پژوهشنامه اقتصاد انرژی ایران
سال هفتم، شماره 25، زمستان 1397 صفحات 40-1

کاربرد یادگیری تقویتی در یک مدل‌سازی عامل‌محور* برای بازار عمده‌فروشی برق ایران

محمدرضا اصغری اسکوئی[1]
فرهاد فلاحی[2]
میثم دوستی‌زاده[3]
سعید مشیری[4]

تاریخ دریافت:1397/02/26 تاریخ پذیرش:1397/06/20

چکیده:

مطالعات اخیر بازارهای عمده‌فروشی برق عموماً براساس مدل‌های چندعاملی است، که در آن‌ها تعادل بازار برپایه رقابت و تعامل عوامل متعدد با یک دیگر به دست می‌آید. از ویژگی‌های اصلی این نوع مدل‌ها، امکان یادگیری عوامل از نتایج رفتار خود و سایرین دریک محیط رقابتی است. در بازار عمده‌فروشی برق، هرعامل یک واحد تولیدکننده برق است که به صورت مستقل و هوشمند با سایر عامل‌ها برای عرضه برق با قیمت‌های پیشنهادی رقابت می‌کند. فرآیند قیمت‌گذاری را می‌توان یک بازی ایستا فرض نمود که هرروز تکرار می‌شود. در این بازی هر عامل قیمت پیشنهادی خود را مستقلا اعلام نموده و بهره‌بردار با توجه به تقاضای بار مصرفی و محدودیت‌ها، بهترین پیشنهادها را انتخاب می‌نماید. عامل به صورت عقلائی عمل نموده و با انتخاب استراتژی مناسب، به دنبال

---

*در تهیه این مقاله از منابع و امکانات پژوهشگاه نیرو استفاده شده که به این وسیله از مسئولین و کارشناسان محترم آن مجموعه به ویژه جناب آقای دکتر کیومرث حیدری قدردانی می شود.

1. استادیار دانشکده علوم ریاضی و رایانه، دانشگاه علامه طباطبائی (نویسنده مسئول)
oskoei@atu.ac.ir
2. دانشجوی دکتری مهندسی برق قدرت، دانشگاه شاهد و پژوهشگر پژوهشگاه نیرو
ffallah@nri.ac.ir
3. استادیار دانشکده فنی و مهندسی دانشگاه لرستان
m.doostizadeh@gmail.com
4. دانشیار اقتصاد، دانشگاه ساسکاچوآن، کانادا
moshiri.s@usask.ca

بیشینه نمودن سود بلندمدت خود است. در این راستا، عامل از قدرت یادگیری و بهبود استراتژی قیمت‌گذاری، که نقش بسیار تعیین کننده در موفقیت عامل دارد، استفاده می‌کند. یادگیری تقویتی یک روش کلاسیک است که در مدل‌های چندعاملی امکان یادگیری مبتنی بر سعی و خطا را فراهم می‌نماید. هدف این مقاله کاربرد و مطالعه روش‌های یادگیری تقویتی در مدل چندعاملی بازار برق ایران و مقایسه آن‌ها با دو استراتژی تصادفی و حریصانه است. در این مطالعه، میزان سود واحدها و زمان رسیدن به حالت تعادل به عنوان ملاک ارزیابی در نظر گرفته شده است. نتایج شبیه سازی نشان می‌دهد، استراتژی یادگیرنده سود عامل‌ها را به طور معناداری افزایش می‌دهد و سرعت همگرائی به حالت تعادل را بیشتر می‌کند.

طبقه‌بندی JEL: C6, C7, Q41

کلیدواژه‌ها: مدل‌سازی عامل‌محور، بازار برق، یادگیری تقویتی، نظریه بازی‌ها، ایران

1. مقدمه

برق به عنوان یکی از منابع انرژی نقش به سزایی دررشد و توسعه اقتصادی اجتماعی کشورها داشته و میزان دسترسی به برق یکی از معیارهای توسعه یافتگی کشور به حساب می آید (تومن و جملکوا[1]، 2003). ساختار سنتی بازارهای برق با توجه به شرایط تولید آن‌ها (به عنوان نمونه برق آبی و یا برق اتمی) در بسیاری از موارد انحصاری بوده و دولت‌ها با توجه به ملاحظات زیست محیطی و آثاراقتصادی واجتماعی برق به طور مستقیم یا غیرمستقیم در آن‌ها دخالت می‌کنند. در سال‌های اخیر با پیشرفت فناوری‌های جدید، بازارهای برق در بسیاری از کشورهای جهان شاهد تجدید ساختار و تحولات اساسی بوده و از یک ساختار متمرکز و یکپارچه به یک بازار آزاد و رقابتی به ویژه در بخش‌های تولید و توزیع تبدیل شده‌اند. هدف اصلی این‌تحول، افزایش کارآیی بازار برق و درنتیجه کاهش قیمت‌ها وافزایش رفاه مصرف کننده است. در عین حال، بازار عمده‌فروشی برق یک بازار نسبتاً پیچیده است و خصوصیاتی مانند ضرورت تعادل لحظه‌ای تولید و مصرف، محدودیت‌های تولید وانتقال، قیود فیزیکی پخش‌بار، هزینه بالای ذخیره انرژی وحساسیت پایین تقاضا نسبت به تغییرات قیمت برپیچیدگی آن افزوده و چالش های خاصی در تنظیم

1.Toman and Jemelkoma (2003)

بازارایجاد می‌کند. علاوه براین‌ها، مسائلی مانند حمایت از محیط زیست، هزینه بالای سرمایه‌گذاری اولیه و استفاده بهینه از منابع نقش مستقیم بر عملکرد بازیگران این بازار دارد. متاسفانه در شرایط رقابتی جدید، مدل های سنتی و کلاسیک که عوامل اقتصادی (تولیدکنندگان یا مصرف کنندگان) را همگون و فاقد تعامل فرض می‌کنند کارآیی لازم را نداشته و نتایج قابل قبولی ارایه نمی‌دهند. بازارهای جدید برق شامل تعداد زیادی از تولیدکنندگان ناهمگون با توجه به نوع فناوری و درجه ریسک پذیریشان هستند که در رقابت با یکدیگر وارد چرخه تولید می‌شوند. مدل‌سازی عامل‌محور[1] (ABM) رفتار واحدهای تولیدی ناهمگون را تبیین و اجازه تعامل و یادگیری در یک محیط پویا را فراهم می‌سازد، اما به علت وسعت مدل و پیچیدگی های زیاد آن نمی‌توان از یک راه حل تحلیلی برای به دست آوردن متغیرها در شرایط تعادلی استفاده کرد. با توجه به پیشرفت فناوری محاسبات، نتایج مدل‌های عامل محور را معمولا با روش‌های شبیه‌سازی در چارچوب سناریوهای مختلف می‌توان ارزیابی کرد.

به طور مشخص‌تر، بازار عمده‌فروشی برق از تعداد زیادی نیروگاه‌های بزرگ وکوچک که در قالبِ شرکت‌های تولیدی فعالیت دارند در کنار شرکت‌هایِ انتقال وتوزیع تشکیل شده‌است و نهاد قانون‌گذار و نهاد مستقل بهره‌بردار سیستم[2] (ISO)، مسئولیت وضع قوانین و کنترل و هدایت بازار را برعهده دارند. واحدهای تولیدکننده روزانه برنامه تولید پیشنهادی خود را برای هر ساعت به نهاد بهره‌بردار ارائه می‌دهند. این نهاد با توجه به پیش‌بینی تقاضای مصرف بیست وچهار ساعت آتی و با اجرای مکانیزم حراج، واحدهای برنده و قیمت بازار را مشخص می‌نماید و سپس با اجرای مکانیزم تسویه، فروش برق و تسویه حساب انجام می‌گیرد. با توجه به ضرورت تعادل مصرف وتولید درکل کشور و همچنین توزیع جغرافیائی واحدهای تولیدی و مصرف‌کننده، در مکانیزم حراج، علاوه بر میزان‌مصرف، قیود اقتصادی و قیود فنی مربوط به تولید، انتقال و توزیع نیز

1.Agent Based Model (ABM)

2.Independent System Operator (ISO)

لحاظ می‌شود. واحدهای تولیدی همواره به دنبال حفظ سهم بازار و کسب سود بیشتر هستند و تلاش می‌کنند با استفاده از بازخورد پیشنهادات قبلی، پیشنهادات جدید را به‌گونه‌ی تهیه کنند که سود خود را بیشینه کنند. در ضمن نهاد قانونگذار و بهره‌بردار نیز با وضع قوانین و اجراء صحیح مکانیزم‌های حراج و تسویه، درصدد حفظ ثبات و منافع خرد وکلان تمام ذینفعان بازار هستند. معمولا تمام اهداف در یک راستا نمی‌باشند و با توجه به هزینه زیاد سرمایه‌گذاری، تصمیم نادرست یا وضع قانون نادرست ممکن است سبب خسارت‌های فراوان شود. بنابراین، مدل‌سازی و شبیه‌سازی واکنش‌های بازار در تصمیم‌گیری‌ها از اهمیت بالائی برخوردار است و با توجه به پیچیدگی‌های بازار، کاربرد روش‌های محاسبات هوشمند برای تحلیل و پیش‌بینی اطلاعات بازار جایگاه ویژه و مهمی به خود اختصاص داده است (اصغری‌اسکوئی 1394).

مدل عامل‌محور با ارائه یک مدل محاسباتی و درنظرگرفتن رفتار عوامل و نحوه تعامل آنها امکان بازنمائی و شبیه‌سازی عملکرد بازار عمده‌فروشی برق را فراهم می‌کند. با طراحی و بکارگیری مدل‌های عامل محور، در واقع می‌توان تاثیر استراتژی‌های مختلف تصمیم‌سازی در واحدهای تولیدی اعم از برنامه‌تولید و سرمایه‌گذاری و همچنین تاثیر وضع قوانین جدید توسط نهادها را پیش از اجراء مورد بررسی و ارزیابی قرارداد. یکی از چالش‌های مهم واحدهای تولیدی، انتخاب استراتژی صحیح پیشنهاد قیمت یا برنامه‌تولید است. پیشنهاد قیمت یا برنامه‌تولید به صورت مستمر و روزانه ارائه می‌شود و واحدهای تولیدی لازم است با توجه به بازخورد بازار درباره ادامه استراتژی یا تغییر آن تصمیم‌گیری کنند. این مقاله به بررسی کاربرد یادگیری تقویتی در ارائه پیشنهاد و مقایسه عملکرد آن با استراتژی‌های دیگر در میزان سود واحدهای تولیدی می‌پردازد.

یادگیری تقویتی[1] از شاخه‌های یادگیری‌ماشین و محاسبات‌هوشمند بوده و یک یادگیری مبتنی بر سعی وخطا است که از روانشناسی رفتارگرایی الهام می‌گیرد. یادگیری

1.Reinforcement Learning

تقویتی در اقتصاد و نظریه‌بازی‌ها بیشتر به بررسی تعادل‌های ایجاد شده تحت عقلانیت محدود عامل‌ها می‌پردازد. در یادگیری ماشین با توجه به این که بسیاری از الگوریتم‌های یادگیری تقویتی از تکنیک‌های برنامه‌نویسی پویا استفاده می‌کنند معمولاً مسئله تحت عنوان یک فرایند تصمیم‌گیری مارکف مدل می‌شود. این مقاله از یک مدل عامل‌محور شناخته شده و معتبر بازار برق ایران استفاده نموده و ضمن ارائه توصیف ریاضی از استراتژی تصمیم‌گیری مبتنی بر یادگیری تقویتی، تاثیر عملکرد آن را بر سود واحدهای تولیدی در طول زمان بررسی کرده و نتیجه را با دو استراتژی دیگر (استراتژی تصادفی و استراتژی حریصانه) مقایسه می‌کند. در این مطالعه هربار اجراء مکانیزم حراج و تسویه یک بازی تکرار ایستا از نوع همکارانه کامل فرض شده است.

ساختار مقاله به شرح زیر است. بعد از مقدمه، در بخش دوم معرفی اجمالی از مدل‌سازی عامل‌محور از بازار برق ارائه می‌شود. بخش سوم پیشینه پژوهشی در حوزه کاربرد یادگیری تقویتی در مدل‌های عامل‌محور بازار برق ارائه می‌گردد. بخش چهارم یادگیری تقویتی برای مدل‌های تک‌عاملی، فرایند مارکفی و مدل‌های چندعاملی معرفی شده و سپس کارکرد این مدل در قالب تئوری بازی و به صورت یک بازی تکرار ایستا بیان می‌شود. در بخش پنجم، مدل بازار برق ایران و مفروضات مهم آن مختصرا شرح داده می‌شود. در بخش ششم نتایج پیاده‌سازی مدل و شبیه‌سازی استراتژی‌های مختلف بر عملکرد بازار ارائه شده و نتایج مورد بحث و بررسی قرار می‌گیرد. بخش هفتم، شامل جمع‌بندی و نتیجه‌گیری است.

