست‌های اتخاذ شده (حالت‌های و اقدام‌های موجود در مسیر) تشکیل می‌شود تا در مرحله تست، به ازای سیاست‌های در نظر گرفته شده در محیط جدید، بهترین سیاست برای رسیدن به بیشترین پاداش تجمعی انتخاب شود.

$$P = \arg\max_p Q \quad (۱۶)$$

تعداد اقدام‌ها و حالت‌ها در روش یافتن بهترین سیاست طبق رابطه (۱۶) می‌تواند هزاران نفر باشد، که مدیریت مقادیر Q را در جدول بسیار ناکارآمد می‌کند. اینجاست که می‌توانیم به جای استفاده از جدول، از شبکه‌های عصبی (DQN[1]) برای پیش‌بینی مقادیر Q برای اقدامات در یک حالت معین استفاده کنیم. به عبارت دیگر، به جای نگاشت یک جفت حالت-اقدام به یک مقدار Q، یک شبکه عصبی حالت های ورودی را به جفت عمل و Q نگاشت می کند. شبکه استفاده شده در این مقاله در شکل ۳ نشان داده شده است.



شکل ۳- شبکه عمیق استفاده شده به عنوان DON

همان‌طور که در شکل ۳ نشان داده شده است، شبکه DQN استفاده شده در این مقاله از سه لایه تمام متصل‌کننده[2] با تعداد ۵۰۰، ۲۵۰ و ۱۲۰ نورون تشکیل شده است که از واحد خطی اصلاح شده[3] (Relu) به عنوان تابع فعال سازی[4] در رابطه (۱۷) و ریشه میانگین مربعات [5] (RMSProp) استفاده می کند.

$$f(x) = \max(0, x) \tag{۱۷}$$

در مرحله نهایی آموزش، سیاست‌های در نظر گرفته و Q متناظر آن‌ها به عنوان ورودی و خروجی به یک شبکه متشکل از سه لایه تمام‌جمع‌کننده با ۵۰۰، ۲۵۰ و ۱۲۰ نورون اعمال می‌شوند. شبکه آموزش‌یافته در مرحله تست، بهترین سیاست رابرای داشتن بیشترین پاداش تعیین می‌کند. مراحل روش پیشنهادی مقابله با تداخلگر برای تسهیم طیف در الگوریتم ۲ نشان داده شده است. در این الگوریتم ابتدا محیط آموزش با تولید موقعیت، سرعت و ضرایب کانال تشکیل می‌شود. ضریب تخفیف با استفاده از دنباله راهنما و بازخورد آن برای هر کاربر محاسبه می‌شود. با اعمال الگوریتم یادگیری تقویتی و محاسبه پاداش تجمعی به ازای تخصیص یا عدم تخصیص بازه‌های فرکانسی به کاربرها، شبکه DQN با تابع بهینه‌سازی انتشار ریشه میانگین مربعات آموزش داده می‌شود.

شبکه DQN آموزش یافته قادر خواهد بود بهترین سیاست از نظر تخصیص یا عدم تخصیص بازه‌های فرکانسی به کاربران را به ازای بیشینه کردن پاداش تجمعی تعیین کند.

[1] Deep Q-network
[2] Fully-connected
[3] Rectified linear unit
[4] Activation function
[5] Root Mean Squared Propagation (RMSProp)

الگوریتم۲- روش پیشنهادی مقابله با تداخلگر برای تسهیم طیف

- **تولید مقادیر اولیه:** موقعیت ، سرعت کاربرها، ضرایب کانال
- **ارسال دنباله راهنما:** X_t
- **دریافت داده بازخوردشده از کاربران در ایستگاه پایه:** Y_t
- **تعیین مقدار ضریب تخفیف با محاسبه همبستگی بین داده راهنما و داده بازخورد داده شده:**

$$\beta^k = [\rho_{X_t^{f1},Y_t^{f1}}, \rho_{X_t^{f2},Y_t^{f2}}, ..., \rho_{X_t^{fT},Y_t^{fT}}]$$

مرحله آموزش:

