بهبود عملکرد حمله در تیم ربات‌های فوتبالیست با استفاده از یادگیری تقویتی

مینا خاکسار[۱]، دانشجو؛ ولی درهمی[۲]، دانشیار؛ مهدی رضائیان[۳]، استادیار

۱- گروه مهندسی کامپیوتر - پردیس فنی و مهندسی - دانشگاه یزد - یزد - ایران - khaksarmina@stu.yazd.ac.ir
۲- گروه مهندسی کامپیوتر - پردیس فنی و مهندسی – دانشگاه یزد - یزد – ایران- vderhami@yazd.ac.ir
۳- گروه مهندسی کامپیوتر - پردیس فنی و مهندسی – دانشگاه یزد - یزد – ایران- mrezaeian@yazd.ac.ir

چکیده: به‌دلیل عدم امکان پیش‌بینی همه وضعیت‌های ممکن برای عامل‌ها در یک سیستم چندعامله‌ی پویا و گسترده، روش‌های یادگیری ماشین، ابزار مناسبی برای کنترل رفتار عامل‌ها می‌باشد. فوتبال شبیه‌سازی شده ربات‌ها یک مسئله شناخته‌شده برای ارزیابی الگوریتم‌های یادگیری ماشین روی سیستم‌های چندعامله است. در این مقاله الگوریتم یادگیری کیو ـ وی (یکی از الگوریتم‌های معروف یادگیری تقویتی) جهت بهبود عملکرد حمله در تیم ربات‌های فوتبالیست دو بعدی بکار گرفته شده‌است. سیگنال تقویتی براساس اینکه بازیکنانی که در حمله دخالت دارند، توپ را جلوی دروازه می‌رسانند، یا اینکه توپ را از دست می‌دهند، تعریف شده‌است و به‌ترتیب عامل‌ها با توجه به وضعیت ذکرشده، جایزه و جریمه دریافت می‌کنند. جهت بهبود عملکرد از ایده تقسیم سیگنال تقویتی متناسب با مقدار خبرگی عامل‌ها در یک سیستم چندعامله استفاده شده‌است. در اینجا میزان خبرگی متناسب با تفاوت ارزش عملِ با بالاترین ارزش با ارزشِ عملِ با کمترین مقدار ارزش در هر حالت تعریف می‌شود. نتایج شبیه‌سازی نشان می‌دهد، بهره‌گیری از ایده تقسیم سیگنال تقویتی براساس خبرگی در آموزش باعث افزایش سرعت در آموزش و بهبود عملکرد عامل‌ها شده‌است.

واژه‌های کلیدی: فوتبال شبیه‌سازی شده ربات‌ها، یادگیری تقویتی، یادگیری کیو ـ وی، سیستم‌های چندعامله، حمله.

Improve performance of attack in the team robots soccer using reinforcement learning

Mina Khaksar[1], Student; Vali Derhami[2], Associate Professor; Mehdi Rezaeian[3], Assistant Professor

1- Computer Engineering Department, Faculty of Engineering, Yazd University, Yazd, Iran, Email: khaksarmina@stu.yazd.ac.ir
2- Computer Engineering Department, Faculty of Engineering, Yazd University, Yazd, Iran, Email: vderhami@yazd.ac.ir
3- Computer Engineering Department, Faculty of Engineering, Yazd University, Yazd, Iran, Email: mrezaeian@yazd.ac.ir

Abstract: Due to the impossibility of predicting all possible states for agents in a wide dynamic multi-agent system, machine learning methods are useful tools to control agent behavior. Simulated Robot Soccer is a well known multi agent benchmark to evaluate machine learning algorithms. In this paper, QV-Learning algorithm (a well known reinforcement learning algorithm) is used to improve the performance of the attack in 2D robots soccer team. The reinforcement signal is defined based on the players involved in the attack can reach the ball in front of goal or lose the ball; They receive positive and negative reward according to the mentioned status, respectively. We use the idea of division the reinforcement signal proportional to the amount of expertness (knowledge) of agents to improve the performance. Here, the expertise is defined as the difference between highest action value and lowest action value in the each state. The simulation results show using the idea of expertise improves the train speed and the performance.

Keywords: Simulated Robot Soccer, reinforcement Learning, QV-Learning, multi-Agent System, attack.

تاریخ ارسال مقاله: ۱۳۹۵/۰۹/۲۹
تاریخ اصلاح مقاله: ۱۳۹۶/۰۱/۰۸ و ۱۳۹۶/۰۶/۰۶
تاریخ پذیرش مقاله: ۱۳۹۶/۰۸/۰۱
نام نویسنده مسئول: ولی درهمی
نشانی نویسنده مسئول: ایران - یزد – صفائیه - بلوار دانشگاه - دانشگاه یزد – ساختمان پردیس فنی و مهندسی۱.

۱- مقدمه

از دیدگاه هوش مصنوعی توزیع‌شده، مسابقه فوتبال ربات‌ها نمونه‌ای از مسائل چندعامله است که مباحث تحقیقاتی متعدد را می‌توان در آن یافت [۱]. هر تیم مجموعه‌ای از ربات‌ها است که باید به‌منظور رسیدن به هدف نهایی و مشترک که همان پیروزی در بازی است با یکدیگر همکاری داشته باشند. به‌ثمررساندن گل و جلوگیری از گل‌زدن توسط حریف را می‌توان به‌عنوان زیرمجموعه‌ای از هدف اصلی پیروزی در نظر گرفت. عامل‌ها در شرایطی که درک درستی از محیط ندارند و دید محلی نسبت به اطراف دارند، باید برای بهینه‌کردن میزان کارایی خود در محیط تلاش کنند. به‌علاوه عامل‌ها باید در یک محیط پویا به‌صورت بی‌درنگ و خصمانه عمل کنند. در نتیجه موفقیت تیمی وابسته به سرعت عملکرد و انعطاف‌پذیری عامل‌ها است. ازآنجایی‌که فضای حالات ممکن برای بازی بسیار بزرگ است، استفاده از یک روش یکنواخت ممکن نیست و لذا طراحی استراتژی‌های مناسب در دستیابی به موفقیت ضروری می‌باشد.

توسعه روش‌های مؤثر برای هماهنگی سیستم‌های چندعامله در محیط‌های خصمانه یکی از جذاب‌ترین چالش‌های علمی است که توسط روبوکاپ[۱] ترویج داده شده‌است و به‌طورکلی توسط لیگ‌های شبیه‌ساز فوتبال حمایت می‌شود. هدف کلی از مکانیسم هماهنگی در این لیگ‌ها کنترل مناسب یک تیم از بازیکنان و داشتن اختیار برای بردن مسابقه در مقابل تیم حریف است.

در محیط چندعامله فوتبال، تصمیم‌گیری برای مهارت‌های سطح بالا توسط عامل‌ها بر عملکرد تیمی تأثیرگذار است. اگرچه هنوز واضح نیست که بیشتر کدام تکنولوژی پایه مانند برنامه‌ریزی تئوری تصمیم‌گیری، یا مهارت‌های سطح پایین، مهم‌ترین نقش را در موفقیت یک تیم روبوکاپ دوبعدی دارد. از طرفی استراتژی‌های موفق و تاکتیک‌های فوتبال تأثیر بسیاری بر مهارت‌های بازیکن دارد. برای اینکه عامل‌ها یک استراتژی را با موفقیت عملی نمایند، باید توانایی پاس‌کاری به تمام نقاط، موقعیت‌یابی[۲] مناسب برای حمله[۳] و دفاع[۴]، علامت‌گذاری[۵] حریف و مهارت‌های سطح پایین‌تر را دارا باشند. از جمله مهارت‌های سطح بالای چندعامله در فوتبال ربات‌ها استراتژی، تاکتیک و شکل‌گیری، مهارت تهاجمی و مهارت تدافعی است. ازآنجایی‌که این مقاله روی مهارت تهاجمی بازیکنان تمرکز دارد، توضیح مختصری از آن در زیر آمده‌است.

حمله به معنای گرفتن توپ و حرکت با توپ به سمت جلو و به سمت دروازه حریف است. استراتژی حمله معمولاً به‌این‌صورت تعریف می‌شود که گردش توپ بین هم‌تیمی‌ها تا زمان فعال‌شدن تابع شوت‌کردن، ادامه دارد [۲]. برای بهبود انتخاب موقعیت‌یابی در طول وضعیت حمله (یعنی تیم صاحب توپ)، بازیکن‌ها باید بهترین موقعیت قابل دسترسی جهت دریافت پاس یا امتیاز گل را پیدا کنند [۱].