2. مدل‌سازی عامل محور بازار برق

مطالعه در مورد نحوه تصمیم‌گیری و تحلیل عملکرد افراد و بنگاه‌ها در سامانه‌های ترکیبی و تعاملی، از قبیل یک جامعه یا یک بازار، همواره موضوع مورد توجه پژوهشگران علوم اجتماعی و حتی مهندسین بوده است. آنها با فرض عقلائی بودن عملکرد افراد و تبادل صحیح اطلاعات، تعامل اجتماعی را در حالت تعادل، بهینه تشخیص می‌دهند. مدل‌سازی

عامل‌محور (ABM) از این دیدگاه، عملکرد یک سامانه پیچیده اقتصادی را براساس رفتار و عملکرد تک‌تک عناصر تشکیل‌دهنده آن، توصیف و تحلیل می‌کند. این دیدگاه با توجه به محدودیت روش‌های تحلیل سنتی و پیشرفت در زمینه محاسبات، بیش از پیش مورد توجه قرارگرفته است. در مدل عامل‌محور، تصمیمات توسط یک ماهیت خودگردان ومستقل به نام عامل[1] گرفته می‌شود. این تصمیمات براساس مجموعه ادراکات عامل از محیط و دانش درونی عامل گرفته می‌شود. البته دانش و معرفت عامل در طول زمان با توجه به تجربه یا به صورت سعی وخطا تکامل یافته و ارتقاء می‌یابد. عامل به صورت عقلائی تصمیم می‌گیرد و قابلیت یادگیری، تطبیق و بازتولید دارد. تصمیم عقلائی، تصمیمی است که از تمامی اطلاعات در دسترس برای رسیدن به بیشینه سود استفاده می‌کند.

مدل عامل‌محور یک مدل محاسباتی است و از تعامل و تاثیر متقابل عامل‌ها بریکدیگر، عملکرد کل پدیده موردنظر را شبیه‌سازی می‌نماید. تعامل وتاثیر متقابل عامل‌ها در سطح خرد، سبب بروز نتایج در سطح بازار می‌گردد. خصوصیت ویژه این نوع مدل، قابلیت تعامل رقابتی ومکرر عامل‌ها است. استفاده از مدل عامل‌محور مزایای متعدد دارد. از جمله اینکه امکان شبیه‌سازی و بازتولید عملکرد سامانه‌های پیچیده را به صورت محاسباتی فراهم می‌کند. خروجی یک سامانه ترکیبی، لزوما حاصل جمع مولفه‌های تشکیل دهنده نیست، بلکه برآیند تاثیر متقابل آنها بر یکدیگر است که مدل عامل‌محور امکان دستیابی به آن را میسر می‌سازد. مدل عامل‌محور امکان بازنمائی و توصیف ساده و طبیعی از یک سامانه ترکیبی و پیچیده را میسر می‌نماید. نهایتا اینکه، این نوع مدل‌سازی قابلیت انعطاف کافی برای نمایش سطوح مختلف از پیچیدگی اعم از تعداد وتنوع عامل‌ها، نحوه عملکرد عقلائی، عدم‌قطعیت، سیاست یادگیری و تکامل عامل‌ها را دارد.

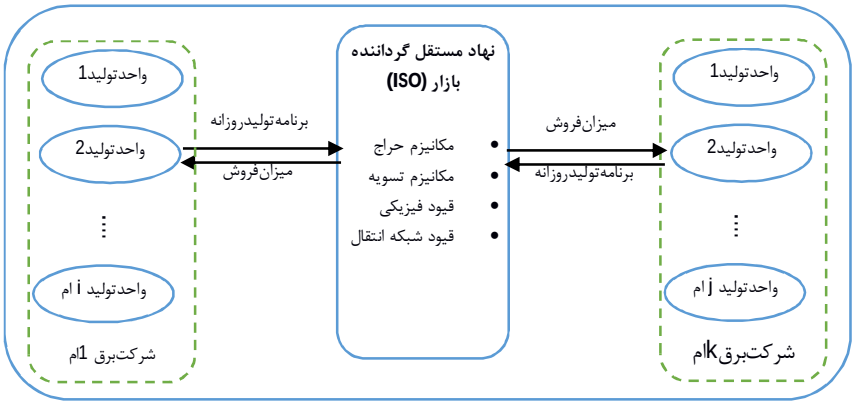
---

1. Agent

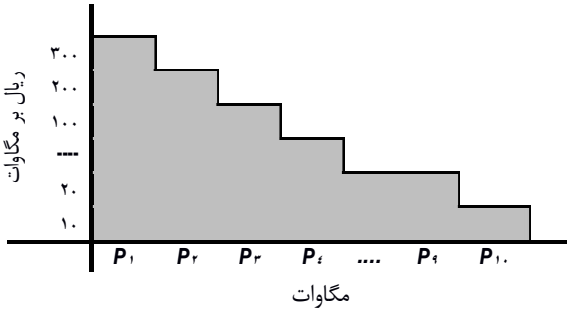
چنانچه اشاره شده مدل‌سازی عامل‌محور یک رویکرد طراحی پائین‌به‌بالا[1] است و در مسائل اقتصادی، که درآن‌ها نتایج بازار از برآیند تاثیر متقابل عوامل خرد شکل می‌یابد، کاربرد فراوان دارد. بازار عمده‌فروشی برق با مدل عامل‌محور به خوبی قابل بازنمائی است. بازیگران اصلی این بازار، واحدهای تولید برق هستند که به صورت مستقل و خودگردان تصمیم می‌گیرند به چه میزان و قیمت نیروی برق به بازار عرضه‌کنند. این واحدها با تبادل شفاف اطلاعات، امکان تعامل با دیگر بازیگران بازار را دارند و با توجه به معیار و قواعد درونی و اطلاعات بیرونی، استراتژی مناسب خود را با هدف بیشینه نمودن سود، انتخاب می‌نمایند. واحدهای تولیدی به لحاظ نوع فناوری ومشخصات فنی به چهار نوع اصلی نیروگاه‌های گازی، بخاری، ترکیبی و آبی‌هسته‌ی تقسیم می‌شوند. انواع عامل‌های دیگر نیز می‌تواند در این مدل وجود داشته باشد، که از جمله شرکت‌های انتقال و توزیع را می‌توان نام برد. در اغلب موارد برای ساده‌سازی مدل، نقش عامل‌های انتقال و توزیع به صورت قیود و قواعد محدودکننده مدل در نظرگرفته می‌شود. ضمنا نهاد بهره‌بردار (ISO) نیز وظیفه کنترل و مدیریت بازار را برعهده دارد. شکل (1) یک نمای از عوامل اصلی تشکیل‌دهنده مدل و جریان تبادل اطلاعات را نمایش می‌دهد. در مدل بازار برق، هرواحد تولیدی (عامل) با ارائه برنامه‌تولید[2] (SF) در مکانیزم حراج شرکت می‌کند. برنامه‌تولید شامل محدوده مشخص از میزان تولید (مگاوات) بازاء قیمت‌های (ریال) مشخص است که برای یک دوره (24 ساعتی) تهیه می‌شود. شکل (2) نمونه یک برنامه‌تولید را نشان می‌دهد. در این نمودار، قیمت پیشنهادی به ازاء میزان تولید به صورت پله‌ای در 10 سطح مشخص شده است. برنامه‌تولید پیشنهادی روزانه توسط هرعامل به نهاد بهره‌بردار ارائه می‌شود. بهره‌بردار با تجمیع برنامه‌تولید پیشنهادی واحدها و باتوجه به پیش‌بینی تقاضای مصرف، در مکانیزم حراج قیمت بازار و فهرست واحدهای برنده و میزان خرید از هریک را مشخص می‌کند. به عبارتی، با مکانیزم حراج، میزان فروش هرواحد که معیار سنجش موفقیت آن

1. Bottom-Up
2. Supply Function (SF)

واحد یا پاداش[1] برنامه‌تولید (SF) است، مشخص می‌شود. معیار پاداش به عنوان بازخورد در تکرارهای بعدی جهت تصحیح یا حفظ استراتژی عامل مورد استفاده قرار می‌گیرد. بنابراین عامل ممکن است استراتژی تصمیم‌گیری خود را با دریافت بازخورد تغییر دهد.



شکل 1: نمائی از مدل عامل‌محور بازار عمده‌فروشی برق: واحدهای تولید برق به صورت عامل مستقل و در قالب گروه شرکت‌های برق می‌باشند و عامل‌های انتقال و توزیع به صورت قیود فیزیکی و فنی لحاظ شده است.



شکل 2: نمونه برنامه‌تولید (SF) پیشنهادی یک واحد به ازای یک ساعت مشخص

1. Reward

بهره‌بردار (ISO) وظیفه اجراء صحیح حداقل دو مکانیزم را دارد: حراج[1] و تسویه‌حساب[2]. از مکانیزم حراج برای تعیین قیمت بازار، لیست برندگان و میزان تولید هرعامل استفاده می‌شود. مکانیزم تسویه‌حساب برای تعیین میزان پرداخت به واحدها مورد استفاده قرارمی‌گیرد. در حال حاضر از دو مکانیزم متفاوت حراج استفاده می‌شود: حداقل هزینه پیشنهادی[3] و حداقل هزینه پرداختی[4]. در اولی انتخاب تولیدکنندگان به‌گونه‌ی است که هزینه کل آنها حداقل گردد و در دومی به‌گونه‌ی انتخاب می‌شوند که هزینه‌های واقعی پرداخت مصرف‌کنندگان حداقل گردد. مکانیزم تسویه حساب نیز به دو روش قابل اجراء است. مکانیزم تسویه بر مبنای پیشنهاد[5] به این صورت است که به تولیدکنندگان منتخب همان مقداری که پیشنهاد داده‌اند، پرداخت می‌شود. درمکانیزم تسویه بر مبنای قیمت تسویه بازار[6] به کلیه تولید کنندگان، قیمت یکسان که همان قیمت تسویه بازار است، پرداخت می‌شود.

چنانچه اشاره شد بکارگیری مدل عامل‌محور، فعالان عرصه بازار را قادر خواهد ساخت که فضای تصمیم‌گیری غیرمتمرکز و نامتجانس را شبیه‌سازی و واکاوی نمایند. عامل‌ها در بستر بازار و بر اساس اهداف، رویه تصمیم‌سازی، و الگوی رفتاری منحصر بفرد خود به کنش‌های متقابل خواهند پرداخت. عامل‌ها قادر خواهند بود که تجربیات گذشته خود را مبنای یادگیری ساخته، رفتار خود را با شرایط محیط تطبیق‌داده، و از فرصت‌های پدید آمده بهره جویند. از این‌رو مدل‌سازی عامل‌محور در زمره روش‌های تطبیقی وپیچیده قرار می‌گیرد. زنجیره این کنش و واکنش عامل‌ها، شبیه‌سازی بازار برق را پیش خواهد راند.

---

1. Auction
2. Settlement
3. Offer Cost Minimization
4. Payment Cost Minimization
5. Pay-As-Bid (PAB)
6. Market Clearance Price (MCP)

موفقیت و یا شکست عامل‌ها مبنای تطبیق راهبرد آن‌ها را شکل داده، و الگوی تکامل رفتار آن‌ها قابلیت تحلیل و بررسی خواهد یافت.

3. پیشینه پژوهش

راشدی و همکاران[1] (2016) یادگیری تقویتی را یک استراتژی پیشنهاد قیمت بهینه برای مدل عامل‌محور بازار عمده‌فروشی برق معرفی نمودند. آنها مدل چندعاملی با مکانیزم تسویه برمبنای قیمت تسویه بازار (MCP) و تعادل تابع‌عرضه[2] (SFE) را استفاده کرده و در قالب یک مسئله نظریه‌بازی با اطلاعات ناکامل و غیرهمکارانه، استراتژی بهینه را با کمک یادگیری تقویتی مشخص نمودند. در این مطالعه از الگوی فرایند تصمیم مارکفی برای مدل تک‌عاملی استفاده شده و یک الگوی تصمیم برای مدل چندعاملی در بازی‌های تصادفی تکرارپذیر ارائه شده است. استراتژی پیشنهاد قیمت برای یک روز بعد بازاء هر واحد تولیدکننده با کمک یادگیری تقویتی در مدل چندعاملی غیرهمکارانه و براساس استاندارد توزیع IEEE-30 پیاده سازی شده است. در این رویکرد یادگیری، برای استراتژی پیشنهاد قیمت، نیازی به تخمین حالت و فضای حالت نیست و هر عامل صرفا برمبنای یادگیری از تجربه اختصاصی خود، پیشنهاد قیمت بعدی را ارائه می دهد.