به ازای هر اپیزود

- **به روزرسانی مکان وسایل نقلیه**

  به ازای هر زمان t

  برای هر کاربر k
- **تعیین وضعیت** $S_t^{(k)}$ **:** انتخاب یک بازه فرکانسی برای کاربر
- **انتخاب اقدام** $A_t^{(k)}$ **:** (تخصیص بازه فرکانسی به کاربر) **از بین اقدام‌ها با** $\varepsilon - greedy$

  - با احتمال ε **یک اقدام تصادفی انتخاب گردد**

  -در غیر این صورت (با احتمال $1-\varepsilon$ **) اقدام متناظر با بیشینه پاداش تجمعی انتخاب گردد:**

$$A_t^{(k)} = \arg\max_a Q_p^{(k)}(S_t^{(k)}, a) = \arg\max_a E_P\{\sum_{m=0}^{\infty} \beta^m R^{(k)}{}_{m+t+1} \mid S_t^{(k)}, a, P\}$$

- **اجرای اقدام تعیین‌شده** $A_t^{(k)}$ **توسط هر وسیله نقلیه** (تخصیص بازه فرکانسی به کاربر) **و مشاهده وضعیت بعدی** $S_{t+1}^{(k)}$ **و محاسبه پاداش** $R_{t+1}^{(k)}$ (ظرفیت کانال)
- **ذخیره** $(S_t^{(k)}, A_t^{(k)}, R_t^{(k)}, S_{t+1}^{(k)})$ **در حافظه** D_k
- {**انتخاب تصادفی mini-batch**} **چند دسته** $(S_j^{(k)}, A_j^{(k)}, R_j^{(k)}, S_{j+1}^{(k)})$ **به عنوان نمونه از** D **انتخاب می‌شود**
- **محاسبه**

$$y_j^{(k)} = \begin{cases} r_j^{(k)} & S_j^{(k)} = \text{terminal} \\ r_j^{(k)} + \beta \max_a Q_P(S_j^{(k)}, a) & S_j^{(k)} = \text{non-terminal} \end{cases}$$

- **آموزش شبکه DQN بر مبنای روش بهینه سازی** RMSProp **و با تابع هزینه** $(y_j^{(k)} - Q_p(S_j^{(k)}, A_j^{(k)}))^2$

مرحله تست:

- **تولید محیط تست:** موقعیت ، سرعت کاربرها، ضرایب کانال
- **به دست آوردن سیاست بهینه** P (تخصیص یا عدم تخصیص بازه‌های فرکانسی به کاربران) **و** Q **بیشینه با استفاده از شبکه DQN**

در الگوریتم $\varepsilon - greedy$ یک مقدار برای ε تعیین می‌شود که در این مقاله ۰/۰۲ در نظر گرفته شده است.

