طراحی سازوکار تدارکات براساس یادگیری Q و با سیاست انتخاب عمل مبتنی بر الگوریتم ازدحام ذرات

زهره کاهه[1]، دکتر رضا برادران کاظم‌زاده[2*]

دانشگاه تربیت مدرس

تاریخ دریافت مقاله: ۱۳۹۴/۰۸/۱۶
تاریخ پذیرش مقاله: ۱۳۹۴/۱۱/۰۳

چکیده

در این مقاله، مسئله مناقصه در یک شرکت خودروسازی برای تدارک قطعات مورد نیاز از تأمین‌کنندگان بالقوه از طریق الگوریتم یادگیری Q حل شده‌است. در این مسئله، خریدار با توجه به پیشنهادات دریافتی از تأمین‌کنندگان بالقوه که شامل قیمت و زمان تحویل پیشنهادی است، سفارش قطعات مورد نیاز خود را به تأمین‌کنندگان تخصیص می‌دهد. هدف خریدار کمینه‌سازی هزینه‌های تدارکات از طریق یادگیری از مناقصات پیشین است. این مسئله به‌صورت یک مسئله تصمیم‌گیری مارکوفی تعریف شده است که در آن هر عمل وابسته به عمل و وضعیت قبلی است. برای حل این مسئله یک الگوریتم یادگیری تقویتی به نام الگوریتم یادگیری Q توسعه داده شده است که در آن از الگوریتم بهینه‌سازی ازدحام ذرات به‌عنوان راهکاری برای یافتن و انتخاب سیاست بهینه برای انتخاب عمل در الگوریتم یادگیری Q استفاده شده‌است. در مقایسه این الگوریتم با حالتی که در آن سیاست انتخاب عمل مطابق با یک الگوریتم حریصانه است، این الگوریتم بسیار کارآمدتر است.

واژه‌های کلیدی: تدارکات، یادگیری تقویتی، الگوریتم یادگیری Q، سیاست انتخاب عمل.

۱- مقدمه

تدارکات در زنجیره تأمین فرآیندی است که طی آن تأمین‌کننده محصولات خود را مطابق با سفارش مشتری ارسال می‌نماید. مدیران باید برای سازوکار تدارکات مواد مستقیم - موادی که به‌طور مستقیم در ساخت محصول نهایی استفاده می‌شود- و مواد غیرمستقیم - موادی که برای پشتیبانی عملیات استفاده می‌شود- و مواد راهبردی و غیر راهبردی تصمیم‌گیری نمایند. برای تدارک هر مورد از مواد مستقیم و غیرمستقیم، و مواد راهبردی و غیر راهبردی مکانیزمی با سودآوری بیشتر برای زنجیره تأمین تعریف نمایند[۱]. به‌عنوان مثال یک شرکت باید تدارکات را به‌گونه‌ای تنظیم نماید که امکان ایجاد توافق برد- برد بین تأمین‌کننده و خریدار برقرار بوده و منافع و محدودیت‌های طرفین در نظر گرفته شود. همچنین هزینه‌ها از جمله هزینه تدارکات، جریمه تأخیر نیز کمینه شود. در واقع هدف تدارکات این است که سفارشات و تحویل‌ها براساس زمان‌بندی و با کمترین هزینه انجام گیرند.

در برخی مطالعات اشاره شده است که طراحی مکانیزم تدارکات در دو دسته مبتنی بر حراج معکوس (مناقصه) و مبتنی بر مذاکره قرار می‌گیرد. در برخی مطالعات، مناقصه (مزایده) نوع خاصی از مذاکره در نظر گرفته شده است[۲]؛ اما در برخی مطالعات دیگر نظیر مطالعه چن و تی‌سنگ[3] بین این دو مفهوم تمایز قائل شده‌اند[۳]. در این مقاله فقط بر مکانیزم تدارکات مبتنی بر مناقصه تمرکز می‌شود. در این سازوکار به‌طور معمول خریدار طی یک فرآیند جستجوی متوالی با فروشندگان در تعامل است، تا زمانی که یک راه‌حل رضایت‌بخش حاصل گردد.

۱- کارشناس ارشد دانشگاه تربیت مدرس، پست الکترونیک: zohreh.kaheh@modares.ac.ir
۲*- دانشیار دانشگاه تربیت مدرس، دانشکده فنی مهندسی، نویسنده پاسخگو، پست الکترونیکی: rkazem@modares.ac.ir، نشانی: تهران، بزرگراه جلال آل احمد، دانشگاه تربیت مدرس، دانشکده فنی مهندسی، بخش مهندسی صنایع

3- Chen & Tseng

عامل‌ها در محیط‌های چالش‌برانگیز استفاده می‌شوند، چالش برانگیز بودن محیط بدین معناست که محیط آنها عموماً (۱) پویا، (۲) غیرقابل پیش‌بینی و (۳) غیرقابل اعتماد است.