راشدی و کبریائی[3] (2014) در یک مطالعه دیگر با استفاده از مفهوم تعادل نش، رفتار بازیگران بازار عمده‌فروشی برق را در دو حالت رقابتی و تعاملی مورد مطالعه قراردادند. آنها برای این بازار، مدل بازی برمبنای تعادل تابع‌عرضه (SFE) را معرفی می‌نمایند. در این مدل تعدادی واحد تولیدکننده وجود دارد که هدف آنها بیشینه نمودن سود است. طبق تعریف تابع عرضه، سطح مطلوب تولید هر واحد را بازاء قیمت‌های متفاوت مشخص

1. Rashedi, *et al.* (2016)
2. Supply function Equilibrium (SFE)
3. Rashedi and Kebriaei (2014)

می‌کند. در بازی رقابتی (ناهمکارانه[1])، هر واحد صرفا هدف بیشینه نمودن سود خود را دنبال می‌کند و نقطه تعادل شرایطی است که سود هرواحد درقبال پیشنهاد سایر واحدها، بیشینه باشد (تعادل نش). دربازی همکارانه، با توجه به این واقعیت که افزایش سود یک واحد سبب کاهش سود واحد دیگر می‌شود، عامل‌ها به دنبال سود بهینه همه‌جانبه[2] هستند و برای نیل به آن برخی قواعد منصفانه باید لحاظ شود. قوانین به نوعی است که به جای حداکثر نمودن صرف سود هر عامل، یک تابع هدف تعریف می‌شود که سود مجموعه را نیز تامین نماید. بنابراین، تابع هدف از دو مولفه تشکیل شده است و علاوه بر سود هر عامل، میزان هماهنگی با سود مجموعه را نیز در بردارد. مولفه دوم ممکن است به صورت افزاینده یا کاهنده تاثیر گذارد. این مدل در دو حالت رقابتی و همکارانه با دو فرض تقاضای مصرف ثابت و تقاضای مصرف متناسب با قیمت، تحلیل شده و نتایج شبیه سازی آن مورد بررسی قرار می‌گیرد. این تحقیق به صورت تحلیلی اثبات می‌کند تعادل نش برای بازی رقابتی وجود دارد. در بازی همکارانه، در حالت تقاضای مصرف ثابت، بسته و محدب بودن تابع هدف شرط لازم و کافی برای وجود نقطه تعادل است. اما در حالت تقاضای مصرف متناسب با قیمت برق این شروط برای وجود نقطه تعادل کافی نمی باشند.

رحیمیان و رجبی‌مشهدی[3] (2010) یک رویکرد تطبیقی یادگیری تقویتی برای بازار عمده‌فروشی برق که به صورت عامل‌محور مدل شده‌است، ارائه می‌دهند. در این تحقیق تلفیق دو رویکرد اکتشاف وبهره‌برداری[4] در یادگیری تقویتی بر اساس خصوصیات لحظه‌ی بازار برق به صورت تطبیقی تنظیم می‌شود. برای این منظور از یک سامانه استنتاج فازی برای تصمیم‌گیری وتنظیم پارامتر یادگیری براساس توان تولید هرعامل استفاده می‌شود. مزایای این روش، تطبیق پارامترهای یادگیری تقویتی بر اساس شرایط بازار وتولید هر

1. Non-cooperative
2. Pareto solution
3. Rahimiyan, *et al*. (2010)
4. Exploration and exploitation

واحد، کاهش حجم محاسبات مربوط به یادگیری، تعیین میزان ریسک هر استراتژی برای پیشنهاد قیمت و تدوین قوانین محاوره‌ی فازی برای انتخاب استراتژی است. بخشی از این تحقیق در رساله رحیمیان (1390) با عنوان رویکرد یادگیری تقویتی با خصوصیت فازی به عنوان یک استراتژی پیشنهاد قیمت در بازار عمده فروشی برق معرفی شده است. این رویکرد با روش قیمت گذاری متداول عامل محور مقایسه شده است.

ناظمی و همکاران (1390) میزان رقابت در بازار عمده‌فروشی برق براساس شاخص‌های ساختاری و نیز پتانسیل بروز رفتار غیررقابتی را مورد مطالعه قرارداده است. در این تحقیق مدلی از بازار برق ارائه شده و ضمن تعیین شاخص‌ها، ضریب همبستگی آنها مربوط به سال 1388 محاسبه شده است. سپس نتایج شبیه سازی مدل و تفاضل تعادل بهینه و رفتار راهبری محاسبه شده است. مدل این پژوهش علاوه بر قیود اقتصادی، قیود فنی را نیز شامل می‌باشد. موید کاظمی و شیخ‌الاسلامی (1393) با ارائه یک مدل عامل‌محور بازار سرمایه‌گذاری صنعت برق، استراتژی مبنی بر یادگیری تقویتی و تاثیر آن بر دینامیک سرمایه‌گذاری را بررسی و مطالعه نموده است. این تحقیق اثر راهبردهای تشویقی برای تامین هزینه‌های تولید و هزینه ثابت را بررسی می کند.

کراوس[1] و همکاران (2006) یک مدل عامل‌محور از بازار عمده‌فروشی برق را درنظر گرفته و ضمن استفاده از استراتژی یادگیری تقویتی شرایط رسیدن به تعادل نش را در صورت وجود، مورد بررسی قرار داده است. این تحقیق با استفاده از نتایج شبیه سازی مدل بازار نشان می‌دهد که در صورت وجود یک نقطه تعادل نش، عملکرد بازار بر اساس استراتژی یادگیری تقویتی به سمت آن همگرا است و در صورت وجود بیش از یک نقطه تعادل، بین آنها نوسان می کند. این تحقیق اثبات می کند که این همگرایی تحت پارامترهای مختلف بازار استوار می‌باشد. نتیجه مهم اینکه در صورت اثبات وجود نقاط تعادل مختلف، عملکرد سیستم نوسانی خواهد بود. مشیری و همکاران (1397) نیز از یک مدل عامل محور

1. Krause, *et al*. (2006)

برای شبیه‌سازی بازار برق ایران و ارزیابی سناریوی تغییر مکانیزم تامین سوخت نیروگاه ها و رقابتی کردن آن استفاده کردند و نشان دادند که تامین سوخت رقابتی در افزایش کارآیی نیروگاه ها تاثیر مثبتی خواهد داشت.

4. یادگیری تقویتی

یادگیری تقویتی، یکی از شاخه‌های یادگیری ماشین است که از روانشناسی رفتارگرایی الهام گرفته و بر رفتارهایی تمرکز دارد که عامل باید برای بیشینه‌کردن پاداش انجام‌دهد. طبق تعریف، رفتار عبارت از نگاشتی است که از مجموعه اداراکات عامل به تصمیم‌سازی برای انجام اقدام مناسب منجر می‌شود و رفتار عقلائی، رفتاری است که با استفاده از تمامی اطلاعات دردسترس به دنبال بیشینه شدن کارآمدی می‌گردد. یادگیری تقویتی یک روش یادگیری برای انتخاب رفتار مناسب براساس پاداش و تنبیه است بدون اینکه لازم باشد نحوه انجام عمل را برای عامل مشخص نمائیم (حاج‌رسولیها 1393).

در یادگیری تقویتی، نوع اقدام عامل از قبل مشخص نمی‌شود، بلکه عامل با جستجوی مبتنی بر سعی وخطا رفتاری را یاد می‌گیرد که بیشترین پاداش را به دست آورده و سود کوتاه‌مدت فدای سود بلندمدت شود. در استراتژی جستجو برای رسیدن به پاداش بیشتر همواره دو رویکرد اصلی وجود دارد: رویکرد بهره‌مندانه[1] (حریصانه) و رویکرد اکتشافی[2] (تصادفی). چالش اصلی ایجاد یک تعادل با ترکیب دو رویکرد فوق است و لذا باید بین کاوش موارد جدید و استفاده از دانش پیشین تناسب ایجاد نمود. لذا در فضای جستجو با توزیع تصادفی یکنواخت باید چندین تکرار انجام داد تا امید ریاضی پاداش بلندمدت را حداکثر نمود. یادگیری تقویتی، در زمینه‌های گوناگونی کاربرد دارد. ازجمله می‌توان از نظریه بازی‌ها، نظریه کنترل، تحقیق در عملیات، نظریه اطلاعات، سامانه چندعاملی، هوش ازدحامی، آمار، الگوریتم ژنتیک، بهینه‌سازی بر مبنای شبیه‌سازی نام برد. یادگیری تقویتی

1. Exploiting (Greedy)
2. Exploring (Random)

در اقتصاد و نظریه بازی بیشتر به بررسی تعادل‌های ایجاد شده تحت عقلانیت محدود عامل‌ها می‌پردازد.

اصولا یادگیری به سه نوع تقسیم‌بندی می‌شود: یادگیری با الگو، یادگیری بدون الگو و یادگیری تقویتی. در یادگیری با الگو، داده‌ها با برچسب، به صورت زوج مشترک، در فرایند یادگیری استفاده می‌شود و معمولا هدف یافتن یک الگو ساخت‌یافته بین داده‌ها است. در یادگیری بدون‌الگو، داده‌ها بدون برچسب و معمولا براساس شباهت بایکدیگر تحلیل می‌شوند. در یادگیری تقویتی، داده‌ها براساس تابع هزینه (پاداش یا تنبیه) تحلیل می‌شوند. یادگیری‌تقویتی با یادگیری با الگو دو تفاوت عمده دارد، نخست اینکه در آن زوج مشترک ورودی وخروجی درکار نیست و رفتارهای ناکارامد نیز از بیرون اصلاح نمی‌شوند و دیگر آنکه تمرکز زیادی روی کارامدی تعاملی وجود دارد که نیازمند رسیدن به یک تعادل بین اکتشاف‌های جدید و بهره‌برداری از دانش اندوخته شده دارد. (حاج‌رسولیها 1393)

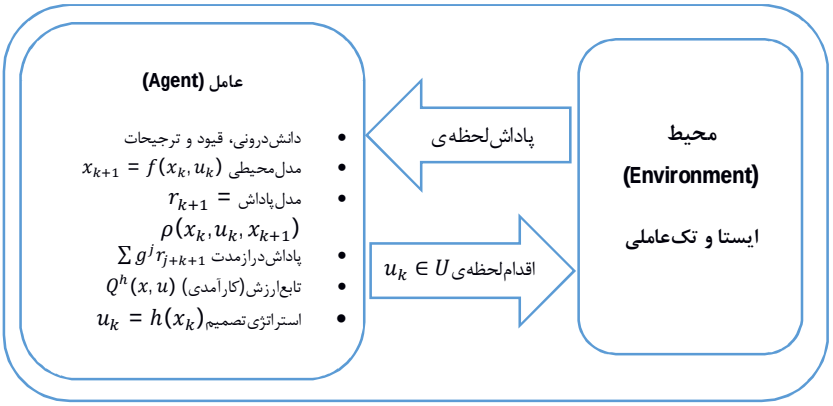
از آنجا که یادگیری تقویتی براساس تعامل با محیط و عامل‌های دیگر شکل می‌گیرد، در ادامه به معرفی یادگیری تقویتی در دو گروه مسائل تک عاملی و چندعاملی اشاره می‌شود.

4-1. یادگیری در مدل تک عاملی

یادگیری در مدل تک‌عاملی، یک فرایند تصمیم مارکفی درنظر گرفته می‌شود. طبق تعریف در (باسینیو و همکاران[1] 2008)، فرایند مارکفی غیراحتمالی، شامل یک چهارتائی $\{X, U, f, \rho\}$ است. X مجموعه فضای حالت و $x_k \in X$ بردار حالت در لحظه k است. U مجموعه متناهی از اقدامات تعریف شده برای عامل است و اقدام عامل در لحظه k به صورت $u_k \in U$ نمایش داده می‌شود. تابع f تابع تبدیلی است که تغییر حالت در اثر اقدام

1.Busoniu, *et al*. (2008)

عامل در لحظه k را به صورت $x_{k+1} = f(x_k, u_k)$ توصیف می‌کند. به دنبال هر اقدام، عامل پاداش اسکالر $r_k \in R$ را دریافت می‌کند که طبق تابع پاداش $r_{k+1} = \rho(x_k, u_k, x_{k+1})$ محاسبه می‌شود. این پاداش صرفا مربوط به حاصل اقدام در لحظه k است که محیط را از حالت k به حالت k+1 منتقل می‌کند و ربطی به پاداش عملکرد درازمدت عامل ندارد. رفتار عامل براساس استراتژی[1] انتخاب اقدام مناسب به صورت $u_k = h(x_k)$ مشخص می‌گردد. این استراتژی چنانچه نسبت به زمان ثابت باشد، ایستا نامیده می‌شود.