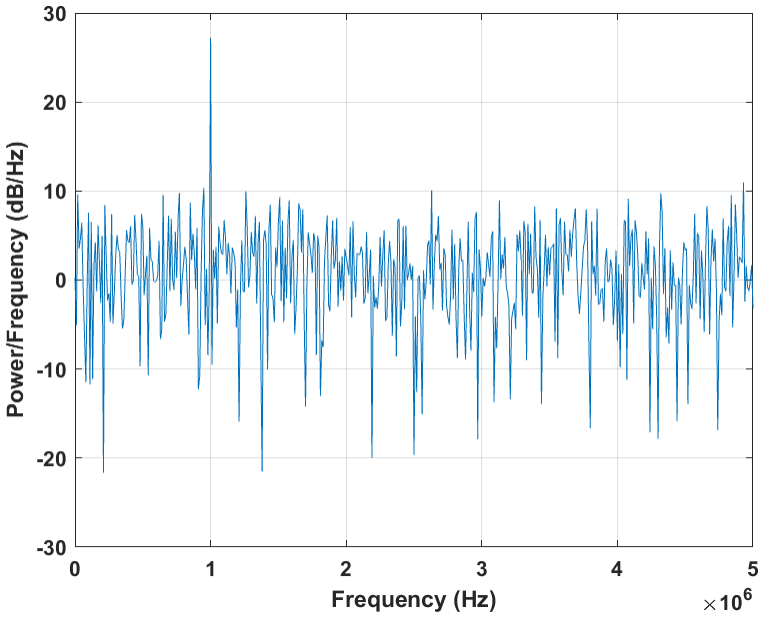
۵- نتایج شبیه‌سازی

در این بخش، نتایج شبیه سازی روش پیشنهادی ارائه شده است. شبکه ارتباطی وسایل نقلیه مطابق [۲۵]، مدل‌سازی شده‌است. پارامترهای به کار رفته در شبیه سازی در جدول ۱ آورده شده است.

جدول ۱- پارامترهای استفاده شده در شبیه‌سازی

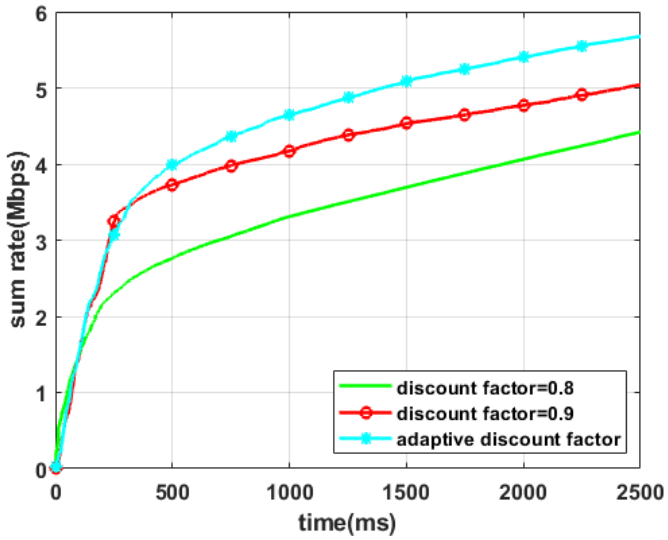
پارامتر	مقدار
تعداد کاربران	۴
پهنای باند	۴ مگاهرتز
توان نویز کانال	۸dBm
محدوده زمانی انتقال	۱۰۰ms
توزیع محوشدگی سایه	توزیع لوگ نرمال
فاصله همبستگی	۵۰m
توزیع محوشدگی سریع	توزیع رایلی
توان تداخلگر	۲۲ dBm
توان کاربر	۲۳dBm
توان نویز	۸dBm
زمان به روز رسانی محوشدگی سریع	۱ ms
زمان به روز رسانی تلفات مسیر و محوشدگی سایه	۱۰۰ ms

در این پژوهش، تداخلگر نقطه‌ای با تولید تابع ضربه در یک فرکانس تصادفی و نیز تداخلگر جاروب به صورت اندازه ضرایب تبدیل فوریه مجموع تابع $\cos(2\pi ft)$ و شیفت‌یافته‌های آن پیاده‌سازی شده است. تداخلگر رگباری حالت کلی‌تر تداخلگرهای نقطه‌ای و جاروب است. چگالی طیفی توان در نظر گرفته شده برای تداخلگر رگباری مورد استفاده در این پژوهش، با استفاده از مدل موجود در Array System Toolbox Phased در برنامه Matlab پیاده‌سازی شده که در شکل ۴ نشان داده شده است.