(۱) پویایی محیط، ناشی از تغییرات سریع آن است، به‌طوری‌که عامل دستیابی به اهدافش نمی‌تواند فرض کند که محیط ایستا باقی می‌ماند.

(۲) غیرقابل پیش‌بینی بودن محیط ناشی از این است که موقعیت‌های بعدی در محیط قابل پیش‌بینی نیستند. این امر اغلب به این دلیل است که یک عامل - نه فقط به علت دانش اندک بلکه گاهی به‌علت عدم دسترسی به اطلاعات- اطلاعات کامل و جامعی از محیط خود ندارد و یا تغییرات محیط ورای اختیار یا دانش عامل می‌باشد.

(۳) غیرقابل اعتماد بودن محیط به این علت است که اعمالی که عامل انجام می‌دهد ممکن است به‌علتی ورای کنترل عامل، با شکست مواجه شود[۴].

عامل‌های هوشمند اساساً بر بعد هوشمندی تمرکز دارند. یک عامل هوشمند می‌تواند برخی جنبه‌های هوش مصنوعی نظیر استدلال، انطباق‌پذیری و یادگیری را در اجرای وظایف خود به کار گیرد. مطابق با سطح هوشمندی عامل‌ها بیم و سگو[1] [۵] عامل‌های هوشمند را به دو دسته تقسیم نمودند:

الف) عامل‌های هوشمند بدون یادگیری ماشینی: این مورد هیچ قابلیت یادگیری ندارند و به سختی می‌توانند تجربیات گذشته و اطلاعات محیط پویا را در نظر بگیرند. این نوع عامل‌ها به سادگی اجرا و تجزیه و تحلیل می‌شوند. بیشتر سیستم‌های مبتنی بر عامل کنونی متعلق به این دسته هستند. آنها با قابلیت استدلال ساده می‌توانند مذاکره را آسان نمایند.

ب) عامل‌های هوشمند با یادگیری ماشینی: در مقابل عامل‌ها با قابلیت یادگیری ماشینی یک پشتیبان تصمیم ساختگی برای انطباق با محیط پویا و رفتار متغییر رقیبان فراهم می‌آورند. به‌طور واضحی داشتن یک راهبرد با سیستم پشتیبان تصمیم مصنوعی برای تطبیق با شرایط پویای محیط و تغییر رفتار رقبا مرجح است. تکنیک‌های مختلف مبتنی بر هوش مصنوعی و روش‌های ابتکاری و فرابتکاری برای فرآهم آوردن رفتار تطبیق‌پذیر برای عامل‌ها نظیر الگوریتم ژنتیک، یادگیری بیزی، و شبکه‌های عصبی به‌وجود آمده‌اند. عامل‌های هوشمند با قابلیت یادگیری قادرند تا راهبردهای مذاکره تطبیق‌پذیر و وابسته به دانش و تجربه خودشان و اطلاعات موجود را ایجاد نمایند. دانش از این نوع می‌تواند به عامل‌ها برای جستجوی فضای جواب به صورت کارا و مؤثر کمک کند و بازده مذاکره را به‌طور کامل بهبود بخشد.

یکی از انواع یادگیری، یادگیری تقویتی است که عبارتست از آموختن اینکه چه عملی انجام گیرد تا یک سیگنال پاداش عددی بیشینه شود. در این الگوریتم برخلاف اکثر الگوریتم‌های یادگیری ماشینی به یادگیرنده گفته نمی‌شود که کدام عمل را انجام دهد، بلکه مشخص می‌شود که انجام کدام عمل بیشترین پاداش را در پی دارد. در موارد پیچیده‌تر نه فقط پاداش‌های فوری مطرح است بلکه حالت‌ها و موقعیت‌های بعدی که انجام عمل منجر به آنها شده و نیز پاداش‌های آتی نیز مورد توجه قرار می‌گیرد. در واقع دو خصوصیت "جستجوی مبتنی بر سعی و خطا" و "پاداش‌های آتی" دو مشخصه اصلی یادگیری تقویتی هستند[۶].