اولین رقابت ربات‌های فوتبالیست در سال ۱۹۹۷ در اوزاکا ژاپن برگزار شد. در این دوره از مسابقات، استراتژی تیم‌ها کاملاً عمومی و سرراست بود. بیشتر تیم‌ها، بازیکن‌ها را در موقعیت ثابت نگه‌داشته و بازیکنان تنها به هنگام نزدیک‌شدن توپ به آن‌ها به طرف توپ حرکت می‌کردند. از این به بعد محققان زیادی به این زمینه علاقه‌مند شدند و روش‌های پیشرفته‌تری برای کنترل این مسئله ارائه کردند. به‌طوری‌که در سال ۱۹۹۸، روبوکاپ۹۸ برای اولین بار میزبان ارزیابی علم سیستم‌های چندعامله در رقابت‌های جهانی بود [۳].

در پژوهش‌های [۶-۴] برای کنترل ربات‌های فوتبالیست از یادگیری تقویتی استفاده شده‌است. پژوهش [۴] روشی است که روی تیم GPR2D تنها برای کنترل عامل صاحب توپ پیاده‌سازی شده‌است. در این روش عامل صاحب توپ برای انتخاب عمل خود براساس موقعیت هم‌تیمی‌ها و حریفان تصمیم‌گیری می‌کند. برای این منظور از الگوریتم یادگیری کیو[۶] استفاده شده‌است. در این پژوهش، ۷ حالت برای محیط و ۷ عمل برای عامل در نظر گرفته شده‌است. عامل‌ها در این روش طی ۱۰ بازی مقابل تیم هلیوس۲۰۱۱ آموزش دیدند. سپس تیم در ۱۰ بازی مقابل هلیوس۲۰۱۱ مورد آزمون قرار گرفت و به نتایج ۴ برد و ۶ مساوی رسید. قابل به ذکر است که ابتدا این روش روی تیم هلیوس۲۰۱۳ پیاده‌سازی گردید و باعث بهبود در نتایج نشد. پژوهش [۵] نیز روی تیم اکسیوم[۷] پیاده‌سازی شده که از یادگیری کیو برای کنترل عملکرد حمله استفاده کرده‌است. در این مقاله ۱۵ عمل برای عامل و $۱۰^۶$ حالت برای محیط در نظر گرفته شده‌است. در این پژوهش در واقع یک الگوریتم جدید برای پاس‌کاری وموقعیت‌یابی پویای عامل‌ها ارائه گردید، ولی به علت فضای حالت بزرگ با مشکل تنگنای ابعاد روبرو شد. در پژوهش [۶] از یادگیری سارسا[۸] تنها برای حل زیر مسئله HFO[۹] استفاده شده‌است. در زیر مسئله HFO به‌جای تمام بازیکنان، یک تیم مدافع و یک دروازه‌بان و یک تیم مهاجم مورد بررسی قرار می‌گیرند. در این پژوهش محیط به‌طور پیوسته در نظر گرفته شده‌است و همچنین برای عامل، ۷ عمل تعریف شده‌است.

در پژوهش [۷] به‌منظور پیش‌بینی موقعیت حریف برای هر بازیکن یک شبکه عصبی[۱۰] سه لایه با ۵ نرون[۱۱] در لایه ورودی و ۱۰ نرون در لایه پنهان و ۲ نرون در لایه خروجی تعریف شده‌است. وضعیت زمین در یک گام زمانی توسط یک بردار ۵ بعدی که شامل موقعیت جاری توپ، کمترین گام زمانی برای رسیدن به توپ میان همه بازیکنان و تخمین موقعیت توپ بعد از کمترین گام زمانی نمایش داده می‌شود. خروجی شبکه عصبی، موقعیت یک بازیکن حریف متناظر با وضعیت زمین را مشخص می‌کند که موقعیت حریف را حدوداً با ۱۰ میلی‌متر خطا پیش‌بینی می‌کند. ولی نتایج آزمایش‌ها با توجه به میانگین خطاها نشان می‌دهد که، روش ارائه‌شده خیلی بهتر از روش‌های پیش‌بینی سنتی عمل نمی‌کند. در [۵] نیز برای تخمین میزان امن‌بودن محیط از شبکه عصبی سه لایه استفاده شده‌است. در این پژوهش ۱۰ نرون در لایه خروجی و ۶ نرون در لایه پنهان و ۱ نرون برای لایه خروجی در نظر گرفته شده‌است. درنهایت این روش به عامل برای انجام پاس‌های قابل اعتمادتر و امن‌تر کمک شایانی کرد.

در پژوهش [۸] از استدلال مبتنی بر مورد[۱۲] و طبقه‌بند بیزی[۱۳] جهت کنترل استراتژی حمله و دفاع استفاده شده‌است. براساس موردهای ثابتی که تعریف شده‌است، با استفاده از طبقه‌بند بیزی، موردی که بیشترین شباهت را با حالت جاری دارد انتخاب می‌کند و براساس آن عمل می‌نماید. با توجه به این‌که در این روش از موارد ثابت استدلال مبتنی بر مورد برای تعیین وضعیت عامل در حالت جاری استفاده می‌شود، با شرایط پویا و پیچیده فوتبال ربات‌ها سازگار نیست.

در پژوهش‌های [۷، ۹] برای کنترل ربات‌های فوتبالیست از جستجوی درختی استفاده شده‌است. در [۹] به منظور کنترل رفتار پاس‌کاری بازیکنان و موقعیت‌یابی بازیکنان درخت جستجوی پیشرو[۱۴] بکار رفته است. در این روش با توجه به ویژگی‌های غیر قابل پیش‌بینی از محیط، موقعیت‌یابی چندگامی و رفتار پاس‌دادن تنها در یک افق خیلی محدود امکان‌پذیر است. پژوهش [۷] روشی است که توسط تیم هلیوس۲۰۱۳ پیاده‌سازی شده‌است. در این پژوهش برای کنترل رفتار تهاجمی عامل‌ها جستجوی اول بهترین[۱۵] بکار رفته است. از آنجایی که جستجوی اول بهترین استفاده شده‌است، همگرایی روش به سیاست بهینه تضمین نمی‌شود.

پژوهش [۱۰] روشی برای ارزیابی استراتژی عامل‌های فوتبال و کسب یک استراتژی مناسب با استفاده از الگوریتم ژنتیک ارائه کرده است. در این مقاله، هر عامل فوتبال دارای یک مجموعه قوانین عمل است که تنها هنگامی که توپ در ناحیه قابل ضربه‌زدن عامل است، استفاده می‌شود. این روش سرعت محاسباتی پایینی دارد و در محیط بی‌درنگ و پیچیده فوتبال ربات‌ها کارایی خوبی ندارد.

پژوهش [۱۱] نسخه بهبودیافته روش [۷] را ارائه کرده‌است که در آن از یک مدل فازی برای ارزیابی هر چه بهتر عمل‌ها استفاده شده‌است. در این پژوهش عامل استراتژی‌های خوب تیم‌های دیگر را یادمی‌گیرد و باعث بهبود توانایی پاس‌کاری عامل‌ها در مقایسه با حالت بدون این سیستم شده‌است.

روش دیگری که در [۱۰] ارائه گردید، از یک‌سری پارامتر برای تصمیم‌گیری جهت علامت‌گذاری و تحت‌فشار گذاشتن حریف در موقعیت حمله استفاده می‌کند. در اینجا تحت‌فشار گذاشتن، یک اقدام تاکتیکی جمعی است که در شرایط غیرتصاحبی توسط بیش از یک بازیکن انجام می‌شود. در اینجا هدف این است که فضا و زمان بازی برای تیمی که در تصاحب است بسته شود، این توسعه حرکات حمله آن‌ها را مشکل ساخته و بدست‌آوردن مجدد توپ را برای تیم آسان می‌سازد.