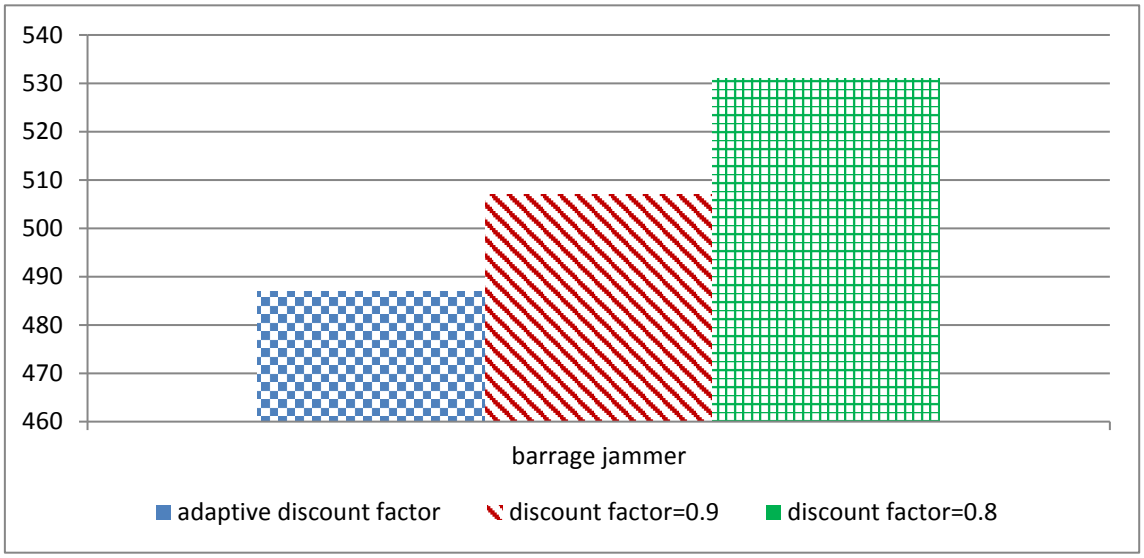
شکل٤- چگالی طیفی توان در نظر گرفته شده برای تداخلگر رگباری

میانگین چگالی طیفی توان نشان داده شده در این نمودار dBm ۲۲ است.گفتنی است برای مقابله با تداخلگر در تسهیم طیف روشی در مقالات ارائه نشده است. از این رو عملکرد خود روش پیشنهادی را در سناریوهای مختلف مورد ارزیابی قرار می دهیم و نیز برای اینکه تکنیک پیشنهادی برای تنظیم ضریب تخفیف یافته را مورد ارزیابی قرار دهیم، از ضریب تخفیف های ثابت در روش پیشنهادی استفاده کرده ایم. نرخ مجموع بدست آمده در مرحله آموزش توسط روش پیشنهادی در حضور تداخلگر رگباری در شکل ۵ نشان داده شده است. توان نویز dBm ۲۲ و توان تداخلگر dBm ۹ است. در این شکل، روش پیشنهادی با ضریب تخفیف تطبیق‌یافته با روش پیشنهادی به ازای ضرایب تخفیف ۰/۸و ۰/۹ مقایسه شده اند.



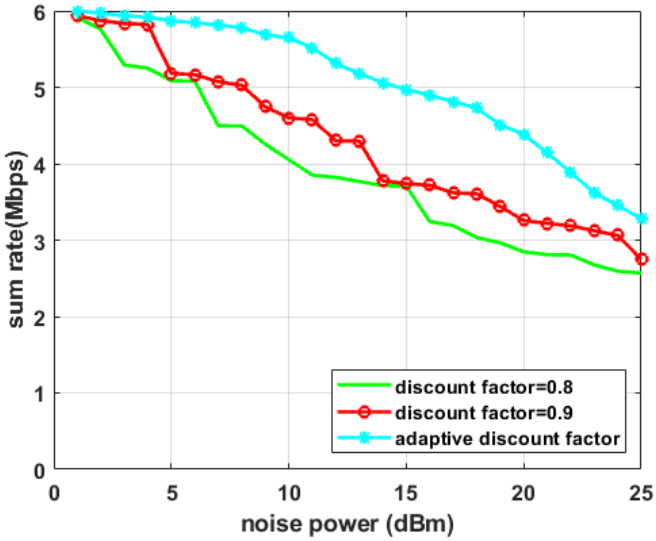
شکل۵- نرخ مجموع بر حسب زمان به ازای ضریب تخفیف‌های متفاوت در حضور تداخلگر رگباری