در این پژوهش مسئله انجام تدارکات، به عنوان یک مسئله تخصیص سفارشات به تأمین‌کنندگان در یک زنجیره تأمین با وجود یک خریدار و تعدادی تأمین‌کننده در نظر گرفته شده است. همچنین فرض شده که هیچ یک از تأمین‌کنندگان از قرارداد سایر تأمین‌کنندگان با خریدار اطلاعی ندارد و در واقع قراردادها از نوع مهر و موم شده هستند. مسئله به این شرح است که تولیدکننده (خریدار) قصد دارد تقاضای خود را برای چند نوع کالا به چند تأمین‌کننده سفارش دهد. به‌این صورت که مجموع سفارشات برای هر نوع کالا برابر تعداد مورد تقاضای آن باشد. همچنین خریدار یک زمان تحویل فرضی را در نظر می‌گیرد. در این مسئله لازم است جوانب مختلف تصمیم‌گیری از جمله پذیرش یا عدم پذیرش پروپوزال‌ها، ایجاد یک تعادل بین قیمت و زمان تحویل پیشنهادی در نظر گرفته شود. از طرفی هر یک از تأمین‌کنندگان یک قیمت برای تعداد حجم پیشنهادی برای هر نوع کالا (برای انواع مختلف کالا جداگانه قیمت داده می‌شود) و زمان تحویل (به روز) را مشخص می‌کنند. خریدار قصد دارد میزان مناسب تخصیص به هر یک از تأمین‌کنندگان را با این فرض که حجم‌های قابل تأمین تامین‌کنندگان مختلف به صورت مستقل از هم از فرآیند پوآسون پیروی می‌کند، را بیاموزد، به‌طوری‌که کمترین هزینه را داشته باشد. در واقع در نظر گرفتن این مسئله به‌صورت یک تصمیم‌گیری مارکوفی با یک قدم با

1- Beam & Segev

این فرض انجام می‌شود که احتمال تغییر حالت فرآیند صرفاً به حالت و عمل قبلی وابسته است و مستقل از خاطره عمل‌های قبلی است. به عبارت دیگر فرض می‌شود که پذیرش یا عدم پذیرش پیشنهاد به پذیرش پیشنهاد قبلی آن وابسته است.

۲- مرور ادبیات:

چهارسوقی[1] و همکاران [۷] یک رویکرد جدید برای تصمیم‌گیری در مورد سیاست سفارش‌دهی اعضای زنجیره تأمین به‌صورت متمرکز پیشنهاد نموده‌اند. آنها زنجیره تأمین را به‌عنوان یک سیستم چند عاملی در نظر گرفته و براساس یادگیری تقویتی فرموله نموده‌اند. نتایج آنها نشان داده است که مکانیزم یادگیری تقویتی نسبت به الگوریتم ژنتیک عملکرد بهتری داشته است. لی[2] و همکاران [۸] یک شرکت با سیاست تولید برای سفارش را در نظر گرفته‌اند که توانایی رد یا قبول سفارشات را دارد و می‌تواند قیمت‌ها و زمان تحویل‌ها را تنظیم نماید. آنها حالتی را که شرکت نیاز دارد تا در رابطه با اینکه چه سفارشاتی را رد یا قبول نماید به‌صورتی‌که یک تعادل بین قیمت و زمان تحویل به‌وجود آورد را در نظر گرفته‌اند. آنها این مسئله را به‌عنوان یک مسئله تصمیم‌گیری مارکوف در نظر گرفته و یک یادگیری تقویتی مبتنی بر الگوریتم یادگیری Q برای این مسئله توسعه داده‌اند. ایما و کورئه[3] [۹] یک الگوریتم یادگیری تقویتی جمعی مبتنی بر بهینه‌سازی انبوه ذرات به‌منظور یافتن سریع سیاست بهینه پیشنهاد دادند به‌صورتی‌که در این الگوریتم چندین عامل وجود دارد که نه فقط از یادگیری خود استفاده می‌کنند بلکه الگوریتم PSO را به‌روزرسانی می‌نمایند. در الگوریتم پیشنهادی آنها مقادیر حالت - عمل براساس بهترین جواب جمعی و جواب شخصی یافت‌شده توسط عامل‌ها به‌روزرسانی می‌شود. گیامونکارو و پنتراندولفو[4] [۱۰] یک رویکرد برای مدیریت موجودی ارائه داد که از طریق تعیین سیاست سفارش‌دهی درصدد بهینه‌سازی عملکرد کل زنجیره تأمین بود.

تانگ[5] و همکاران [۱۱] یک خط تولید تقاضا محور را به همراه مرکز فروش مرتبط با آن به‌عنوان یک سیستم دو سطحی در نظر گرفتند. این سیستم ویژگی‌های دریافت غیرقطعی مواد، تقاضای تصادفی مشتری، زمان پردازش نامعین و موجودی بافر محدود را در بر می‌گیرد. آنها یک الگوریتم ترکیبی یادگیری Q - شبیه‌سازی تبرید برای این مسئله پیشنهاد دادند.