یکی از مزیت‌های یادگیری تقویتی سادگی در پیاده‌سازی است که با تعریف سه فاکتور حالت، عمل و سیگنال تقویتی برای مسئله موردنظر، به‌سادگی به حل مسئله می‌پردازد. اگرچه این برای دیگر روش‌های یادگیری ماشین ساده نیست [۱۲]. به همین علت در این مقاله برای کنترل و بهبود مهارت‌های حمله که هدف این تحقیق است از یادگیری تقویتی استفاده شده‌است. در مقاله حاضر تاکید بر روی سیستم‌های متشکل از چند عامل خودمختار است که می‌توانند در محیط‌های بی‌درنگ، نویزی، نیاز به همکاری و خصمانه با اهداف متقابل عمل کنند.

در بسیاری از کارهای پیشین، الگوریتم یادگیری کیو و سارسا استفاده شده‌است. در این مقاله سعی بر این است که با به‌کارگیری الگوریتم یادگیری کیو ـ وی[۱۶] عامل‌های درگیر در حمله در فوتبال ربات‌های دو بعدی آموزش ببینند. آزمایش‌هایی که در همین زمینه [۱۳] روی عملکرد الگوریتم‌های یادگیری کیو و کیو ـ وی انجام گرفت، نشان از برتری الگوریتم یادگیری کیو ـ وی نسبت به الگوریتم کیو برای حل مسئله ربات‌های فوتبالیست دارد. ازآنجایی‌که اختلاف انواع روش‌های یادگیری تقویتی در نحوه به‌روزرسانی توابع ارزش می‌باشد، از طرفی یادگیری کیو ـ وی از تابع ارزش حالت برای به‌روزرسانی تابع ارزش حالت ـ عمل خود استفاده می‌کند، دارای نتایج بهتری نسبت به دیگر روش‌های یادگیری تقویتی است. روش‌های ارائه‌شده در این مقاله بر روی آخرین کد پایه در دسترس از تیم هلیوس (هلیوس۲۰۱۳) پیاده‌سازی شده‌است. لازم به ذکر است این تیم در رقابت‌های چندساله اخیر در مسابقات جهانی روبوکاپ جزو تیم‌های اول و دوم بوده است.

ساختار این نوشتار در پنج بخش صورت گرفته است. در بخش دوم مفاهیم مربوط به یادگیری تقویتی و روش انتخاب عمل شرح داده می‌شود. همچنین توضیحاتی در ارتباط با محیط شبیه‌سازی که استفاده شده‌است، بیان شده‌است. بخش سوم شرح فرمول‌گذاری کردن مسئله آمده‌است. در بخش چهارم پیاده‌سازی و نتایج روش بیان شده‌است. و بخش آخر جهت بررسی نتایج و بیان پیشنهادات در نظر گرفته شده‌است.

۲- مفاهیم اولیه

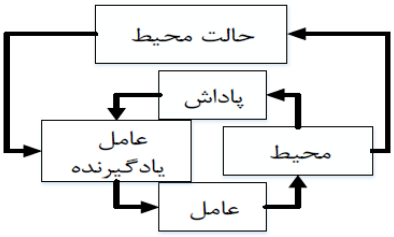
در این بخش به تشریح مفاهیم و روش‌های موردنیاز برای درک بهتر مطالب این مقاله پرداخته شده‌است.

۲-۱- یادگیری تقویتی

یادگیری تقویتی یک شاخه از یادگیری ماشین است که عامل یادگیرنده تلاش می‌کند تا انجام یک وظیفه مشخص را در یک محیط مشخص یاد بگیرد. هدف عامل بیشینه‌کردن مجموع پاداش دریافتی در طول اجرا می‌باشد [۱۴]. در یادگیری تقویتی یک یا چند عامل با توجه به تقابلی که با محیط دارند یک عمل را در یک حالت داده‌شده انتخاب می‌کنند تا به یک سیاست بهینه برسند. سیاست بهینه یادگرفته شده برای انجام این روش باید بهترین عمل را از بین عمل‌های قابل اجرا در هر موقعیت انتخاب کند [۱۵]. ساختار کلی یادگیری تقویتی در شکل ۱ قابل مشاهده است.

یک مسئله یادگیری تقویتی توسط سه فاکتور حالت، عمل و سیگنال تقویتی فرمول‌گذاری می‌شود. یک وضعیت در محیط توسط حالت مشخص می‌شود. حالت $s_t \in S$ وضعیتی از محیط در زمان t را تعریف می‌کند. در اینجا S مجموعه‌ای از حالت‌های ممکن است. عامل یادگیرنده در هر گام زمانی مجموعه‌ای از عمل‌هایی که می‌تواند در حالت جاری انتخاب کند، دریافت می‌کند. عامل یک عمل $a_t \in A$ را انتخاب می‌کند که A مجموعه‌ای از عمل‌های ممکن است. انجام یک عمل عامل را به

حالت جدید در گام زمانی بعدی منتقل می‌کند و عامل یک سیگنال تقویتی $r_t \in R$ به عنوان نتیجه‌ای از عملش دریافت می‌کند [۱۴].



شکل ۱: تعامل عامل و محیط در الگوریتم یادگیری تقویتی [۱۶]

تقریباً همه الگوریتم‌های یادگیری تقویتی با تخمین تابع ارزش حالت (یا حالت ـ عمل) سروکار دارند که تابع ارزش تخمین زده‌شده توسط عامل، نحوه انتخاب عمل عامل در یک حالت داده‌شده را مشخص می‌کند. باید توجه داشت که تابع‌های ارزش با توجه به سیاست‌های خاص تعریف می‌شوند. یک سیاست، π ، یک نگاشت از حالت $s \in S$ به عمل $a \in A$ است [۱۴].

به‌طور رسمی ارزش حالت s تحت سیاست π به‌صورت $V_\pi(s)$ نشان داده می‌شود، مقدار بازگشتی موردانتظار در صورتی‌که عامل با شروع از حالت $s_t = s$ تحت سیاست π به دست می‌آورد را نشان می‌دهد. تابع ارزش حالت، $V_\pi(s)$، به‌صورت رابطه (۱) تعریف می‌گردد.

$$V_\pi(s) = E_\pi[\sum_{k=0}^{\infty}\gamma^k r_{t+k+1} \,|s_t = s] \qquad (۱)$$

در اینجا $E_\pi[.]$ امید مقدار پاداش مورد انتظاری است که عامل بر اثر دنبال کردن سیاست π کسب می‌کند. $\gamma \in [0,1]$ نرخ تنزل می‌باشد که اهمیت نسبی پاداش‌های بلندمدت را نسبت به پاداش‌های کوتاه‌مدت تنظیم می‌کند. متغیر t نشان‌دهنده هر گام زمانی است. همچنین باید توجه داشت که ارزش حالت پایانی در هر صورتی، صفر می‌باشد.

به‌طور مشابه ارزش عمل a در حالت s تحت سیاست π ، به‌صورت $Q_\pi(s,a)$ نشان داده می‌شود، که مقدار بازگشتی موردانتظار با شروع از حالت $s_t = s$ و انتخاب عمل $a_t = a$، و دنبال‌کردن سیاست π می‌باشد. تابع ارزش حالت ـ عمل $Q_\pi(s,a)$ به‌صورت رابطه (۲) تعریف می‌شود.

$$Q_\pi(s,a) = E_\pi[\sum_{k=0}^{\infty}\gamma^k r_{t+k+1} \,|s_t = s, a_t = a] \qquad (۲)$$

در این مقاله الگوریتم یادگیری کیو ـ وی پیاده‌سازی شده‌است و نتایج آن مورد بررسی قرار گرفته است. در ادامه این بخش نحوه عملکرد الگوریتم یادگیری کیو ـ وی و نحوه تخمین تابع ارزش حالت ـ عمل مربوط به آن آمده‌است.