مطابق شکل پاداش در شرایط ضریب تخفیف تطبیق‌یافته نسبت به ضریب تخفیف‌های ثابت همگرایی سریع‌تری دارد. دلیل این امر انتخاب طیفهای کمتر تخریب شده توسط روش پیشنهادی است. در این شرایط کاربرانی که در طیف مورد نظر در معرض تداخلگر هستند، همبستگی کمتری بین سیگنال ارسالی و سیگنال بازخورد داده شده تجربه می‌کنند و این مقدار همبستگی به عنوان ضریب تخفیف تاثیر مشارکت این کاربران برای انتخاب به عنوان اقدام آینده در پاداش را کاهش می‌دهد. بنابراین همگرایی سریع‌تری با استفاده از ضریب تخفیف تطبیق‌یافته رخ می‌دهد. تعداد تکرار برای رسیدن به همگرایی در صورت وجود تداخلگر رگباری برای ضریب تخفیف ثابت ۰/۸، ۰/۹ و ضریب تخفیف تطبیق‌یافته در شکل ۶ مقایسه شده است.



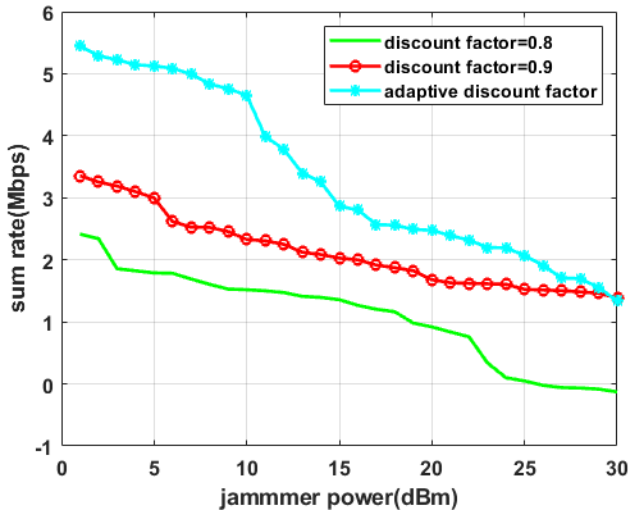
شکل ٦- تعداد تکرار برای همگرایی پاداش تجمیعی در سه نوع تداخلگر

همانطور که در شکل ۶ مشخص شده است، روش تسهیم طیف امن با ضریب تخفیف تطبیق‌یافته تعداد تکرار کمتری را برای رسیدن به همگرایی به نرخ مطمین مطلوب تجربه می‌کند. از داده‌های موجود در جدول ۲ می توان نتیجه گرفت که روش پیشنهادی با ضریب تخفیف تطبیق‌یافته نسبت به روش پیشنهادی با ضریب تخفیف ثابت، سرعت آموزش الگوریتم ی دارد. روند نرخ مجموع بر حسب توان نویز در حضور تداخلگر رگباری در شکل ۷ نشان داده شده است. توان تداخلگر برای رسم این نمودار ۹ dBm است.



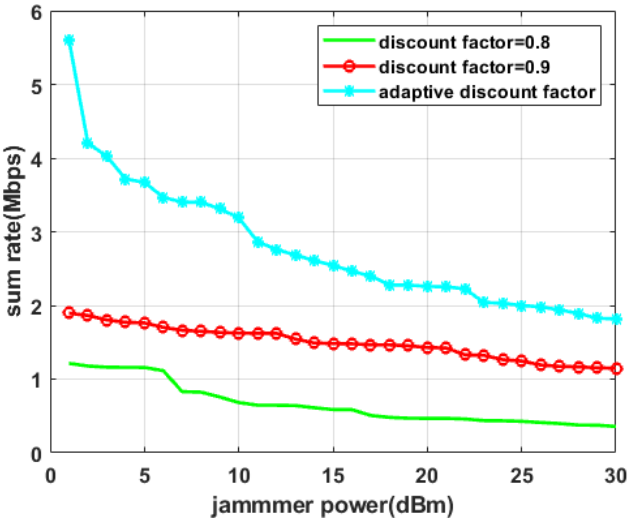
شکل۷- نرخ مجموع برحسب توان نویز در تداخلگر رگباری