فو و فو[6] [۱۲] یک سیستم چندعاملی یکپارچه با استدلال آگاه از زمینه و موضوع[7] برای بهبود مدیریت همکارانه زنجیره تأمین پیشنهاد نمودند. سیستم پیشنهادی آنها موجب یکپارچگی بیشتر و تقویت یادگیری و انطباق با محیط زنجیره تأمین می‌شود.

مرتضوی[8] و همکاران [۱۳] از طریق یک سیستم مبتنی بر عامل و با تعبیه الگوریتم یادگیری تقویتی موجب بهبود یکپارچگی در یک زنجیره تأمین چهار سطحی با تقاضای نامعین شدند.

در ادبیات چندین نوع طراحی مختلف برای عامل‌ها وجود دارد. درحالتی‌که عامل، پاداشی مرتبط با مطلوبیت خود دریافت می‌کند، دو نوع طراحی مختلف برای عامل در نظر گرفته می‌شود[۱۴]:

- عامل یک تابع مطلوبیت براساس وضعیت خودش یا تاریخچه وضعیت‌هایی که داشته، یاد می‌گیرد و آن را برای انتخاب عملی که خروجی آن بیشترین مطلوبیت مورد انتظار را داشته باشد، استفاده می‌نماید.
- عامل یک تابع عمل - ارزش یاد می‌گیرد که مطلوبیت مورد انتظار این عمل را در موقعیت داده شده مشخص کند. این نوع یادگیری که به‌عنوان یادگیری Q معروف است، نوع خاصی از یادگیری تقویتی نحسوب می‌شود.

در این مقاله مسئله تدارکات مبتنی بر مناقصه در چند مرحله به‌صورت یک مسئله تصمیم‌گیری مارکوفی تعریف می‌شود و سیاست انتخاب عمل در الگوریتم یادگیری Q براساس بهترین مقادیر شخصی و جهانی ذرات در الگوریتم PSO به‌روزرسانی می‌شود که نسبت به حالتی که انتخاب عمل به‌صورت حریصانه انجام می‌شود، بسیار کاربردی‌تر است. رویکرد مبتنی بر تعبیه یک الگوریتم یادگیرنده در عامل خریدار (برپاکننده مناقصه) به‌منظور یادگیری درباره انتخاب پروپوزال‌ها در چند دوره مناقصه پیاپی در حوزه مطالعات مربوط به حراج معکوس (مناقصه) یک نوآوری محسوب می‌شود. همچنین به‌رغم سایر مطالعات که فقط براساس قیمت‌های پیشنهادی هستند، می‌توان از نوآوری‌های

1- Chaharsooghi
2- Li
3- Iima & Kuroe
4- Giannoccaro & Pontrandolfo
5- Tang
6- Fu & Fu
7- Context-aware Reasoning
8- Mortazavi

دیگر در این زمینه به در نظر گرفتن هزینه‌های ناشی از دیرکرد در انتخاب پروپوزال تأمین‌کننده اشاره کرد.