۲-۱-۱- یادگیری کیو ـ وی

یادگیری کیو ـ وی یک روش یادگیری تقویتی برسیاست[۱۷] است که تعمیمی از ایده‌های دو روش یادگیری کیو و سارسا می‌باشد. در این روش به‌جای نگهداری یک تابع ارزش، همزمان هر دو تابع ارزش حالت V و تابع ارزش حالت ـ عمل Q نگهداری می‌شود. که همین مسئله باعث بهبود عملکرد در روش یادگیری کیو ـ وی نسبت به دو روش یادگیری کیو و سارسا شده‌است [۱۴]. قاعده به‌روزرسانی تابع ارزش حالت و به‌روزرسانی تابع ارزش حالت ـ عمل به‌ترتیب در روابط (۳) و (۴) برای یادگیری کیو ـ وی آمده‌است:

$$V(s_t) = V(s_t) + \beta[r_t + \gamma V(s_{t+1}) - V(s_t)] \qquad (۳)$$

$$Q(s_t, a_t) = Q(s_t, a_t) + \alpha[r_t + \gamma V(s_{t+1}) - Q(s_t, a_t)] \qquad (۴)$$

در اینجا متغیرها به‌صورت زیر تعریف می‌شوند:

- s_t حالت جاری متناظر
- a_t عمل انتخاب شده در حالت s_t
- r_t سیگنال تقویتی دریافتی بر اثر انتخاب عمل a_t در حالت s_t
- s_{t+1} حالت بعدی
- γ ضریب تخفیف
- α و β نرخ آموزش

الگوریتم ۱ شبه کد الگوریتم یادگیری کیوـ وی را نمایش می‌دهد.

الگوریتم ۱: الگوریتم یادگیری کیو ـ وی

Initialize Q(s,a) arbitrarily
Initialize V(s) arbitrarily
Repeat (for each episode):
  Initialize S
  Repeat (for each step of episode):
  Choose a_{t+1} from s_{t+1} using policy derived from Q (e.g. ε-gready)
  take action a, observe r_{t+1}; s_{t+1}
  $Q(s_t, a_t) = Q(s_t; a_t) + [r_{t+1} + \gamma V(s_{t+1}) - Q(s_t, a_t)]$
  $V(s_t) = V(s_t) + \beta[r_{t+1} + \gamma V(s_t+1) - V(s_t)]$
  $s_t \longleftarrow s_{t+1}$
Until s_t is Terminal

۲-۲- انتخاب عمل به روش شبه‌حریصانه

یکی از سیاست‌ها که به‌طور گسترده مورد استفاده قرار می‌گیرد روش شبه‌حریصانه[۱۸] می‌باشد. در این روش با احتمال ε-1 (ε یک عدد حقیقی مثبت بین صفر و یک) عملی برگزیده می‌شود که دارای بالاترین مقدار ارزش باشد. دیگر عمل‌ها با احتمال ε با شانس مساوی انتخاب می‌شوند. احتمال انتخاب هر عمل در سیاست شبه‌حریصانه در رابطه (۶) آمده‌است [۱۴].

$$p(a) = \begin{cases} 1 - \varepsilon + \frac{\varepsilon}{N} & a = arg\max_{b \in A}(Q(s,b)) \\ \frac{\varepsilon}{N} & otherwise \end{cases} \qquad (۶)$$

در رابطه بالا N تعداد عمل‌ها می‌باشد و $arg\max_{b \in A}(Q(s,b))$ شماره عمل با بیشینه ارزش را برمی‌گرداند.

۲-۳- معرفی محیط شبیه‌ساز ربات‌های فوتبالیست

سیستم شبیه‌سازی که پیاده‌سازی‌ها و آزمون‌های این مقاله در آن صورت گرفته است، rcsoccersim15.1.1 می‌باشد. این سیستم شبیه‌ساز از سه بخش اصلی سرور فوتبال، نمایشگر فوتبال و نمایشگر مجدد بازی تشکیل شده‌است [۳]. سیستم سرور، نرم افزار اصلی در طراحی و مدل‌سازی محیط فوتبال می‌باشد. تمام اطلاعات مربوط به حسگرها، عملگرها، الگوریتم‌های ایجاد خطا و الگوهای مورد استفاده در

برقراری ارتباط میان عامل‌ها توسط این سیستم، تولید، پردازش و بین عامل‌ها، منتقل می‌شوند.

سیستم شبیه‌ساز، دو نمایشگر بازی برخط[۱۹] و برون خط[۲۰] نیز به‌همراه دارد که به‌طور مستقیم با سرور در تقابل است و اطلاعات ایجاد شده از سوی سرور را نمایش می‌دهد. روش نمایش برخط درهنگام اجرای یک مسابقه واقعی و روش نمایش برون‌خط در زمان آزمون و یا مطالعه روی نحوه عملکرد یک تیم در یک بازی انجام شده مورد استفاده قرار می‌گیرد. یکی دیگر از اجزای شبیه‌ساز، نمایشگر مجدد بازی است که اطلاعات مربوط به یک بازی انجام شده را مجدداً توسط نمایشگر به روش برون‌خط نمایش می‌دهد.

۳- روند انجام پژوهش و روش پیشنهادی

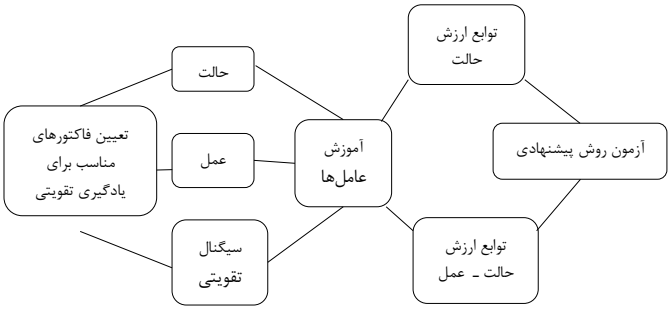
با توجه به مطالعاتی که صورت گرفت و آزمایشاتی که انجام شد یادگیری تقویتی برای بهبود عملکرد حمله تیم در نظر گرفته شد. از جمله ویژگی‌های یادگیری تقویتی که کارایی مناسب آن برای کار روی محیط پیچیده و بی‌درنگ فوتبال ربات‌ها را اثبات می‌کند، می‌توان به موارد زیر اشاره نمود:

- با تعریف سه فاکتور حالت، عمل و سیگنال تقویتی قابل فرمول‌گذاری است.
- یادگیری با یک معیار عددی (سیگنال تقویتی) صورت می‌گیرد.
- به عامل توانایی کسب رفتارهای پیشرفته با دانش کم را می‌دهد.
- دارای هزینه محاسباتی پایین می‌باشد.

روند کلی کار و الگوریتمی که در هر دو روش ارائه شده به‌ترتیب در شکل ۲ و الگوریتم ۲ آمده‌است. با توجه به شکل ۲ و الگوریتم ۲ در ابتدا باید سه فاکتور اصلی حالت، عمل و سیگنال تقویتی و همچنین مقادیر اولیه توابع ارزش حالت و ارزش حالت ـ عمل مشخص گردد. سپس عامل در طول بیست بازی طی مراحل آموزش مختلف آموزش می‌بینند تا به سیاست نهایی دست پیدا کند. سیاست نهایی نحوه انتخاب عمل عامل را در حالت‌های مختلف تعیین می‌کند. نحوه تعیین فاکتورهای یادگیری تقویتی که ایده اصلی این مقاله برای بهبود عملکرد حمله در فوتبال ربات‌ها می‌باشد در زیر بخش‌های بعدی آمده‌است.

الگوریتم۲: الگوریتم پیشنهادی

۱- گسسته‌سازی حالت‌ها و عمل‌ها
۲- تعیین سیاست دلخواه برای هر حالت و مقداردهی پارامترها
۳- شروع حلقه ۱ (به ازای هر بازی)
۴- شروع حلقه ۲ (به ازای هر مرحله آموزش)
۵- انتخاب عمل برای هر حالت
۶- به‌کارگیری عمل‌ها در محیط
۷- دریافت ادراک از محیط
۸- تعیین حالت و سیگنال تقویتی
۹- به‌روزرسانی توابع ارزش حالت و ارزش حالت ـ عمل
۱۰- اگر در حالت پایانی قرار ندارد یا توپ از منطقه جریمه خارج‌نشده برو به گام ۴
۱۱- اگر بازی به پایان نرسیده برو به گام ۳
۱۲- برگرداندن سیاست



شکل۲: روند انجام پژوهش

۳-۱- فرمول‌گذاری کردن مسئله

جهت پیاده‌سازی الگوریتم یادگیری کیو ـ وی در ابتدا لازم است تا فضای حالتی که عامل قادر به درک آن است گسسته‌سازی شود. ازآنجایی‌که حالت‌هایی که عامل صاحب توپ و عامل‌های بدون توپ درک می‌کنند با هم متفاوت است، عامل‌ها در دو دسته عامل صاحب توپ و عامل بدون توپ در نظر گرفته شد. جهت گسسته‌سازی حالت‌ها با توجه به اطلاعاتی که عامل از محیط اطراف دارد، برای عامل صاحب توپ یک بردار ۶ بیتی و برای عامل بدون توپ یک بردار ۳ بیتی جهت تعیین ویژگی‌های محیط برای حالت متناظر عامل در نظر گرفته شده‌است. در زیر شرح کاملی از ویژگی‌های در نظر گرفته‌شده و نحوه گسسته‌سازی فضای حالت آمده‌است.