همان‌طور که در شکل مشخص است، با افزایش توان نویز، نرخ مجموع به دلیل رابطه عکس با توان نویز کاهش پیدا می‌کند، اما در این شرایط، تسهیم طیف با ضریب تخفیف تطبیق‌یافته عملکرد بهتری نسبت به ضریب تخفیف ثابت دارد و نرخ مجموع ارائه شده توسط ضرایب تطبیق یافته بهتر از ضرایب ثابت است. با افزایش توان نویز سفید، مقدار همبستگی برای فرکانس‌هایی که نویز بیشتری دارند افزایش پیدا می‌کند. در این بخش می خواهیم نمودارهای نرخ مجموع برحسب توان تداخلگر را برای انواع مختلف تداخلگر ترسیم کنیم. برای این منظور، ۱۰۰ الگوی تصادفی مختلف برای چگالی طیفی توان تولید شده است، سپس با تغییر توان تداخلگر، نرخ مجموع برای هر کدام از ۱۰۰ الگو محاسبه شده و متوسط نتایج در منحنی‌ها گزارش شده است. توان نویز در حضور هر سه تداخلگر ۲۲ dBm در نظر گرفته شده است. متوسط نرخ مجموع بر حسب توان‌های مختلف در حضور تداخلگر نقطه‌ای در شکل ۸ نمایش داده شده است.



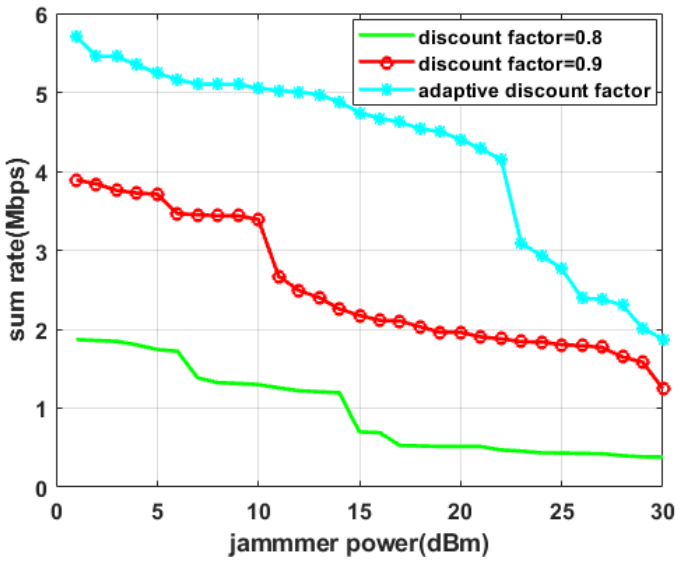
شکل۸- نرخ مجموع به ازای توان تداخلگر نقطه‌ای

همان‌طور که در شکل ۸ نشان داده شده است، با افزایش توان تداخلگر نقطه‌ای، مقدار نرخ مجموع روند کاهشی دارد چرا که با افزایش توان تداخلگر نرخ مجموع کاهش می یابد. نرخ مجموع بر حسب توان‌های مختلف در حضور تداخلگر جاروب در شکل ۹ نمایش داده شده است.



شکل۹- نرخ مجموع به ازای توان تداخلگر جاروب

با توجه به نمودار شکل ۹، هر سه نمودار روند نزولی دارند و روش تطبیق یافته نسبت به دو روش دیگر نرخ مجموع بالاتری در تمام توان های تداخلگر تولید می کند. نرخ مجموع بر حسب توان‌های مختلف در حضور تداخلگر رگباری در شکل ۱۰ نمایش داده شده است.



شکل۱۰- نرخ مجموع به ازای توان تداخلگر رگباری

همان‌طور که در شکل ۱۰ نشان داده شده است، نرخ مجموع با افزایش توان تداخلگر در حضور تداخلگر رگباری، روند کاهشی دارد که این امر به دلیل تاثیر معکوس توان تداخلگر بر نرخ ارسال داده است. همچنین ملاحظه می‌کنیم که روش پیشنهادی با ضریب تطبیق یافته به نرخ مجموع بالاتری نسبت به ضریب تخفیف ثابت دست پیدا می‌کند. نکته قابل توجه در این نتایج آن است که نرخ مجموع ارائه شده توسط روش پیشنهادی با افزایش توان تداخلگر نیز قابل قبول است و به ویژه تکنیک ضریب تطبیق یافته سبب مقاومت مناسب روش پیشنهادی در برابر تداخلگر در توان‌های مختلف می‌شود.