شکل 3: نمائی از تعامل مدل تک‌عاملی که به صورت فرایند مارکفی غیراحتمالی توصیف می‌شود.

هدف یک عامل عقلائی، انتخاب اقدام مناسب در جهت بیشینه نمودن پاداش عملکرد در یک بازه درازمدت است. این معیار به صورت امید مجموع پاداش‌ها در بازه نامتناهی طبق رابطه (1) محاسبه می‌شود.

$$R_k = E\left\{\sum_{j=0}^{\infty} \gamma^j r_{k+j+1}\right\} \qquad (1)$$

1. Policy

در رابطه (1) r_k پاداش لحظه‌ای و R_k امید پاداش تجمعی است. ضریب γعدد مثبت و کوچکتر از واحد است که ضریب‌تنزیل[1] نامیده می‌شود و میزان تاثیر پاداش لحظه‌ای در پاداش تجمعی را مشخص کرده و ضمنا سبب کراندار شدن آن می‌گردد. وظیفه عامل، انتخاب یک اقدام مناسب در لحظه kاست به طوریکه سبب بیشینه شدن پاداش تجمعی شود، لذا لازم است هر لحظه تخمینی از محاسبه پاداش تجمعی داشته باشد. برای این منظور تابع ارزش اقدام[2] که به اصطلاح تابع کیو[3] و به اختصار از این پس تابع ارزش نیز نامیده می‌شود، طبق رابطه (2) تعریف می‌گردد.

$$Q^h(x,u) = E\left\{\sum_{j=0}^{\infty}\gamma^j\, r_{k+j+1} \mid x_k = x, u_k = u, h\right\} \qquad (2)$$

تابع ارزش همواره وابسته به استراتژی تصمیم عامل (h) است. شکل 3 نمائی از تعامل مدل تک‌عاملی و مولفه‌های اصلی را نشان می‌دهد. مقدار بهینه تابع ارزش، در بیشینه آن، به صورت $Q^*(x,u) = max_h Q^h(x,u)$تعریف می‌شود. اثبات می‌شود که بیشینه تابع ارزش، طبق معادله بلمن[4] قابل محاسبه است (باسینیو و همکاران[5] 2008). براساس معادله بلمن، مقدار بهینه تابع، همواره معادل مجموع پاداش لحظه‌ای و مقدار بهینه تابع ارزش انتظاری در لحظه‌ی بعدی است. براین اساس، استراتژی بهینه، یک استراتژی غیراحتمالی و حریصانه[6] است که در هر لحظه، اقدامی را انتخاب می‌کند که بیشترین تابع ارزش را مطابق رابطه (3) محقق سازد.

$$h(x) = argmax_h Q^h(x,u) \qquad (3)$$

---

1. Discount Factor
2. Action-Value Function
3. Q-function
4. Bellman Optimality Equation
5. Busoniu, *et al.* (2008)
6. Greedy Policy

بنابراین یک عامل یادگیرنده، لازم است ابتدا تابع ارزش را هرلحظه تخمین زده و سپس براساس استراتژی حریصانه اقدامی را انتخاب کند که تابع ارزش را بیشینه نماید. لذا می‌توان گفت که یادگیری تقویتی، عبارت از تخمین تابع ارزش بهینه و یافتن استراتژی مناسب براساس آن است. تخمین تابع ارزش به دوصورت قابل انجام است: مبتنی بر مدل و بی‌نیاز از مدل. روش‌های برنامه‌ریزی پویا و اکتشافی (هیوریستیک)، یادگیری مبتنی بر مدل فضای حالت هستند. روش بی‌نیاز از مدل که به نام یادگیری‌کیو[1] نیز شناخته می‌شود، بسیار مفید و پرکاربرد است. یادگیری‌کیو یک روش محاسبات تکرارشونده است. در این روش، تخمین مقدار بهینه Q* هر لحظه طبق رابطه (4) براساس پاداش لحظه‌ای بروزرسانی می‌شود.

$$Q_{k+1}(x_k,u_k) = Q_k(x_k,u_k) + \alpha_k[r_{k+1} + g.max_{u'}Q_k(x_{k+1},u') - Q_k(x_k,u_k)]$$
$$Q_{k+1}(x_k,u_k) = (1-\alpha_k)Q_k(x_k,u_k) + \alpha_k[r_{k+1} + g.max_{u'}Q_k(x_{k+1},u')] \quad (4)$$

در رابطه (4) تخمین تابع ارزش، بدون نیاز به تابع تبدیل f و تابع پاداش ρ انجام می‌گیرد، لذا جزء روش‌های عاری از مدل شناخته می‌شود. ضریب یادگیری α_k که یک عدد مثبت و کوچکتر از واحد است، نرخ بروزرسانی تابع ارزش بوده و سرعت همگرائی را تنظیم می‌کند. عبارت داخل کروشه، تفاضل تخمین $Q_k(x_k,u_k)$ بازاء دو لحظه متوالی k و k+1 را مشخص می‌کند که به اصطلاح نرخ تغییرات زمانی[2] نامیده می‌شود. اثبات می‌شود که مقدار Q در رابطه (4) تحت شرایطی به مقدار بهینه آن (Q*) همگرا می‌شود (باسینیو، 2008). یک شرط مهم همگرائی این است که تمام اقدام‌های ممکن برای حالت‌های محتمل (احتمال غیرصفر) لحاظ گردیده و جستجو برای یافتن مقدار بیشینه

1. Q-learning
2. Temporal difference

تغییرات‌زمانی به صورت فراگیر باشد. بنابراین لازم است گاهی اوقات، اقدامی متفاوت از استراتژی حریصانه انتخاب شود که به آن استراتژی تصادفی یا اکتشافی گفته می‌شود. برای مثال، به ازاء هر اقدام، استراتژی حریصانه با احتمال ε و استراتژی تصادفی با احتمال $1-\varepsilon$ اختیارشود. رویکرد دیگر برای تامین شرط همگرائی، استراتژی اکتشافی بلتزمن[1] است که طبق آن در حالت x اقدام u با احتمال رابطه (5) مشخص می‌شود.

$$h(x,u) = \frac{e^{Q(x,u)/\tau}}{\sum_{u'} e^{Q(x,u')/\tau}} \qquad (5)$$

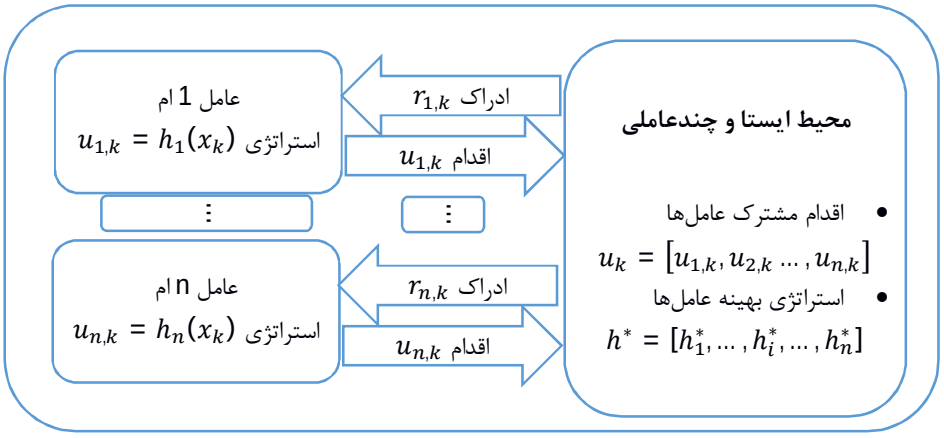
در رابطه (5) پارامتر $\tau \in (0,\infty)$ که اصطلاحا درجه‌حرارت نامیده می‌شود، میزان تصادفی یا حریصانه بودن استراتژی انتخاب را معین می‌کند. بازاء $\tau \dashrightarrow 0$ استراتژی انتخاب کاملا حریصانه و بازاء $\tau \dashrightarrow \infty$ استراتژی انتخاب کاملا تصادفی خواهد بود. استراتژی انتخاب، بازاء مابین این دو مقدارحدی، ترکیبی وابسته به پارامتر τ خواهد بود. هرچه روند یادگیری پیشرفت کند، درجه حرارت بیشتر می‌گردد.

2-4. یادگیری در مدل چند عاملی

حالت تعمیم یافته از فرایند تصمیم مارکفی در مدل چندعاملی، تبدیل به بازی‌احتمالاتی[2] می‌گردد. در یک مدل مشتمل بر n عامل (شکل 4)، تغییر حالت از x_k به x_{k+1} به سبب اقدام مشترک[3] تمام عامل‌ها به‌صورت $u_k = [u_{1,k}, u_{2,k} \dots, u_{n,k}]$ تعریف می‌گردد. دراین تعریف $u_{i,k}$ اقدام عامل i ام در لحظه k ام است. اقدام‌مشترک و تغییرحالت، سبب پاداش $r_{i,k+1} = \rho_i(x_k, u_{i,k})$ برای عامل i ام می‌شود. ضمنا هر عامل، استراتژی مخصوص به خود برای انتخاب اقدام دارد که به صورت $u_{i,k} = h_i(x_k)$ نمایش داده می‌شود. در ادبیات نظریه‌بازی[4]چنانچه تابع پاداش تمام عامل‌ها یکسان باشد،

1. Boltzmann
2. Stochastic game
3. Joint action
4. Game theory

$\rho_1 = \rho_2 = .. = \rho_n$ بازی همکارانه[1] کامل نامیده می‌شود و چنانچه تنها دو عامل وجود داشته باشد و $\rho_1 = -\rho_2$ این بازی رقابتی[2] کامل یا مجموع صفر[3] نامیده می‌شود. بازی مخلوط[4] ترکیب‌های متنوعی است که نه‌همکارانه کامل و نه‌رقابتی کامل باشد. چنانچه ملاحظه می‌شود، فرایند تصمیم مارکفی و یادگیری تقویتی در مدل‌های چندعاملی وارد نظریه بازی‌ها شده و براساس ادبیات نظریه‌بازی قابل بیان وترکیب می‌شود. از این پس برای اختصار و توصیف دقیق از اصطلاحات نظریه بازی استفاده خواهد شد (باسینیو و همکاران، 2008).



شکل 4: نمائی از مدل چندعاملی مشتمل بر n عامل.

3-4. بازی ایستا و تکراری

یک نوع خاص از بازی‌ها، مدلی است که مجموعه فضای حالت تهی است ($X = \emptyset$) و پاداش صرفا وابسته به اقدام هرلحظه عامل دارد. به‌عبارتی پاداش صرفا تابع اقدام مشترک

1. Cooperative
2. Competitive
3. Zero-sum
4. Mixed games

عامل‌ها است $r_{i,k+1} = \rho_i(u_k)$. تهی بودن مجموعه فضای حالت به معنی این است که مدل، حافظه ندارد و تصمیم‌گیری بازاء هراقدام (در محیط ایستا)، تکرار یک بازی مستقل است. البته با این تفاوت که در هر تکرار، از پاداش تکرارهای قبل در ساخت تصمیم بعدی می‌توان استفاده نمود. در این نوع بازی، از آنجا که فضای حالت وجود ندارد، استراتژی هر عامل تنها وابسته به مقدارثابتی است که به بازی تخصیص داده شده است $u_{i,k} = h_i(x = constant)$. عملکرد فرایند یادگیری تقویتی در این نوع بازی، معطوف به تعیین این مقدار ثابت در هرتکرار است، که مستقیما در استراتژی و تبعا در میزان پاداش لحظه‌ای و تجمعی تاثیر دارد.