بردار ۶ بیتی که جهت گسسته‌سازی حالت‌های عامل صاحب توپ در نظرگرفته شد به شرح زیر می‌باشد:

- بیت اول: نشان‌دهنده قابلیت پاس‌دادن توسط عامل صاحب توپ در حالت جاری است.
- بیت دوم: نشان‌دهنده حضور یا عدم‌حضور عامل صاحب توپ روبروی دروازه است.
- بیت سوم: نشان‌دهنده وضعیت نزدیک‌ترین بازیکن حریف است، یعنی بیان می‌کند موقعیت حریف پشت یا جلو بازیکن است.
- بیت چهارم: اگر فاصله نزدیک‌ترین بازیکن حریف تا بازیکن صاحب توپ کمتر از ۲ متر باشد این بیت ۱ می‌شود.
- بیت پنجم: اگر فاصله نزدیک‌ترین بازیکن حریف تا بازیکن صاحب توپ بین ۲ تا ۴ متر باشد این بیت ۱ می‌شود.
- بیت ششم: اگر فاصله نزدیک‌ترین بازیکن حریف تا بازیکن صاحب توپ بیشتر از ۴ متر باشد این بیت ۱ می‌شود.

ازآنجایی‌که تنها یکی از ۳ بیت آخر به‌طور همزمان یک می‌شود، ۲۴ حالت به دست می‌آید، که بر حسب مقادیر از ۰ تا ۲۳ شماره‌گذاری می‌شوند.

بردار ۳ بیتی که جهت گسسته‌سازی حالت‌های عامل‌های بدون توپ در نظرگرفته شد به شرح زیر می‌باشد:

- بیت اول: بیانگر این است که بازیکن سریع‌ترین بازیکن به توپ (یعنی نزدیک‌ترین بازیکن به بازیکن صاحب توپ) هست یا خیر
- بیت دوم: اگر بازیکن توانایی دریافت پاس از بازیکن هم‌تیمی خود را داشته باشد این بیت یک می‌شود.
- بیت سوم: نشان دهنده وضعیت نزدیک‌ترین بازیکن حریف است که پشت یا جلو بازیکن قرار دارد.

در نتیجه ۸ حالت برای عامل بدون توپ به وجود می‌آید که از ۰ تا ۷ شماره‌گذاری می‌شود.

مجموعه عمل‌های عامل صاحب توپ به شرح زیر است:

-Slow_Forward: حرکت آهسته همراه با توپ

- Fast_Forward: حرکت سریع همراه با توپ

- Pass: پاس‌دادن توپ به نزدیک‌ترین بازیکن هم‌تیمی

- Dribble: تلاش برای دریبل‌کردن بازیکن حریف

- Hold_ball: نگه‌داشتن توپ

- Kick: ضربه مستقیم به توپ به سمت دروازه حریف

در حالت‌هایی که بیت سوم صفر باشد یعنی نزدیک‌ترین بازیکن حریف پشت سر بازیکن صاحب توپ قرار دارد عمل Dribble از عمل‌های عامل حذف شده‌است. زیرا عمل دریبل‌کردن حریف تنها هنگامی‌که بازیکن حریف جلوی توپ قرار دارد معنا پیدا می‌کند. بنابراین برای عامل صاحب توپ ۶ عمل درحالتی‌که نزدیک‌ترین بازیکن حریف جلوی عامل قرار دارد و ۵ عمل درحالتی‌که نزدیک‌ترین بازیکن حریف پشت سر عامل قرار دارد در نظر گرفته شده‌است.

مجموعه عمل‌های عامل‌های بدون توپ به شرح زیر است:

Get_Ball: موقعیت‌یابی در بهترین موقعیت جهت به‌دست‌آوردن توپ

Get_Open: موقعیت‌یابی پایه تیم

ازآنجایی‌که قسمت اصلی حمله در منطقه جریمه صورت می‌گیرد (مستطیل قرمز رنگی که در شکل ۳ مشاهده می‌کنید)، در این شبیه‌سازی نیز عامل تنها با ورود توپ به این محوطه آموزش می‌بینند و از روش یادگیری تقویتی ارائه‌شده استفاده می‌کند.

در این مقاله تمام پیاده‌سازی‌ها و بهبودها بر روی کد پایه هلیوس۲۰۱۳ انجام شده‌است. روش یادگیری کیو ـ وی با انتخاب عمل به روش شبه‌حریصانه پیاده‌سازی شده‌است. با توجه به مطالعاتی که روی پژوهش‌های پیشین انجام شد، برای پارامترهای مربوط به الگوریتم یادگیری کیو ـ وی یعنی، γ، α و β به ترتیب مقادیر ثابت ۰/۸، ۰/۵ و ۰/۵ و برای پارامتر ε در روش انتخاب عمل شبه‌حریصانه مقدار ۰/۲ در نظر گرفته شد. بازیکنان حمله با دو رویکرد مختلف، به‌طور جداگانه آموزش دیدند و عملکرد آن‌ها نسبت به یکدیگر و تیم پایه هلیوس۲۰۱۳ مورد آزمون قرار گرفت. شرح مفصلی از روش‌های پیاده‌سازی شده در ادامه آمده‌است.

شکل ۳: نمایش منطقه جریمه، ناحیه آموزش عامل صاحب توپ

۳-۱-۱- توزیع یکسان سیگنال تقویتی

در اولین روشی که پیاده‌سازی شد تمام عامل‌های یادگیری تقویتی به‌طور یکسان سیگنال تقویتی دریافت می‌کنند. درواقع به‌هنگام رسیدن سیگنال پاداش یا جریمه تمام عامل‌های درگیر در حمله سیگنال تقویتی را به‌طور یکسان دریافت می‌کنند. با ورود توپ به منطقه جریمه آموزش عامل شروع می‌شود و یک مرحله آموزش آغاز می‌شود و اگر هر یک از سه حالت پایانی زیر اتفاق بیفتد مرحله به پایان می‌رسد و عامل‌ها سیگنال تقویتی به شرح زیر را دریافت می‌کنند.

- اگر بر اثر عمل توسط عامل، توپ از منطقه جریمه خارج شود، عامل جریمه ۵- دریافت می‌کند.
- اگر بر اثر عمل عامل، توپ از دست تیم خودی خارج شود و به دست تیم حریف برسد، عامل جریمه ۵۰- را دریافت خواهد کرد.
- اگر عامل در حالتی قرار بگیرد که روبروی دروازه باشد و حریف در مقابل آن قرار نداشته باشد و قابلیت پاس دادن هم وجود نداشته باشد (حالت‌های ۶،۷ و ۸). در این‌صورت عامل عمل kick را انتخاب می‌کند و پاداش ۱۰۰ را دریافت خواهدکرد. دراین‌مورد جهت همکاری بهتر عامل‌ها و پاس‌کاری بیشتر بین عامل‌ها قابلیت پاس‌دادن مورد بررسی قرار گرفته‌است.

به‌علاوه ازآنجایی‌که در یادگیری تقویتی عامل به‌دنبال بیشینه‌کردن پاداش دریافتی در طول اجرا می‌باشد، برای اینکه عامل در گام زمانی کمتری به گل برسد در هر گام زمانی جریمه ۰/۵- برای عامل در نظر گرفته شده‌است.

عامل‌ها در طول ۲۰ بازی مقابل تیم هلیوس۲۰۱۳ آموزش دیدند تا تابع Q عامل‌ها به سیاست نهایی همگرا شود. جهت همگرایی سریع‌تر

آموزش به سیاست نهایی در انتهای هر بازی از تابع ارزش حالت ـ عمل به‌دست‌آمده توسط هر یک از بازیکنان حمله در حالت بدون توپ میانگین گرفته و به عنوان تابع ارزش حالت ـ عمل اولیه عامل در بازی بعد در نظر گرفته‌شد.