٦- نتیجه‌گیری

این مقاله یک روش ضد تداخلگر پنهان بر اساس تسهیم فرکانس معرفی می‌کند که مکانیزم آن کاهش احتمال تخصیص زیرباند فرکانسی مورد حمله تداخلگر به کاربرها است. یک چارچوب یادگیری تقویتی چندعاملی طراحی شده است که با تجزیه و تحلیل همبستگی داده انتقالی بین ایستگاه پایه و کاربر، زیرباندهای کمتر تخریب شده توسط تداخلگر را به طور غیر مستقیم به دست می‌آورد. روش پیشنهادی بر اساس تکنیک های یادگیری عمیق تقویتی چند عاملی و وابسته ساختن مقدار ضریب تخفیف به میزان تخریب زیر باند فرکانسی، امکان دسترسی سریع‌تر کاربران به طیف‌های مناسب و در نتیجه مقابله با تداخلگر را فراهم می‌کند.

مراجع:

[1] Y. Li *et al.*, "On the performance of deep reinforcement learning-based anti-jamming method confronting intelligent jammer," *Appl. Sci.*, vol. 9, no. 7, pp. 1–15, 2019, doi: 10.3390/app9071361.
[2] Naparstek, Oshri, and Kobi Cohen. "Deep multi-user reinforcement learning for distributed dynamic spectrum access." IEEE transactions on wireless communications 18, no. 1 (2018): 310-323.
[3] L. Jia, F. Yao, Y. Sun, Y. Niu, and Y. Zhu, "Bayesian Stackelberg Game for Antijamming Transmission With Incomplete Information," *IEEE Commun. Lett.*, vol. 20, no. 10, pp. 1991–1994, 2016, doi: 10.1109/LCOMM.2016.2598808.
[4] M. K. Hanawal, M. J. Abdel-Rahman, and M. Krunz, "Joint Adaptation of Frequency Hopping and Transmission Rate for Anti-Jamming Wireless Systems," *IEEE Trans. Mob. Comput.*, vol. 15, no. 9, pp. 2247–2259, 2016, doi: 10.1109/TMC.2015.2492556.
[5] L. Xiao, T. Chen, J. Liu, and H. Dai, "Anti-jamming transmission stackelberg game with observation errors," *IEEE Commun. Lett.*, vol. 19, no. 6, pp. 949–952, 2015, doi: 10.1109/LCOMM.2015.2418776.
[6] H. Zhu, C. Fang, Y. Liu, C. Chen, M. Li, and X. S. Shen, "You Can Jam but You Cannot Hide: Defending Against Jamming Attacks for Geo-Location Database Driven Spectrum Sharing," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 10, pp. 2723–2737, 2016, doi: 10.1109/JSAC.2016.2605799.
[7] Liu, Xin, Yuhua Xu, Luliang Jia, Qihui Wu, and Alagan Anpalagan. "Anti-jamming communications using spectrum waterfall: A deep reinforcement learning approach." IEEE Communications Letters 22, no. 5 (2018): 998-1001.
[8] B. Wang, Y. Wu, K. J. R. Liu, and T. C. Clancy, "An anti-jamming stochastic game for cognitive radio networks," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 4, pp. 877–889, 2011, doi: 10.1109/JSAC.2011.110418.
[9] Y. Wu, B. Wang, K. J. R. Liu, and T. C. Clancy, "Anti-jamming games in multi-channel cognitive radio networks," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 1, pp. 4–15, 2012, doi: 10.1109/JSAC.2012.120102.