دربازی ایستا یا تکراری، استراتژی تصادفی و چارچوب احتمالی، تقریبا اجتناب ناپذیر است. این موضوع به ویژه در توصیف ریاضی تعادل نش بسیار تعیین کننده است. فرض کنید در یک مدل چندعاملی، بهترین اقدام عامل i ام (که همزمان با اقدام سایر عامل‌ها است) و بیشترین پاداش انتظاری را به دنبال دارد با h_i^* مشخص شود، رابطه (6) برای هرتکرار بازی به صورت مستقل قابل کاربرد است (لذا اندیس k دیگر مورد استفاده قرار نمی‌گیرد).

$$E\{r_i|h_1,\dots,h_i,\dots,h_n\} \le E\{r_i|h_1,\dots,h_i^*,\dots,h_n\}\forall h_i \qquad (6)$$

تعادل نش[1] یک مجموعه استراتژی مشترک $h^* = [h_1^*,\dots,h_i^*,\dots,h_n^*]$ است که انتخاب هر عامل بهترین انتخاب نسبت به سایر عامل‌ها باشد. تعادل نش، شرایطی را توصیف می‌کند که تمام عامل‌ها پاداش بیشینه ممکن را کسب نموده و تغییر استراتژی به سود هیچکدام نیست. اثبات شده است که هر بازی ایستا حداقل یک و بعضا بیش از یک نقطه تعادل دارد. تعادل نش دراغلب برنامه‌های یادگیری تقویتی به عنوان هدف فرایند یادگیری درنظر گرفته می‌شود. لذا درشبیه‌سازی مدل چندعاملی، که عامل‌ها شروع به

1 .Nash Equilibrium

تعامل با یکدیگر نموده و با توجه به الگوریتم یادگیری از عملکرد قبلی، اقدامی را انتخاب می‌کنند که پاداش بیشتری حاصل کنند، پس از گذشت زمان مشخصی به شرایطی می‌رسند که مجموع پاداش عامل‌ها ثابت شده و تغییر نمی‌کند، این حالت تعادل نش است (باسینیو، 2008).

4-4. یادگیری مدل چندعاملی در حالت همکارانه کامل

در بازی همکارانه کامل، عامل‌ها تابع پاداش یکسان دارند ($\rho_1 = \rho_2 = .. = \rho_n$) و هدف یادگیری، بیشینه‌سازی پاداش تجمعی مشترک است ($R = \sum R_{i,k}$). در این نوع بازی، عامل‌ها به دنبال سود بهینه همه‌جانبه هستند و شرایط به نوعی است که به جای صرفا بیشینه نمودن سود هر عامل، یک تابع هدف کلی تعریف می‌شود که سود مجموعه را تامین نماید. با فرض وجود یک کنترل‌کننده مرکزی، این مسئله نیز به یک مسئله فرایند تصمیم مارکفی در فضای مشترک استراتژی تبدیل می‌شود. در این حالت، یادگیری براساس تخمین تابع ارزش مشترک رابطه (7) و با استراتژی حریصانه، می‌تواند به پاسخ بهینه همگرا باشد. شروط همگرائی قبلا اشاره شده است.

$$Q_{k+1}(x_k, u_k) = Q_k(x_k, u_k) + \alpha[r_{k+1} + g.max_{u'} Q_k(x_{k+1}, u') - Q_k(x_k, u_k)] \quad (7)$$

باتوجه به اینکه عامل‌ها مستقل هستند، هماهنگی آنها علیرغم یادگیری موازی و همزمان براساس یک تابع ارزش مشترک، هنوز یک چالش می‌تواند باشد. در استراتژی حریصانه، عامل باید به گونه‌ای اقدام کند که تابع ارزش طبق رابطه (8) حداکثر شود.

$$h_i^*(x) = arg\max_{u_i} \max_{u_1 \dots u_{i-1} u_{i+1} \dots u_n} Q^*(x, u) \quad (8)$$

چنانچه اشاره شد، همگرائی وابسته به نوع استراتژی انتخاب عامل است. درعمل استراتژی حریصانه با مشکل مواجه شده و اغلب پاسخ نهائی به صورت بهینه محلی و نه ضرورتا بهینه سراسری حاصل می‌شود. لذا در اینجا حل مسئله در سه حالت بدون

هماهنگی، با هماهنگی و هماهنگی غیرمستقیم، طرح می‌گردد. درروبکرد یادگیری بدون هماهنگی، ایده اصلی این است که لزوما تابع ارزش بهینه، یکتا نیست. لذا در یادگیری تیمی، هر عامل براساس رابطه (7) تخمین جداگانه اختصاصی را داشته و براساس رابطه (8) استراتژی جداگانه را با هدف بیشینه نمودن پاداش، تعیین می‌نماید. الگوریتم یادگیری توزیعی[1](لوئر و مارتین[2] 2000) در حل مسائل همکارانه کامل در حالت غیراحتمالی به‌کاربرده می‌شود. در این الگوریتم، هر عامل استراتژی اختصاصی h_i و تابع ارزش اختصاصی Q_iرا دارد که صرفا وابسته به اقدام همان عامل u_i است. ضمنا تابع ارزش اختصاصی، تنها زمانی بروزرسانی می‌شود که روند آن مطابق رابطه (9) افزایشی باشد.

$$Q_{i,k+1}(x_k, u_{i,k}) = max\{Q_{i,k}(x_k, u_{i,k}), r_{k+1} + g\max_{u_i} Q_{i,k}(x_{k+1}, u_i)\} \quad (9)$$

این شرط تضمین‌کننده آن است که تابع ارزش اختصاصی براساس رابطه (10) همواره حداکثر تابع ارزش مشترک را تقریب می‌کند.

$$Q_{i,k}(x, u_i) = \max_{u_1 \dots u_{i-1} u_{i+1} \dots u_n} Q_k(x, u) \quad \forall k, u = [u_1, \dots, u_n]^T \quad (10)$$

بنابراین مطابق رابطه (11) در یادگیری توزیع شده، استراتژی اختصاصی هر عامل فقط و فقط زمانی تغییر می‌کند که مستقیما سبب ارتقاء تابع ارزش گردد.

$$h_{i,k+1}(x_k) = \begin{cases} u_{i,k} & \max_{u_i} Q_{i,k+1}(x_k, u_i) > \max_{u_i} Q_{i,k}(x_k, u_i) \\ h_{i,k}(x_k) & otherwise \end{cases} \quad (11)$$

به عبارتی، رویه رابطه (11) تضمین می‌کند که استراتژی اشتراکی $h^* = [h_1^*, \dots, h_i^*, \dots, h_n^*]$ همواره براساس تابع ارزش سراسری Q* بهینه خواهد بود. به

1. Distributed Q-Learning
2. Lauer and Martin (2000)

سادگی قابل اثبات است که به شرط مثبت بودن مقدار پاداش و تخصیص مقدار اولیه صفر به تابع ارزش اختصاصی هرعامل ($\forall i,k\ \ Q_{i,0}=0, r_{i,k}>0$) استراتژی اختصاصی عامل‌ها به سمت تابع ارزش مشترک بهینه همگرا خواهد شد (باسینیو و همکاران، 2008). در مدل بازار برق ایران که در قسمت بعد شرح داده شده است، شرایط فوق وجود دارد و استراتژی یادگیری توزیع شده به کار برده می‌شود.

5. مدل بازار برق ایران

هدف این مقاله بررسی تاثیر یادگیری تقویتی توزیع شده در عملکرد عامل‌های بازار برق ایران است. مدل نرم‌افزاری بازار برق ایران (فلاحی، 1391) که یک مدل چندعاملی است در پژوهشگاه نیرو پیاده‌سازی شده و سال‌هاست به صورت نرم‌افزار اصلی شبیه‌سازی و هدایت بازار مورد استفاده قرارگرفته است. مشخصه‌های این نرم‌افزار براساس داده‌های واقعی تنظیم شده و نتایج شبیه‌سازی بسیار نزدیک به عملکرد واقعی بازار برق ایران است (فلاحی،1392). چنانچه در بخش‌های قبل توضیح داده شد، قیمت‌گذاری در بازار عمده‌فروشی برق، بازاء هرساعت درشبانه‌روز، توسط نهاد بهره‌بردار (ISO) بازار انجام می‌شود. با توجه به مفروضات این مدل، فرایند تعیین قیمت را می‌توان یک بازی ایستا فرض نمود که هرساعت تکرار می‌شود. در این بازی هرعامل قیمت پیشنهادی خود را به طور مستقل اعلام نموده و بهره‌بردار با توجه به تقاضای بار مصرفی، بهترین پیشنهادها را انتخاب می‌نماید. لازم به توضیح است که فرایند یادگیری سبب تغییر در استراتژی تصمیم سازی عامل می‌گردد، لکن روی مکانیزم بازار (محیط) که مستقل از عامل است، تاثیری از لحاظ پویایی ندارد. لذا ایستا و تکراری بودن صرفا مربوط به محیط و مکانیزم حراج و تسویه است. یادگیری سبب پویائی رفتار عامل می شود، اما مکانیزم‌های اجرائی بازار را تغییر نمی‌دهد.

دراکثر سیستم‌های بهم پیوسته قدرت، گرچه بیشتر توان مورد نیاز توسط واحدهای حرارتی تامین می‌شود، طیف متنوعی از واحدها با فناوری و ظرفیت‌های مختلف در شبکه

موجود است. در مدل مورد مطالعه ما واحدهای از نوع بخاری، گازی، سیکل ترکیبی، آبی، هسته و دیزلی وجود دارد، که در این تحقیق تنها چهار نوع اصلی مورد بررسی قرارگرفته است. استراتژی‌های عملیاتی مختلفی برای انتخاب ترکیب بهینه از واحدها در جهت برآوردن تقاضا وجود دارد، که درساعات مختلف روز نیز متغیر است. البته اغلب استراتژی مبتنی بر معیارهای اقتصادی مقدم بر معیارهای دیگر است. به بیان دیگر، معیار مهمی که در بهره‌برداری سیستمهای قدرت وجود دارد، برآورده شدن تقاضای مصرف با حداقل هزینه و با بهره‌گیری از ترکیب بهینه نیروگاه‌های مختلف می‌باشد. به فرآیند تعیین زمان‌بندی و برنامه‌ریزی ورود وخروج واحدهای تولیدی اصطلاحا مشارکت واحدها[1] (UC) اطلاق می‌شود. هدف اصلی UC حداقل‌کردن کل هزینه عملیاتی تولید است در حالی که همه محدودیت‌های متناظر با این مسئله نیز برآورده شوند. محدودیت‌ها به دو دسته تقسیم می‌شوند: (1) محدودیت‌های تولیدکنندگان، نظیر نرخ شیب وحداقل زمان روشن و خاموش بودن واحد و ... (2) محدودیت‌های سیستمی، نظیر میزان انرژی و ذخیره مورد نیاز سیستم، محدودیت شبکه انتقال و .... در تابع هدف نیز هزینه‌های مرتبط با تولید انرژی، شیب، هزینه‌های راه‌اندازی و خاموش‌کردن واحد به همراه تاثیر این تصمیمات روی درآمد یا هزینه‌های مشتریان نیز باید درنظر گرفته شوند. با توجه به تابع هدف ومحدودیت‌های مطرح شده، UC یک مساله برنامه‌ریزی عدد صحیح مختلط غیرخطی با بزرگ مقیاس است (فلاحی، 1392).

نهاد بهره‌بردار (ISO) به طور کلی دو وظیفه مهم برای اجرای بازار بر عهده دارد: مکانیزم حراج و مکانیزم تسویه حساب. از مکانیزم حراج برای انتخاب پیشنهادات و همچنین تعیین میزان انرژی و خدمات جانبی استفاده می‌شود. سپس مکانیزم تسویه حساب برای تعیین میزان پرداخت به تولیدکنندگان منتخب مورداستفاده قرارمی‌گیرد. درحال حاضر از دو مکانیزم حراج متداول در بازارهای عمده فروشی استفاده می‌شود: (1) مکانیزم

1. Unit Commitment (UC)

کمینه هزینه پیشنهادی[1] که در این روش انتخاب تولیدکنندگان به گونه‌ی است که هزینه کل آنها حداقل گردد. (2) مکانیزم کمینه هزینه پرداخت[2] که در این روش تولیدکنندگان به گونه‌ی انتخاب می‌شوند که هزینه‌های واقعی پرداخت مصرف‌کنندگان حداقل گردد. در مرحله دوم، بازارها با دو روش تسویه می‌شوند: (1) روش پرداخت بر مبنای پیشنهاد[3](PAB) که در این روش به تولیدکنندگان منتخب همان مقداری که خودشان پیشنهاد داده‌اند پرداخت می‌شود. (2) روش پرداخت بر مبنای قیمت تسویه بازار[4](MCP) که در این روش به تولیدکنندگان منتخب، قیمت یکسانی که همان قیمت تسویه بازار است پرداخت می‌شود. مسئله برنامه‌ریزی ورود و خروج واحدها (UC) و توزیع اقتصادی‌بار[5] (ED) جزء مسائل برنامه‌ریزی عدد صحیح هستند که هدفشان حداقل‌کردن هزینه تولید با توجه به تقاضای مصرف، ظرفیت‌واحد، احتیاجات ذخیره و سایر محدودیتهای واحدهاست. این مسائل عموماً در رده مسائل پیچیده NP-hard قرار می‌گیرند. اما چون جز مسائل جدایی‌پذیر هستند به شکل کارآمدی با استفاده از روش‌های تجزیه وبرنامه‌ریزی عددی قابل حل هستند و درحال حاضر بسته‌های نرم‌افزاری پیشرفته‌ای برای حل این دسته از مسائل در دسترس است (فلاحی، 1392). توضیحات بیشتر در مورد تعریف محدودیت ها و روش های عددی مربوط به حل مسئله، خارج از موضوع اصلی این مقاله است لذا در ادامه تنها به معرفی عامل‌ها و فرایند یادگیری جهت ارائه پیشنهاد قیمت آنها می‌پردازیم.