۳-۱-۲- تقسیم سیگنال تقویتی براساس خبرگی

دانش در عامل یادگیری تقویتی نمایانگر این است که عامل کاوش بیشتری روی محیط داشته‌است و تجربیات بیشتری کسب کرده‌است. درواقع دانش بیشتر نشان‌دهنده این است که عامل جفت حالت ـ عمل بیشتری را ملاقات کرده‌است. معیاری که در این مقاله جهت بررسی میزان خبرگی عامل‌ها در نظر گرفته شده‌است، به‌صورت رابطه (۷) است که در آن دانش عامل با محاسبه اختلاف میان بیشینه و کمینه مقادیر در تابع ارزش حالت ـ عمل عامل در یک حالت داده‌شده، سنجیده می‌شود. با توجه به این رابطه در واقع عاملی که اختلاف میان بیشینه و کمینه مقادیر در تابع ارزش حالت ـ عملش کمتر باشد دانش کمتری دارد.

$$C_i^K = \max_{i=1,\dots,n}(Q_i(s_t,a)) - \min(Q_i(s_t,a)) \qquad (۷)$$

ایده اصلی در این روش براین‌اساس بود که، به‌هنگامی‌که عملکرد یک سیستم چندعامله به‌درستی نمی‌باشد تمام عامل‌ها مقصر نمی‌باشند، و عامل با دانش کمتر احتمال خطای بیشتری دارد. بر همین اساس جهت تقسیم بهتر سیگنال تقویتی بین عامل‌ها و آموزش بهتر بازیکنان به هنگام رسیدن جریمه ۵۰- در طول بازی‌ها هر عامل براساس دانشی که دارد مطابق با رابطه (۸) جریمه دریافت می‌کند و بقیه سیگنال‌های تقویتی مشابه با روش قبل توزیع می‌گردد. به واقع در این روش عاملی که دانش کمتری دارد جریمه بیشتری نسبت به عاملی که دانش بیشتری دارد دریافت می‌کند. این امر باعث می‌شود عامل‌ها تلاش بیشتری برای کسب دانش پیدا کنند و با سرعت بیشتری به سیاست نهایی همگرا شوند. به‌خصوص در ابتدای آموزش که عامل‌ها دانش کمتری دارند باعث افزایش سرعت آموزش می‌گردد. عامل‌ها در این روش نیز در طول ۲۰ بازی در مقابل تیم هلیوس۲۰۱۳ آموزش دیدند تا تابع Q به سیاست نهایی همگرا گردد.

$$P_i = \left(1 - \frac{\max(Q_i(s_t,a)) - \min(Q_i(s_t,a))}{\max(\max(Q_j(s_t,a)) - \min(Q_j(s_t,a)))}\right) * (-50) \qquad (۸)$$
$$i = 1,\dots,n \quad j = 1,\dots,n$$

۴- نتایج پیاده‌سازی

برای آزمون روش‌هایی که در بالا شرح داده‌شد ۶۰ بازی به‌طور جداگانه بین روش توزیع یکسان سیگنال تقویتی بین عامل‌ها و تیم هلیوس۲۰۱۳ و ۶۰ بازی میان روش تقسیم سیگنال تقویتی بین عامل‌ها براساس خبرگی و تیم هلیوس۲۰۱۳ برگزار شد. نتایج بازی‌های انجام‌شده جهت آزمون روش‌ها به‌صورت جدول ۱ به‌دست آمد. جهت مقایسه بهتر بین نتایج، ستون امتیاز آمده‌است. برای محاسبه امتیاز هر تیم در طول ۶۰ بازی، امتیاز ۳ برای برد، امتیاز ۱ برای مساوی و امتیاز ۰ برای باخت در نظر گرفته شده‌است و مجموع آن‌ها محاسبه گردیده‌است.

جدول ۱: نتایج بازی روش‌های ارائه شده مقابل تیم هلیوس۲۰۱۳

تیم‌ها	برد	مساوی	امتیاز
توزیع یکسان سیگنال تقویتی|هلیوس	۳|۴۹	۸	۱۷|۱۵۵
تقسیم سیگنال تقویتی براساس خبرگی|هلیوس	۳|۵۱	۶	۱۵|۱۵۹

نتایج پیاده‌سازی نشان می‌دهد که روش‌های ارائه‌شده نتایج بهتری درمقابل تیمی که آموزش ندیده است، دارند و باعث بهبود عملکرد کلی تیم شده‌است. همچنین روش تقسیم سیگنال تقویتی براساس خبرگی باعث می‌شود، عامل سریع‌تر به هدف و سیاست نهایی برسد و باعث بهبود در سرعت آموزش عامل‌ها شده‌است. به‌علاوه روش ارائه‌شده باعث افزایش دانش به‌دست‌آمده در طول یادگیری به‌خصوص در اوایل یادگیری و در نتیجه افزایش سرعت یادگیری می‌شود.

به‌منظوراینکه صحت نتایج بالا یعنی سرعت بیشتر آموزش در روش تقسیم سیگنال تقویتی براساس خبرگی نسبت به روش توزیع یکسان سیگنال تقویتی نشان داده‌شود، آموزش بجای ۲۰ بازی با ۱۵ بازی تکرار شد. پس از ۱۵ آموزش در مقابل تیم هلیوس۲۰۱۳، عملکرد هر دو تیم مقابل هم با ۶۰ بازی مورد آزمون قرار گرفت. نتایج این ۶۰ آزمون در جدول ۲ آمده‌است. جدول ۲ نشان‌دهنده نتایج آزمون ۱ است، یعنی حالتی‌که در آن تیم‌ها طی ۲۰ بازی آموزش دیدند. جدول ۲ همچنین نشان‌دهنده نتایج آزمون ۲ است، یعنی حالتی‌که تیم‌ها طی ۱۵ بازی آموزش دیدند. در هر دو آزمون، دو تیم آموزش دیده در حالت توزیع یکسان سیگنال تقویتی و حالت تقسیم سیگنال تقویتی براساس خبرگی مقابل هم در ۶۰ بازی مورد آزمون قرار گرفتند.

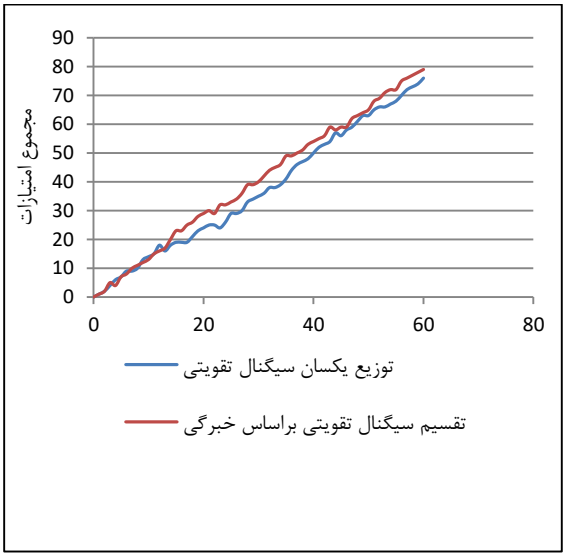
نتایج جدول ۲ نشان‌دهنده همگرایی سریع‌تر روش تقسیم سیگنال تقویتی براساس خبرگی به سیاست نهایی نسبت به روش توزیع یکسان سیگنال تقویتی می‌باشد. همان‌طورکه پیداست روش تقسیم سیگنال تقویتی براساس خبرگی موفق به کسب امتیاز بیشتری نسبت به روش توزیع یکسان سیگنال تقویتی در ۱۵ آموزش شد که نشانگر عملکرد بهتر و همگرایی سریع‌تر به سیاست نهایی در روش تقسیم سیگنال تقویتی براساس خبرگی می‌باشد.