[10] S. Machuzak and S. K. Jayaweera, "Reinforcement learning based anti-jamming with wideband autonomous cognitive radios," *2016 IEEE/CIC Int. Conf. Commun. China, ICCC 2016*, pp. 1–5, 2016, doi: 10.1109/ICCChina.2016.7636793.

[11] A. P. Badia *et al.*, "Never Give Up: Learning Directed Exploration Strategies," pp. 1–28, 2020.

[12] Y. Xiao, J. Hoffman, T. Xia, and C. Amato, "Learning Multi-Robot Decentralized Macro-Action-Based Policies via a Centralized Q-Net," *Proc. - IEEE Int. Conf. Robot. Autom.*, pp. 10695–10701, 2020, doi: 10.1109/ICRA40945.2020.9196684.

[13] K. K. Nguyen, T. Q. Duong, N. A. Vien, N.-A. Le-Khac, and M.-N. Nguyen, "Non-Cooperative Energy Efficient Power Allocation Game in D2D Communication: A Multi-Agent Deep Reinforcement Learning Approach," *IEEE Access*, vol. 7, pp. 100480–100490, 2019, doi: 10.1109/access.2019.2930115.

[14] H. Li, "Multiagent Q -learning for aloha-like spectrum access in cognitive radio systems," *Eurasip J. Wirel. Commun. Netw.*, vol. 2010, 2010, doi: 10.1155/2010/876216.

[15] H. H. Chang, H. Song, Y. . Zhang, H. He, and L. Liu, "Distributive dynamic spectrum access through deep reinforcement learning: A reservoir computing-based approach," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 1938–1948, 2019, doi: 10.1109/JIOT.2018.2872441.

[16] X. Chen, Z. Zhao, and H. Zhang, "Stochastic power adaptation with multiagent reinforcement learning for cognitive wireless mesh networks," *IEEE Trans. Mob. Comput.*, vol. 12, no. 11, pp. 2155–2166, 2013, doi: 10.1109/TMC.2012.178.

[17] S. B. Janiar and V. Pourahmadi, "Deep-reinforcement learning for fair distributed dynamic spectrum access in wireless networks," *2021 IEEE 18th Annu. Consum. Commun. Netw. Conf. CCNC 2021*, 2021, doi: 10.1109/CCNC49032.2021.9369536.

[18] V. François-Lavet, P. Henderson, R. Islam, M. G. Bellemare, and J. Pineau, *An introduction to deep reinforcement learning*, vol. 11, no. 3–4. 2018.

[19] X. Zhu *et al.*, "Dynamic Spectrum Anti-Jamming with Reinforcement Learning Based on Value Function Approximation," *IEEE Wirel. Commun. Lett.*, pp. 1–5, 2022, doi: 10.1109/LWC.2022.3228045.

[20] X. Liu, Y. Xu, L. Jia, Q. Wu, and A. Anpalagan, "Anti-jamming communications using spectrum waterfall: A deep reinforcement learning approach," *IEEE Commun. Lett.*, vol. 22, no. 5, pp. 998–1001, 2018, doi: 10.1109/LCOMM.2018.2815018.

[21] X. Chen, C. Wang, Z. Zhou, and K. Ross, "Randomized Ensembled Double Q-Learning: Learning Fast Without a Model," pp. 1–25, 2021.

[22] H. van Seijen, M. Fatemi, and A. Tavakoli, "Using a logarithmic mapping to enable lower discount factors in reinforcement learning," *Adv. Neural Inf. Process. Syst.*, vol. 32, no. NeurIPS, pp. 1–11, 2019.

[23] X. Song, P. Willett, S. Zhou, and P. B. Luh, "The MIMO radar and Jammer games," *IEEE Trans. Signal Process.*, vol. 60, no. 2, pp. 687–699, 2012, doi: 10.1109/TSP.2011.2169251.

[24] "Black, Paul E. 'greedy algorithm, Dictionary of Algorithms and Data Structures.' US Nat. Inst. Std. & Tech Report 88 (2012): 95.," vol. 88, p. 2012, 2012.

[25] conformance specification Radio, User Equipment UE. "3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); User Equipment (UE) conformance specification Radio transmission and reception." (2011).