درمدل بازار برق ایران، مجموعا 550 واحد تولیدکننده داریم که بازار حضور دارند. مشخصات فنی واحدهای تولیدی در پنج نوع فناوری شامل نیروگاه‌های بخاری، گازی، سیکل ترکیبی، آبی و هسته‌ی با جزئیاتی ازقبیل حداقل ظرفیت ($P_{i,min}$) و حداکثر ظرفیت ($P_{i,max}$) تولید، متوسط هزینه تولید در ساعت (AVC_i)، برنامه تعمیرات سالانه و سایر

1. Offer Cost Minimization
2. Payment Cost Minimization
3. Pay-As-Offer or Pay-As-Bid
4. Pay-at-MCP
5. Economic Despatching

موارد در مدل تعریف شده‌اند. هرعامل یا واحدتولید کننده، حداکثر 10 گزینه برای پیشنهاد قیمت دارد. براساس قوانین، این 10 گزینه براساس ضریبی از متوسط هزینه تولید هر واحد طبق رابطه (12) محاسبه می‌شود. این ضریب بزرگتر از یک و کوچکتر از دو می‌باشد.

$$h_i(j) = \left(1 + \frac{2 \times j}{10}\right) AVC_i, j = 0,1 \dots 8 \qquad (12)$$

هرعامل حداقل سه استراتژی برای قیمت گذاری دارد: تصادفی، حریصانه و یادگیرنده. در استراتژی تصادفی، حرکت اکتشافی است و قیمت پیشنهادی عامل i ام بازاء تکرار k+1 ام بازی به صورت تصادفی (13) مشخص می‌شود.

$$u_{i,k+1} = h_i[uniform.random(0,8)] \qquad (13)$$

در استراتژی غیرتصادفی، حرکت جهت‌دار است، لذا ابتداء لازم است تابع ارزش توزیعی، به صورت اختصاصی بازاء هرعامل، طبق رابطه (14) تخمین زده شود.

$$Q_{i,k+1}(j) = Q_{i,k}(j) + a[r_{i,k+1} - Q_{i,k}(j)] \ \forall j \qquad (14)$$

در این رابطه، تابع ارزش اختصاصی $Q_{i,k}(j) = Q_{i,k}[u_{i,k} = h_i(j)]$ و پاداش اختصاصی $r_{i,k+1} = \rho_i(u_k)$ برای عامل iام حاصل از قیمت‌گذاری در لحظه kام است. پارامتر آلفا، که عددی مثبت وکمتر از واحد است، ضریب یادگیری است. چنانچه قبلا اشاره شد، ضریب یادگیری نرخ همگرائی تخمین تابع ارزش به مقدار بهینه را مشخص می‌کند. استراتژی حریصانه، قیمتی را انتخاب می‌کند که بیشینه تابع ارزش تخمینی را برای تکرار بعدی سبب شود. لذا قیمت پیشنهادی عامل iام در تکرار k+1ام از رابطه (15)

محاسبه می‌شود. یاداور می‌شود در این نوع بازی، فضای حالت تهی است و مسئله از نوع بازی تکرار است.

$$u_{i,k+1} = h_i\left[\underset{j}{\text{argmax}}\, Q_{i,k+1}(j)\right] \qquad (15)$$

در استراتژی یادگیرنده، ترکیبی از استراتژی تصادفی و حریصانه مطابق رابطه (16) برای قیمت گذاری انجام می‌شود. این ترکیب براساس معادله بلتزمن براساس معیار درجه حرارت τ تنظیم می‌شود.

$$u_{i,k+1} = \begin{cases} h_i\left[\underset{j}{\text{argmax}}\, Q_{i,k+1}(j)\right] & if\ uniform.random(0,1) \le \tau_k \\ h_i[uniform.random(0,8)] & otherwise \end{cases} \qquad (16)$$

در این تحقیق معیار درجه حرارت به صورت تجربی براساس رابطه (17) تعیین می‌شود.

$$\tau_k = min\left(\theta, \frac{12k + 500}{12k + 5200}\right) \qquad (17)$$

در رابطه فوق k شمارنده تکرار بازی و پارامتر θ =0/93 سقف احتمال استراتژی تصادفی را مشخص می‌کند (فلاحی، 1392). در پایان این بخش و قبل از بررسی نتایج شبیه‌سازی، مجددا تاکید می‌شود که مدل مورد استفاده در این پژوهش برگرفته از مدل ارائه شده توسط پژوهشگاه نیرو است و این مقاله اثر استراتژی‌های مختلف تصمیم‌سازی را مطالعه می‌کند. در این مدل اثر منافع متقابل مصرف کننده و تولید کننده، در قالب برخی قیود[1] لحاظ شده است که از جمله رابطه (12) که سقف قیمت پیشنهادی عامل‌ها را مشخص می‌کند، می‌توان نام برد. ضمنا عدم امکان تبانی بین تولیدکنندگان و برونزا بودن

1. Constraints

مصرف مورد تاکید می‌باشد. با لحاظ این مفروضات و قیود، این مدل به شکل بازی همکارانه توصیف می‌شود، که عامل‌ها در تابع پاداش مشترک هستند، اما در زمینه سهم بازار رقابت دارند. دربازی همکارانه، عامل‌ها به دنبال سود بهینه همه‌جانبه هستند و برای نیل به آن برخی قواعد باید لحاظ شود. قوانین به نوعی است که به جای صرفا بیشینه نمودن سود هر عامل، یک تابع هدف تعریف می‌شود که سود مجموعه را نیز تامین نماید.

6. نتایج شبیه‌سازی مدل

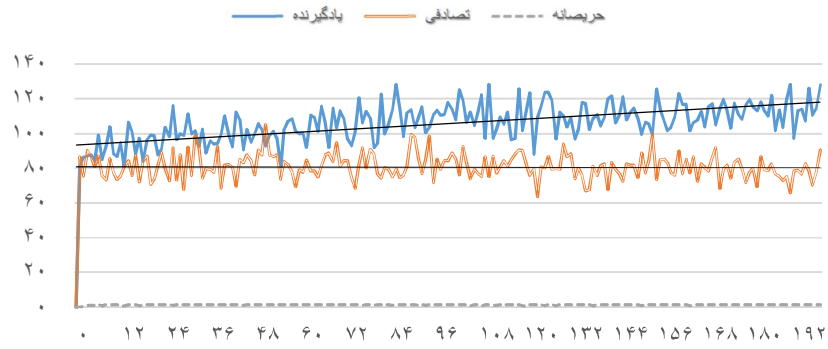
در این بخش نتایج شبیه‌سازی که برای مطالعه و بررسی تاثیر استراتژی یادگیرنده در عملکرد واحدها و سود عملیاتی آنها انجام شده است، ارائه می‌شود. شبیه‌سازی مدل بازار برق ایران، در حالت تسویه بر اساس قیمت پیشنهادی (PAB) و نحوه مشارکت واحدها (UC) در طول 200 تکرار با استراتژی‌های مختلفی انجام‌شده و نتایج ثبت شده است. با توجه به تفاوت ماهیت فنی و اقتصادی واحدها، ارزیابی‌ها به تفکیک فناوری نیروگاه‌ها انجام گرفته‌است. معیارارزیابی عملکرد استراتژی‌ها، براساس پاداش یا سود فروش واحدها بازاء هر مگاوات ثبت شده است.

6-1. تاثیر استراتژی یادگیری در سود واحدهای تولیدی

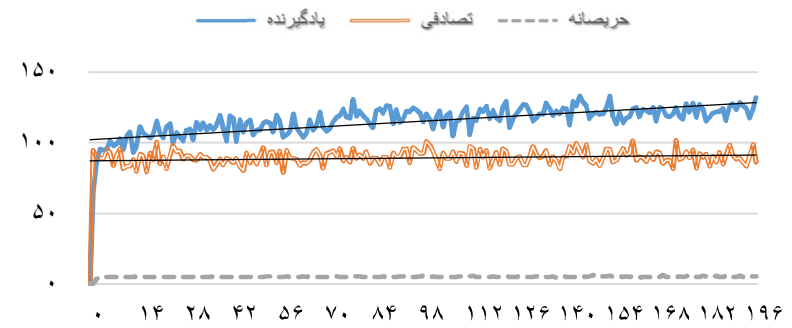
شکل‌های (5) تا (8) میانگین پاداش یا سود بازاء هرمگاوات برای واحدهای تولیدی چهارفناوری اصلی، به ترتیب نیروگاه‌های بخاری، گازی، سیکل‌ترکیبی و آبی را نمایش می‌دهند. نتایج مربوط به 200 تکرار فرایند قیمت‌گذاری است که بر اساس سه استراتژی تصادفی، حریصانه و یادگیرنده عامل‌ها تصمیم‌گیری نموده و قیمت پیشنهادی را ارائه می‌دهند. از بررسی این چهار نمودار مشاهده می‌شود که در تمام واحدهای تولیدی، اعم از بخاری، گازی، ترکیبی و آبی، سود حاصل از استراتژی صرفا حریصانه بسیار کمتر از دو استراتژی دیگر است. از لحاظ نظری اشاره شد که این روش دریک نقطه کمینه محلی متوقف شده وبه پاسخ بهینه تقریب نمی‌کند. مقایسه سود واحدهای تولیدی نشان می‌دهد،

سود حاصل از استراتژی صرفا تصادفی یک روند تقریبا ثابت دارد در حالیکه سود حاصل از استراتژی یادگیرنده یک روند افزایشی یکنواخت را دنبال می‌کند. شکل (9) نموداری از شیب تقریب‌خطی تغییرات سود را نمایش می‌دهد. چنانچه مشخص است، افزایش سود ناشی از استراتژی یادگیرنده در نیروگاه‌های حرارتی (بخاری، گازی و ترکیبی) بیش از ده درصد است که به صورت قابل توجهی بیشتر از استراتژی صرفا تصادفی است. در نیروگاه‌های آبی افزایش سود با استراتژی یادگیرنده بیش از 2درصد است که در مقایسه با استراتژی تصادفی بیشتر است.

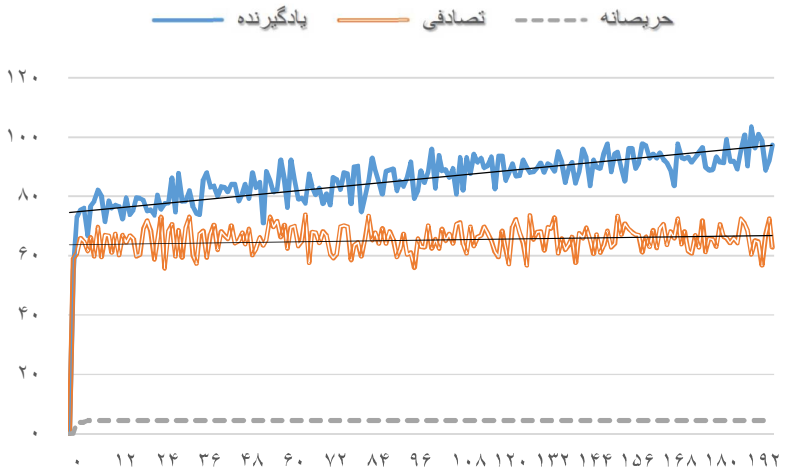
اساسا سهم تولید نیروگاه‌های آبی در بازار ایران نسبت به نیروگاه‌های حرارتی کمتر است. شکل(10) سهم بازار فناوری‌های مختلف را نمایش می‌دهد. طبق نتایج شبیه‌سازی، سهم نیروگاه‌های بخاری، گازی و ترکیبی هریک حدود 30 درصد است و مجموع سهم نیروگاه‌های آبی وهسته‌ای کمتر از 10 درصد است. شکل (11) نتایج تحلیل آماری مقایسه سود ناشی از استراتژی یادگیرنده با دو استراتژی دیگر را نشان می‌دهد. براساس این تحلیل توزیع پراکندگی پاداش واحدهای گازی بازاء استراتژی یادگیرنده به صورت معناداری متمایز و بیشتر از استراتژی حریصانه، تصادفی است.