جدول ۲: نتایج ۶۰ تست روش تقسیم سیگنال تقویتی براساس خبرگی در مقابل روش توزیع یکسان سیگنال تقویتی

تیم‌ها	برد	مساوی	باخت	امتیاز
آزمون ۱ (آموزش تیم‌ها طی ۲۰ بازی)	۲۲	۲۱	۱۷	۸۷
آزمون ۲ (آموزش تیم‌ها طی ۱۵ بازی)	۴۴	۱۰	۶	۱۴۲

برای مقایسه بهتر روش‌ها نمودار شکل ۴ آمده‌است که مجموع امتیازات کسب‌شده توسط تیم آموزش‌دیده با هر دو روش توزیع یکسان سیگنال تقویتی و توزیع سیگنال تقویتی براساس خبرگی، که در این

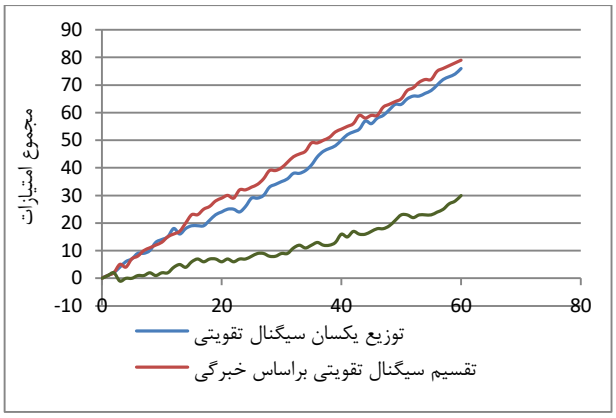
مقاله ارائه گردید، در مقابل تیم هلیوس۲۰۱۳ (بعد از ۲۰ بازی آموزشی) را نشان می‌دهد. در اینجا منظور از مجموع امتیارات، اختلاف گل زده توسط روش ارائه‌شده از گل خورده از تیم هلیوس۲۰۱۳ در طول ۶۰ بازی می‌باشد.



شکل ۴: مقایسه مجموع امتیازات کسب‌شده روش‌های ارائه‌شده در ۶۰ بازی مقابل تیم هلیوس۲۰۱۳

چندین مرحله شبیه‌سازی انجام‌شد. در اولین مرحله شبیه‌سازی‌ها ابتدا عامل‌های صاحب توپ آموزش دیدند[۱۳]. نتایج مقایسه با بهترین نتایج از این مرحله از شبیه‌سازی در نمودار شکل ۵ قابل مشاهده است.

همچنین جهت انجام آزمون‌های بیشتر روی روش‌های ارائه‌شده دو روش پیشنهادی ۶۰ بار مقابل کد پایه آزادشده توسط تیم رایت‌ایگل[۲۱] در سال ۲۰۱۳ قرار گرفت. تیم رایت‌ایگل در روبوکاپ‌های سال‌های اخیر جز تیم‌های برتر جهان بوده‌است. نتایج حاصل از این ۶۰ بازی به‌ترتیب در جدول ۳ و نمودار شکل ۶ قابل مشاهده است.

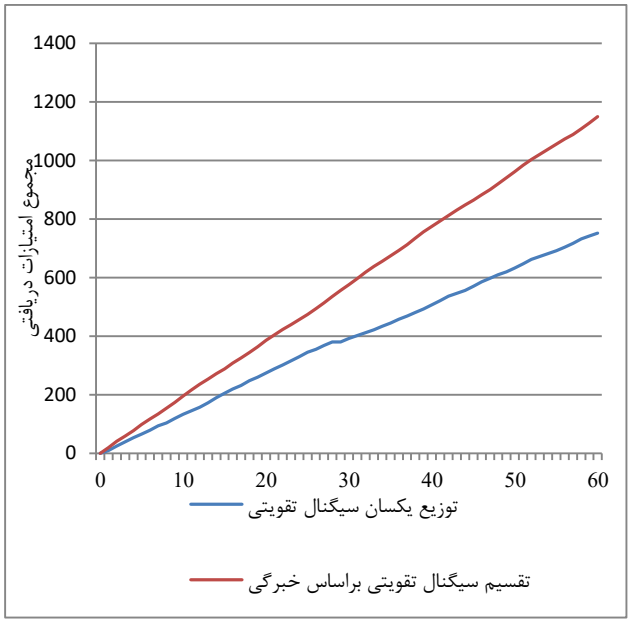


شکل ۵:مقایسه مجموع امتیازات کسب‌شده در ۶۰ بازی مقابل تیم هلیوس۲۰۱۳

نتایج آزمون‌ها حاکی از آن است که روش‌های ارائه‌شده توانایی مقابله با تیم‌های برتر جهانی و حریفان مختلف را دارد و بهبود خوبی روی عملکرد تیم داشته‌است. همچنین عامل‌ها توانایی انتخاب بهترین عمل در هر حالت را با کمک سیاستی که در طول آموزش کسب کردند، دارند.

جدول ۳: نتایج بازی روش‌های ارائه‌شده مقابل تیم رایت‌ایگل۲۰۱۳

تیم‌ها	برد	مساوی	امتیاز
توزیع یکسان سیگنال تقویتی|رایت‌ایگل	۰|۵۹	۱	۱|۱۷۸
تقسیم سیگنال تقویتی براساس خبرگی|رایت‌ایگل	۰|۶۰	۰	۰|۱۸۰

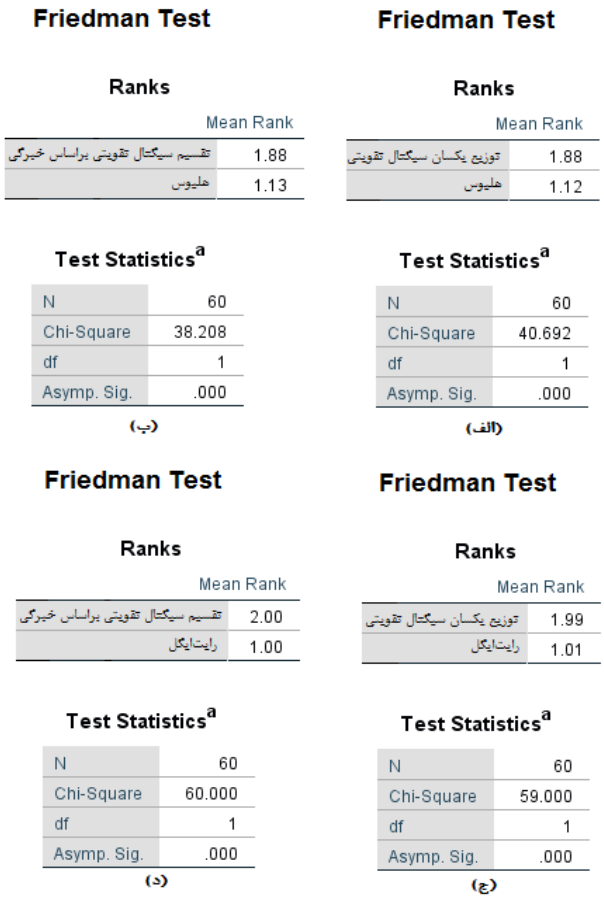


شکل ۶: مقایسه مجموع امتیازات کسب‌شده روش‌های ارائه‌شده در ۶۰ بازی مقابل تیم رایت‌ایگل۲۰۱۳

برای مقایسه بهتر نتایج از لحاظ آماری، از آزمون فریدمن جهت رتبه‌بندی روش‌ها استفاده شده‌است. آزمون فریدمن که به آزمون تحلیل واریانس معروف است، یکی از آزمون‌های آماری است که از آن برای مقایسه میانگین رتبه‌های K متغیر یا گروه استفاده می‌شود. در اینجا با اعمال آزمون فریدمن برروی نتایج آزمون‌ها به رتبه‌بندی روش‌ها پرداخته شده‌است. برای این منظور تعداد گل‌هایی که هر تیم در طول آزمون زده‌اند، ملاک رتبه‌بندی قرار گرفته‌است. نتایج آزمون که با استفاده از نرم‌افزار spss محاسبه شده‌است، به شرح شکل ۷ است.

در شکل ۷ جدول Rank میانگین رتبه‌های هر تیم را نشان می‌دهد. جدول Test Statistics محتوی نتایج اصلی آزمون است که مقدار آماره کی‌دو با یک درجه آزادی را نشان می‌دهد. همچنین سطح معنی‌داری آزمون را برابر با مقدار صفر نشان می‌دهد که بیانگر این است که بین میانگین رتبه‌ها اختلاف معناداری وجود دارد. با توجه به جداول Rank در شکل ۷ هر دو روش ارائه‌شده امتیاز بیشتری مقابل حریفان خود

کسب نمودند. براین‌اساس نتایج آزمون فریدمن بیانگر عملکرد بهتر روش‌های ارائه‌شده می‌باشد.