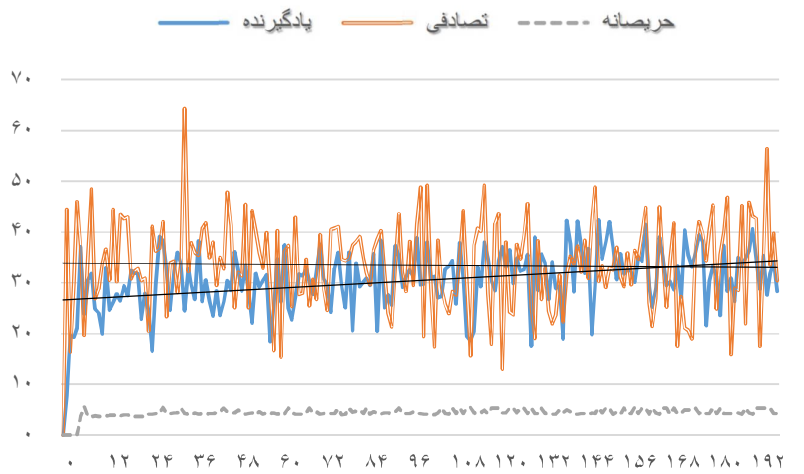
شکل 5: پاداش (سود) بازاء هرمگاوات در واحدهای بخار براساس استراتژی حریصانه، تصادفی و یادگیرنده در 200 تکرار بازی



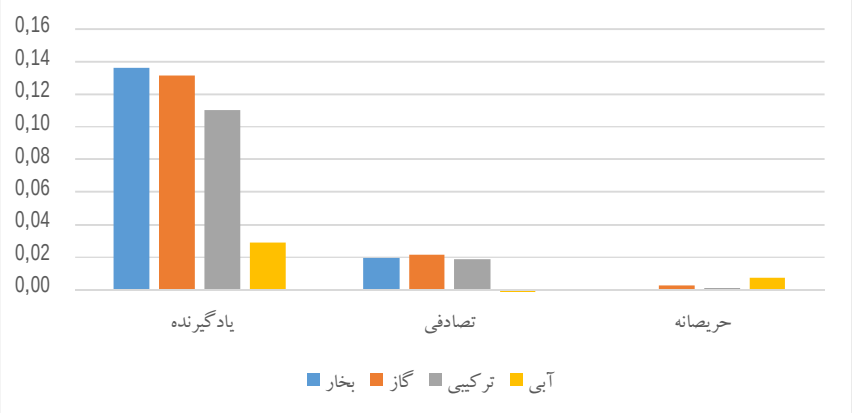
شکل 6: پاداش (سود) بازاء هرمگاوات در واحدهای گازی براساس استراتژی حریصانه، تصادفی و یادگیرنده در 200 تکرار بازی



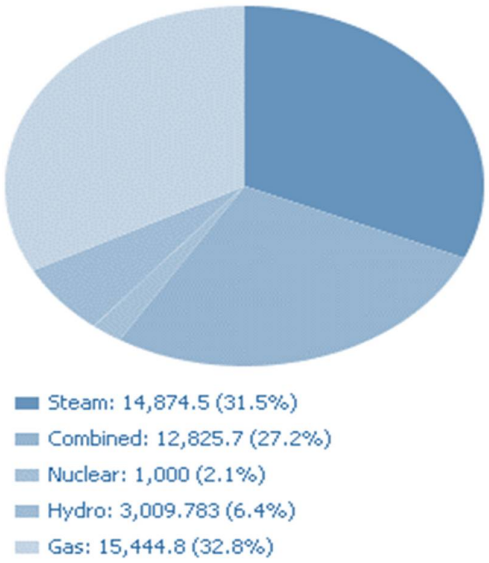
شکل 7: پاداش (سود) بازاء هرمگاوات واحدهای سیکل‌ترکیبی براساس استراتژی حریصانه، تصادفی و یادگیرنده در 200 تکرار



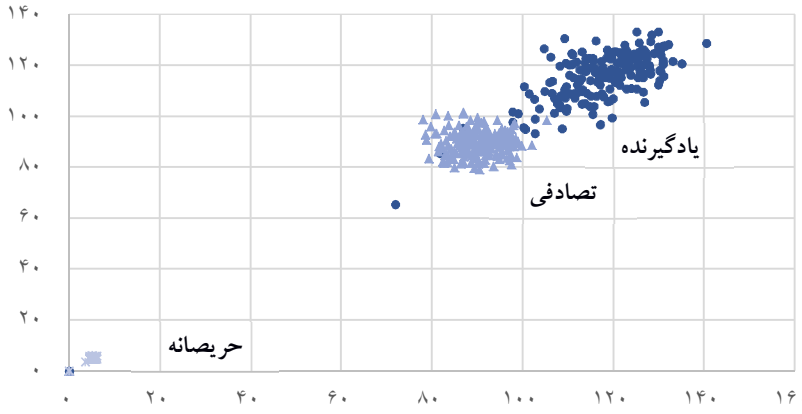
شکل 8: پاداش (سود) بازاء هرمگاوات واحدهای آبی براساس استراتژی حریصانه، تصادفی و یادگیرنده در 200 تکرار بازی



شکل 9: شیب تغییرات خطی پاداش (سود) واحدها براساس استراتژی حریصانه، تصادفی و یادگیرنده در طول 200 تکرار بازی



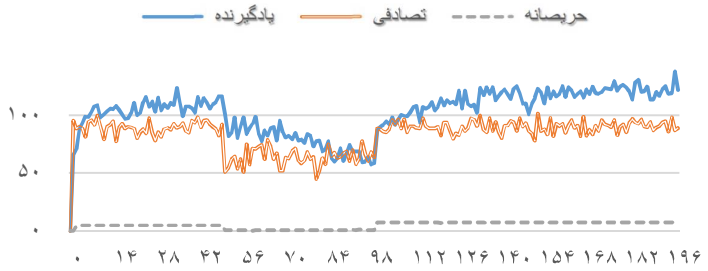
شکل 10: سهم بازار فناوری‌های مختلف در شبیه‌سازی بازار برق ایران



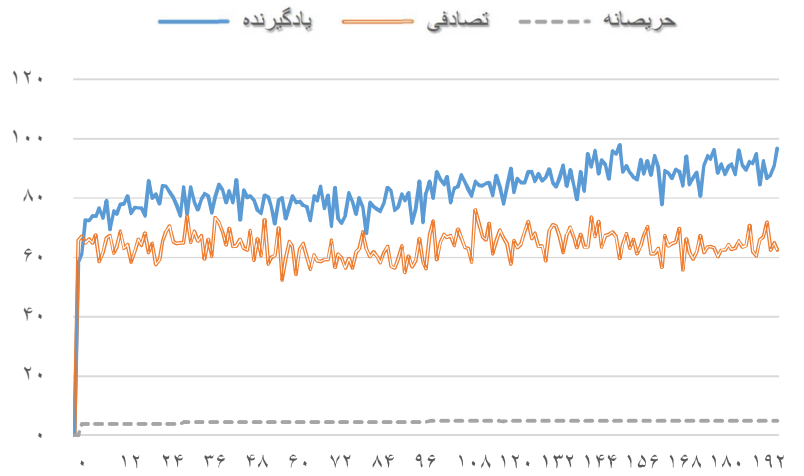
شکل 11: پراکندگی پاداش (سود) براساس استراتژی حریصانه، تصادفی و یادگیرنده در 200 بازی

2-6. تاثیر شوک تغییر تقاضای بار مصرفی

آزمون دوم، با هدف بررسی عملکرد عامل‌ها در دوره گذرا و روند رسیدن به وضع پایدارطراحی گردید. در واقع عملکرد بازار تا رسیدن به نقطه تعادل در این آزمون مورد بررسی قرار می‌گیرد. این آزمون نیز شامل 200 بار تکرار بازی قیمت‌گذاری است، با این تفاوت که در تکرار پنجاه‌ام تقاضای بارمصرفی ناگهان به سطح 80٪ کاهش داده می‌شود و مجددا در تکرار صدم به میزان بار کامل بازگردانده می‌شود. این تغییرات بازاء سه استراتژی مورد نظر، جداگانه مورد بررسی قرارگرفته و نتایج ثبت شده‌است. تبعا تغییر بار، تاثیر مستقیم در پیشنهاد قیمت‌ها خواهد داشت. در این آزمون واکنش واحدها و سود آنها به تفکیک فناوری مورد بررسی قرارگرفته است. شکل (12) روند تغییر سود واحدهای گازی را بازاء سه استراتژی حریصانه، تصادفی و یادگیرنده نشان می‌دهد. مشخص است استراتژی یادگیرنده چه در حالت بار کامل و چه در حالت کاهش بار ناگهانی، سود بیشتری برای واحدها تامین می‌نماید. البته واحدهای گازی به سبب ماهیت فناوری، اینرسی کمتری نسبت به تغییرات دارند. شکل (13) واکنش نیروگاه‌های سیکل‌ترکیبی را به این تغییرات نشان می‌دهد. مشخص است شوک بار، تاثیر محسوسی بر عملکرد وسود این نوع واحدها نداشته و البته طبق انتظار اولیه استراتژی یادگیرنده همواره قالب است. این موضوع درمورد نیروگاه‌های آبی نیز مشابه است.



شکل12: پاداش (سود) واحدهای گازی با وجود شوک تغییر تقاضا به ازاء استراتژی حریصانه، تصادفی و یادگیرنده

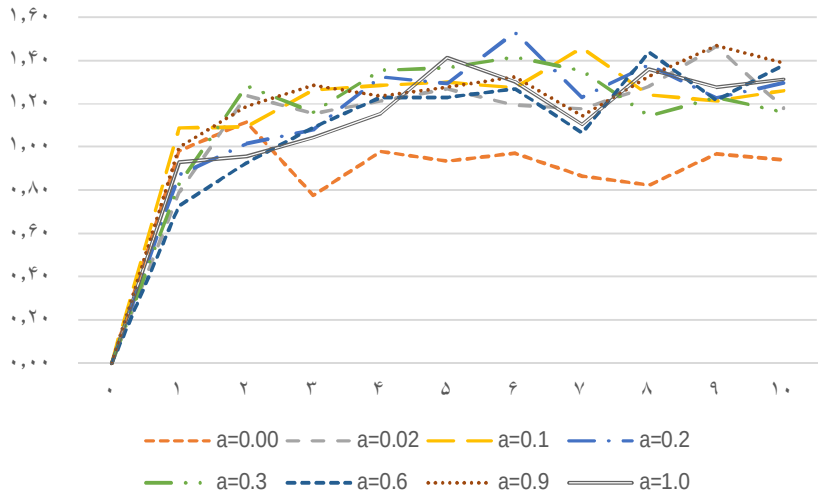


شکل13: پاداش (سود) واحدهای سیکل‌ترکیبی با وجود شوک تغییر تقاضا به ازاء استراتژی حریصانه، تصادفی و یادگیرنده

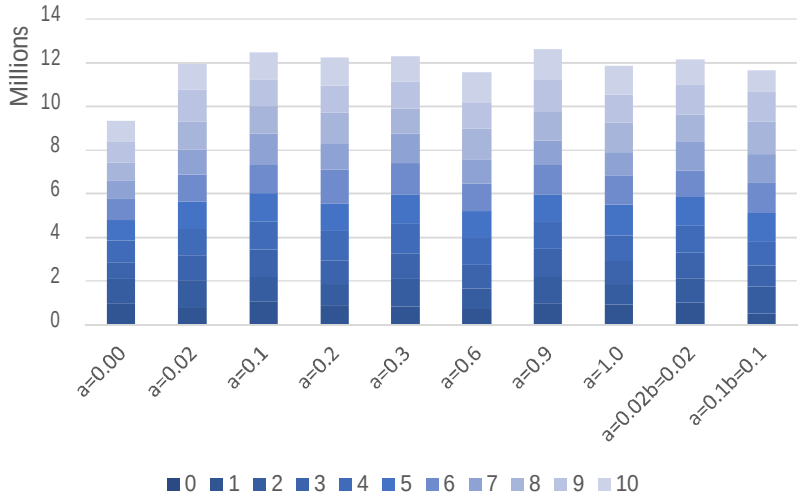
3-6. تاثیر ضریب یادگیری

ضریب یادگیری، چنانچه شرح داده شد، در سرعت همگرائی و رسیدن به حالت تعادل در قبال تغییرات در سیستم، تاثیر دارد. این موضوع در شکل (14) مورد مطالعه قرارگرفته است. این شکل روند تغییر سود واحدهای بخار را در 10 تکرار اول به تفکیک مقادیر مختلف ضریب یادگیری نشان می‌دهد. چنانچه مشاهده می‌شود تفاوت معناداری بین نتایج با ضریب یادگیری صفر و غیرصفر دیده می‌شود، اما تفاوتی معنادار بین اثر مقادیر غیرصفر وجود ندارد. شکل (15) مجموع سود ثبت شده واحدهای بخار را در 10 تکرار اول به تفکیک ضرایب یادگیری نشان می‌دهد. در شکل اخیر شیب استراتژی تصادفی نیز به منظور مقایسه و درک تاثیر ضریب یادگیری، اضافه شده است. ضریب یادگیری در میزان سود یا پاداش اثری ندارد و تنها می‌تواند در سرعت همگرائی موثر باشد. جمع‌بندی این است که ضریب یادگیری غیرصفر که سبب تاثیرپذیری فرایند پیشنهاد قیمت از مقدار

پاداش در تکرارهای قبلی است، تاثیر به سزائی در افزایش سود واحدها دارد اما اندازه آن ضریب، تاثیر معناداری در میزان سود ندارد.



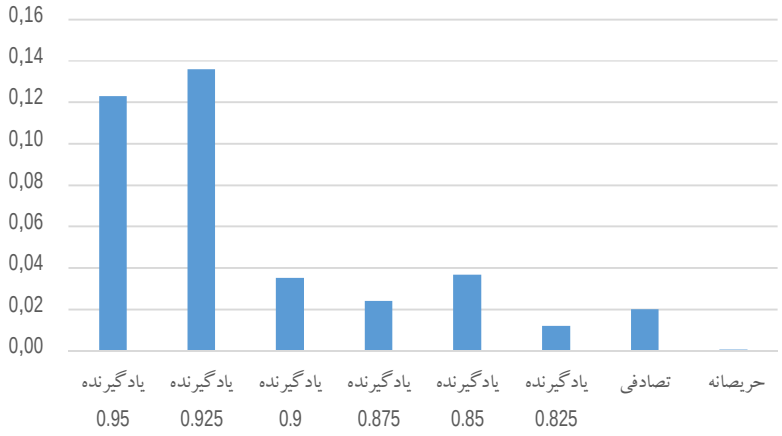
شکل14: تغییرات پاداش در 10 تکرار اول بازی به ازاء ضرایب یادگیری متفاوت (واحدهای بخار)



شکل15: مجموع پاداش واحدهای بخار در 10 تکرار اول بازاء مقادیر مختلف ضریب یادگیری

4-6. تاثیر پارامتر درجه حرارت

چنانچه در بخش‌های نظری یادگیری شرح داده شد، معادله بهینه‌سازی بلتزمن یک معیار نمادین به نام درجه‌حرارت معرفی می‌کند که میزان تلفیق دو استراتژی تصادفی و حریصانه را تعیین می‌کند. این معیار در مدل ما به صورت روابط (16) و (17) تعریف شده است. نام درجه‌حرارت از الگوریتم تبرید شبیه‌سازی[1] شده گرفته است که به تدریج که تعداد تکرار افزوده می‌شود از میزان استراتژی تصادفی کاسته شده و بر میزان استراتژی حریصانه افزوده می‌شود. شکل (16) اثر ناشی از ترکیب‌های مختلف از دو استراتژی تصادفی و حریصانه را در قالب شیب خطی سود عامل‌ها نمایش می‌دهد. این نمودار مشخص می‌کند که با دستکاری تغییرات درجه‌حرارت در طول مدت تکرار بازی، همگرائی تابع ارزش در استراتژی یادگیرنده مختل شده در نتیجه بر قیمت‌گذاری عامل‌ها و شیب افزایشی سود آنها به شدت تاثیر منفی می‌گذارد.



شکل 16: شیب تغییرات خطی پاداش واحدها براساس استراتژی یادگیرنده با درجه حرارت (θ) مختلف و مقایسه آن با استراتژی حریصانه، تصادفی

1. Simulated Annealing (SA)

7. جمع‌بندی و نتیجه‌گیری

پیشرفت‌های فناوری دهه‌های اخیر، بازارهای برق را از حالت تمرکز به سوی رقابتی شدن هدایت کرده است. هرچند رقابتی شدن بازارها موجب کارآیی بیشتر و در نتیجه افزایش رفاه جامعه می‌شود، اما با توجه به ویژگی‌ها و پیچیدگی‌های خاص بازار برق، چالش‌های نوینی نیز به همراه دارد. بازارهای برق سنتی معمولا توسط تولیدکنندگان انحصاری و تحت نظارت دولت تنظیم می‌شوند، اما میزان تولید و قیمت در بازارهای رقابتی از رقابت و تعامل تولیدکنندگان در یک بازار ایجاد شده توسط نهاد تنظیم بازار تعیین می‌شوند. باتوجه به برقراری تعادل لحظه‌ای در بازار برق و عدم امکان ذخیره‌سازی برق در مقیاس وسیع، بازارهای برق به صورت یک روز یا یک ساعت قبل تنظیم می‌شوند. این بازار شامل تعداد زیادی تولیدکننده ناهمگن با فناوری‌ها و استراتژی‌های متفاوت و نهاد مستقل تنظیم بازار به عنوان حراج‌گر است که در چارچوب قواعد از پیش تعیین شده عمل می‌کنند. مدل‌سازی چنین بازاری با توجه به ناهمگونی عوامل مستقل تصمیم‌گیر و تعامل و یادگیری آنها در طول زمان با مدل‌سازی‌های کلاسیک که عوامل بازار را همگون فرض می‌کنند امکان‌پذیر نیست. مدل‌سازی عامل‌محور و شبیه‌سازی عملکرد بازار برق به عنوان یک رویکرد کارآمد برای مطالعه رفتار بازیگران بازار و تاثیر تعامل آنها بر یکدیگر و بر تعادل بازار در سال های اخیر مطرح شده است. این مدل‌ها توانایی تبیین رفتار عوامل ناهمگون و تعامل آنها را در یک محیط پویا که منجر به تعادل بازار می‌شود را دارند، اما به دلیل گستردگی و پیچیدگی آنها از شبیه‌سازی و تحلیل سناریو به جای حل معادلات سیستم استفاده می‌شود.

یکی از چالش‌های مهم در بازارهای برق رقابتی، مطالعه رفتار عامل‌های تولیدی ناهمگن و استراتژی پیشنهاد قیمت آنها برای یک روز بعد است. هر عامل عقلائی با هدف حداکثر نمودن سود درازمدت، به دنبال انتخاب استراتژی مناسب برای پیشنهاد قیمت در مکانیزم حراج است. در این تحقیق سه استراتژی حریصانه، تصادفی و یادگیرنده مورد مطالعه و بررسی قرار گرفته است. مدل مورد استفاده یک مدل چندعاملی از بازار برق ایران است که مکانیزم حراج در آن یک بازی ایستا و تکراری در نظر گرفته شده است. فرایند

یادگیری تقویتی توزیع‌شده برای یادگیری عامل از نتیجه بازی‌های قبلی مورد استفاده قرار گرفته است. ارزیابی‌ها در حالت تعادل صورت گرفته و نتایج نشان می‌دهد سود حاصل از یادگیری تقویتی، بصورت معناداری بیش از استراتژی حریصانه و تصادفی است. در حالت شوک و تغییرات بار مصرفی، یادگیری تقویتی باعث همگرائی و تعادل شبکه می‌گردد. از محدودیت های این مطالعه می توان به عدم امکان تشکیل ائتلاف و تبانی در بازار و در نظرنگرفتن حالت بازی همکارانه هماهنگ‌شده درمدل و همچنین برونزا فرض کردن میزان مصرف روز بعد اشاره کرد که هر یک از آن ها می توانند موضوعات تحقیقات آتی قرار گیرند.

8. منابع

الف) فارسی

اصغری‌اسکوئی، محمدرضا (1394)، پیش‌بینی سری‌های زمانی مالی با کمک شبکه عصبی تاخیری، *فصلنامه پژوهش‌های اقتصادی ایران*، سال 15، شماره 57، صص 75 – 108.

رحیمیان، مرتضی (1390)، سنجش و تحلیل بازار در بازار برق به کمک اقتصاد محاسباتی عامل محور، رساله دکتری، دانشگاه فردوسی مشهد.

ناظمی، علی، خوش‌اخلاق، رحمان، عمادزاده، مصطفی، شریفی، علی‌مراد (1390)، برآورد قدرت بازار در بازار برق عمده فروشی ایران، *تحقیقات مدل‌سازی اقتصادی*، دوره1، شماره4، صص 31 – 55.

مویدکاظمی، حمیدرضا و شیخ‌الاسلامی، محمدکاظم (1393)، ارزیابی راهبردهای توسعه تولید از دید رگولاتور با در نظر گرفتن دینامیک سرمایه‌گذاری بازیگران بازار»، بیست و هشتمین کنفرانس بین المللی برق، تهران، ایران، صص 1- 7.

فلاحی، فرهاد (1391)، «گزارش تحلیلی بازار برق ایران: کاربست شبیه‌سازی عامل محور در مطالعات بازار برق»، گزارش‌های پژوهشکده انرژی و محیط زیست پژوهشگاه نیرو، گروه پژوهشی اقتصاد و مدیریت برق.

فلاحی، فرهاد (1392)، «بررسی و تعیین مدل خرید انرژی الکتریکی در بازار روزانه از دید بهره بردار مستقل سیستم»، گزارش‌های پژوهشکده انرژی و محیط زیست پژوهشگاه نیرو، گروه پژوهشی اقتصاد و مدیریت برق.

مشیری، سعید، مروت، حبیب، ونصیری ، عباس (1397)، بررسی تاثیر افزایش قیمت سوخت بر قیمت برق با استفاده از مدلسازی عامل بنیان بازار برق، *فصلنامه مطالعات اقتصاد انرژی*، سال 14، شماره 56، صص 12-22.

حاج‌رسولیها، حسین (1393). هوش مصنوعی پیشرفته، تالیف راسل و نورینگ، جلد دوم، چاپ سوم، تهران: نشر نیازدانش.

ب) انگلیسی

Rashedi, N., Tajeddini, M. A., and Kebriae, H. (2016), Markov Game Approach for Multi-Agent Competitive Bidding Strategies in Electricity Market, *IET Generation, Transmission & Distribution*, Vol.10, No.15, pp. 3756-3763.

Rashedi, N., and Kebriaei, H. (2014), Cooperative and Non-Cooperative Nash Solution for Linear Supply Function Equilibrium Game. *Applied Mathematics and Computation*, Vol. 244, pp.794-808.

Rahimiyan, M., and Rajabi Mashhadi, H. (2010), An Adaptive Q-Learning Algorithm Developed for Agent-Based Computational Modeling of Electricity Market. *IEEE in Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, Vol.40, Issue.5, pp. 547-556.

Krause, Th., Vdovina Beck, E., Cherkaoui, R., Germond, A., Andersson, G. and Ernst, D. (2006), A Comparison of Nash Equilibria Analysis and Agent-Based Modeling for Power Markets. *International Journal of Electrical Power & Energy Systems*, Vol.28, Issue.9, pp.599-607.

Busoniu, L., Babuska, R. and De Schutter B.(2008), A Comprehensive Survey of Multiagent Reinforcement Learning. *IEEE in Transactions on Systems, Man, And Cybernetics-Part C: Applications and Reviews*, Vol.38, Issue.2, pp.1-18.

Lauer, M., and Riedmiller, M. (2000), an Algorithm for Distributed Reinforcement Learning in Cooperative Multi-Agent Systems. In

Proceedings of the Seventeenth International Conference on Machine Learning.

Toman, M. and Barbora, J. (2003), Energy and Economic Development: an Assessment of the State of Knowledge, Discussion Paper 03–13, Resources for the Future, Washington.