شکل ۷: نتایج آزمون فریدمن، الف) روش توزیع یکسان سیگنال تقویتی مقابل تیم هلیوس، ب) روش تقسیم سیگنال تقویتی براساس خبرگی مقابل تیم هلیوس، ج) روش توزیع یکسان سیگنال تقویتی مقابل تیم رایت‌ایگل، د) روش تقسیم سیگنال تقویتی براساس خبرگی مقابل تیم رایت‌ایگل

۵- نتیجه‌گیری

در این مقاله روشی برای بهبود عملکرد حمله در تیم شبیه‌ساز فوتبال ربات‌ها با استفاده از یادگیری تقویتی ارائه‌شد. به‌همین‌منظور الگوریتم یادگیری کیو ـ وی روی تیم ربات‌های فوتبالیست پیاده‌سازی شد و به آموزش عامل‌های حمله پرداخته شد. برای آموزش عامل‌ها از دو روش مختلف جهت تقسیم سیگنال تقویتی بین عامل‌ها استفاده شد. در روش اول سیگنال تقویتی به‌طور یکسان بین عامل‌ها توزیع گردید و در روش دوم از ایده تقسیم سیگنال تقویتی برحسب میزان خبرگی عامل‌ها بهره گرفته‌شد. براساس این ایده عامل با دانش کمتر جریمه بیشتری نسبت به عامل با دانش بیشتر دریافت می‌کند. نتایج پیاده‌سازی‌ها نشان داد که هر دو روشی که برای آموزش عامل‌های حمله در این مقاله پیاده‌سازی شد نتایج بهتری نسبت به حالتی که عامل‌ها آموزش ندیده‌اند، داشت. همچنین باعث بهبود عملکرد تیم در مقابله با حریفان مختلف شد. به‌علاوه در روش دوم که عامل‌ها براساس سطح خبرگی خود سیگنال تقویتی دریافت کردند، همگرایی به سیاست نهایی سریع‌تر صورت گرفت و باعث افزایش سرعت در آموزش عامل‌ها گردید. با توجه به پیاده‌سازی‌ها روش دوم یعنی تقسیم سیگنال تقویتی براساس خبرگی، عملکرد بهتری نسبت به روش اول یعنی توزیع یکسان سیگنال تقویتی داشت. به‌طورکلی روش تقسیم سیگنال تقویتی براساس خبرگی باعث افزایش سرعت آموزش عامل‌ها در طول مراحل یادگیری و بهبود عملکرد تیم نسبت به تیمی که عامل‌های حمله آن آموزش ندیده‌اند، شد.

روشی که در این مقاله ارائه شد تنها برای بهبود عملکرد بازیکنان حمله در حساس‌ترین منطقه حمله یعنی منطقه نزدیک به دروازه ارائه شد. برای کارهای آینده می‌توان مسئله یادگیری تقویتی را به‌نحوی برای آموزش کل بازیکن‌ها و در کل زمین بازی فوتبال تعریف کرد. همچنین می‌توان با تعریف جزئی‌تر مسئله روی رفتارهای سطح پایین عامل‌ها، روش ارائه‌شده را برای کنترل رفتار بازیکنان در فوتبال شبیه‌ساز سه‌بعدی تعمیم داد.

مراجع

[1] F. Almeida, N. Lau, and L. P. Reis, ”A Survey on Coordination Methodologies for Simulated Robotic Soccer Teams,” RoboCup Symposium, 2010.

[2] M. Alavi, M. F. Tarazkouhi, A. Azaran, A. Nouri, S. Zolfaghari, and H. R. S. Boroujeni, ”Robocup 2012- Soccer Simulation League 2D Soccer Simulation Riton,” Robot Soccer World Cup, Springer Berlin Heidelberg, 2013.

[3] M. Chen et. al., (2003) RoboCup Soccer Server for Soccer Server Tersion cefc and later, [Online], Available: http://wwfc.cs.virginia.edu/documentation/manual.pdf [jol. 11, 2015]

[4] J. R. F. Neri, M. R. Zatelli, C. H. F. dos Santos, and J. A. Fabro, ”A Proposal of QLearning to Control the Attack of a 2D Robot Soccer Simulation Team,” Robotics Symposium and Latin American Robotics Symposium (SBR-LARS), pp. 174–178, 2012.

[5] M. Ghazanfari, S. O. Shirkhorshidi, and F. Samsamipour, ”Axiom 2013 Team Description Paper,” Robot Soccer World Cup, Springer Berlin Heidelberg, vol. 8371, 2014.

[6] S. Kalyanakrishnan, Y. Liu, and P. Stone, ”Half field offense in RoboCup soccer: A multiagent reinforcement learning case study,” Robot Soccer World Cup, Springer Berlin Heidelberg, vol. 4434, pp. 72–85, 2008.

[7] H. Akiyama, T. Nakashima, and K. Yamashita, ”Helios2013 team description paper,” Robot Soccer World Cup, Springer Berlin Heidelberg, vol. 8371, 2014.

[8] T. Sirinivasan, K. Aarthi, S. A. Meenakshi, and M. Kausalya, ”Cbrrobosoc: An efficient planning strategy for robotic soccer using case based reasoning,” International Conference on Computational Intelligence for Modeling Control and Automation, and International Conference on Intelligent Agents, Web Technologies and Internet Commerce, pp. 113–119, 2006.

[9] A. Bai, H. Zhang, G. Lu, M. Jiang, and X. Chen, ”WrightEagle 2D Soccer Simulation Team Description,” Robot Soccer World Cup, Springer Berlin Heidelberg, vol. 7500, 2013.

[10] S. Marian, D. Luca, B. Sarac, and O. Cotarlea, ”OXSY 2014 Team Description,” Robot Soccer World Cup, Springer Berlin Heidelberg, 2015.

[11] H. Akiyama, T. Nakashima, and K. Yamashita, ”HELIOS2014 Team Description Paper,” Robot Soccer World Cup, Springer Berlin Heidelberg, 2015.

[12] M. Yoon, *Developing basic soccer skills using reinforcement learning for the RoboCup Small Size League*, Master Thesis, Stellenbosch University, pp.11, March 2015.

[۱۳] مینا خاکسار، ولی درهمی و مهدی رضائیان، «بهبود عملکرد حمله در تیم ربات‌های شبیه‌ساز فوتبال با استفاده از یادگیری تقویتی»، دومین کنفرانس محاسبات تکاملی و هوش جمعی، دانشگاه شهید باهنر، کرمان، اسفند ۹۵.

[14] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*, MIT press Cambridge, 1998.

[۱۵] عادل اکبری مجد، حسین شایقی، حمید محمدنژاد، عبداله یونسی، «کنترل کننده مقاوم تطبیقی بار فرکانس مبتنی بر یادگیری تقویتی برای یک سیستم قدرت به‌هم پیوسته شامل SMES» مجله مهندسی برق دانشگاه تبریز، جلد ۴۷، شماره ۲، صفحات ۳۸۱-۳۹۰، تابستان ۱۳۹۶.

[۱۶] مریم رمضانیان لنگرودی، سیدمازیار میرحسینی مقدم، بهنام علیزاده، «استفاده از روش یادگیری رقابتی برای قیمت‌دهی استراتژیک شرکت‌های تولید براساس LMP در بازار برق»، مجله مهندسی برق دانشگاه تبریز، جلد ۴۷، شماره ۲، صفحات ۵۳۷-۵۴۹، تابستان ۱۳۹۶.

زیرنویس‌ها

۱ RoboCup
۲ Positioning
۳ Offensive
۴ Defensive
۵ Marking
۶ Q - Learning
۷ Axiom
۸ Sarsa
۹ Half Field Offensive
۱۰ Neural Network
۱۱ Neuron
۱۲ Case Based Reasoning
۱۳ Bayesian Classifier
۱۴ Forward Search Tree
۱۵ Best First Search
۱۶ QV-Learning
۱۷ On-Policy
۱۸ ε-Greedy
۱۹ Online
۲۰ Offline
۲۱ WrightEagle