۳- یادگیری تقویتی

مسئله یادگیری تقویتی مکانیزم یادگیری حاصل از تعاملات را به‌منظور دستیابی به اهداف بیان می‌کند. در این مسئله یادگیرنده و تصمیم‌گیرنده عامل نام دارد. موجودیت‌هایی که تعامل عامل با آنها صورت می‌پذیرد محیط نامیده می‌شود. در این تعامل به‌طور دائم عامل (در مسئله ما خریدار) اعمالی راانتخاب می‌نماید و محیط (تأمین‌کنندگان یا پیشنهاددهندگان در مناقصه) به این اعمال پاسخ می‌دهد و موقعیت جدید را به عامل عرضه می‌نماید. نتیجه برقراری ارتباط عامل با محیط پاداش‌ها (ارزش‌های عددی خاص که عامل سعی در بیشینه کردن آنها در طی زمان دارد) و دستیابی به موقعیت‌های جدید است[۱۵]. در هر گام زمانی عامل یک نگاشت از حالات به احتمالات انتخاب هر عمل ممکن را به‌کار می‌برد. این نگاشت سیاست عامل نامیده می‌شود. سیاست عامل با $\Pi_t(s,a)$ نشان داده می‌شود که عبارت است از احتمال اینکه در دوره t حالت سیستم s باشد و عمل a انجام شود. روش‌های یادگیری تقویتی مشخص می‌نماید که عامل چگونه سیاست‌های خود را بر طبق تجربیات خود تغییر دهد. هدف عامل به‌طورکلی بیشینه کردن میزان کلی پاداش‌ها در دراز مدت است. هر چیزی که به‌صورت دلخواهانه امکان تغییر توسط عامل را نداشته باشد، جزیی از محیط محسوب می‌شود. در حقیقت ممکن است عامل درباره چگونه کار کردن محیط بداند اما آن را تحت کنترل نداشته باشد. بنابراین حدفاصل عامل- محیط نشان‌دهنده محدودیت عامل در کنترل است و نه دانش آن. یکی از مهم‌ترین نقاط قوت یادگیری تقویتی توسعه الگوریتم یادگیری Q می‌باشد[۱۶]. رابطه به‌روزرسانی این الگوریتم در ساده‌ترین فرم این الگوریتم که الگوریتم یادگیری Q تک قدمی نامیده می‌شود، به صورت فرمول ۱ تعریف می‌شود[۱۷]، که در آن مقدار Q(s,a) عبارت است از مجموع پاداش‌های دریافت شده هنگامی که از حالت S شروع و عمل a را انجام و خط‌مش موجود را دنبال نموده باشد. اندیس t و t+1 اشاره به دوره قبلی و دوره فعلی دارد. α نرخ یادگیری و نشان‌دهنده میزان تأثیر دوره‌های آتی است که مقداری بین صفر تا کوچک‌تر یا مساوی یک می‌گیرد. $R(s_t)$ پاداش مربوط به s_t و γ پارامتر ارزش زمانی است که مقداری بزرگ‌تر یا مساوی صفر و کوچک‌تر از یک می‌باشد.

(۱)

$$Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \alpha\left[R(s_t) + \gamma \max_{a_{t+1}} Q(s_{t+1},a_{t+1}) - Q(s_t,a_t)\right]$$

در این الگوریتم تابع ارزش- عمل آموخته شده Q، به‌طور مستقیم و مستقل از سیاستی که پیگیری شده است، مقدار Q* را تخمین می‌زند. این مطلب آنالیز الگوریتم را ساده می‌نماید و اثبات همگرایی را ممکن می‌سازد [۱۵]. الگوریتم یادگیری Q در ساده‌ترین صورت به صورت زیر می‌باشد:

- متغیر γ و ماتریس پاداش محیط تنظیم می‌شود.
- مقادیر تابع ارزش با مقادیر صفر مقداردهی اولیه می‌شود.
- برای هر تکرار:

✓ موقعیت اولیه به صورت تصادفی انتخاب می‌شود.

✓ تا زمانی‌که به موقعیت هدف دست نیافته است:

❖ از میان تمام عمل‌های ممکن برای موقعیت فعلی یکی انتخاب می‌شود.

❖ با استفاده از عمل ممکن گذر به حالت بعدی لحاظ می‌شود.

❖ مقدار بیشینه Q برای موقعیت بعدی براساس تمام اعمال ممکن به‌دست آورده می‌شود.

❖ حالت بعدی به‌عنوان حالت جاری در نظر گرفته می‌شود.

$$Q = \left[Q_{1,1}, Q_{1,2}, Q_{1,3}, \ldots, Q_{5,25}, Q_{5,26}\right]$$

۴- جزئیات طراحی مکانیزم تدارکات مبتنی بر یادگیری عامل خریدار

در مکانیزم پیشنهادی، خریدار با شبیه‌سازی محیط مناقصه، به‌منظور دستیابی به یک توافق و براساس یک الگوریتم یادگیری به‌صورت تک عاملی به‌دنبال انجام تدارکات با هدف کاهش هزینه‌های خود است. جزئیات الگوریتم یادگیری Q که نوعی یادگیری تقویتی است در ادامه مطرح شده‌اند:

۴-۱- حالت سیستم

در چارچوب یادگیری تقویتی، عامل تصمیمات خود را مبتنی بر تابعی از سیگنال‌های دریافت شده از محیط که حالت محیط نام دارد، بنا می‌کند. در واقع براساس حالت محیط، حالت عامل یادگیرنده مشخص می‌شود. در مسئله ما ورودی‌های عامل خریدار عبارتند از قیمت‌ها و زمان‌های تحویل دریافت شده از تأمین‌کنندگان. خریدار براساس شبیه‌سازی محیط باید بهترین تصمیم را برای تغییر حالت

از حالت فعلی به حالت بعدی لحاظ نماید. حالت عامل عبارت است از مقدار تقاضای (باقیمانده تقاضا) خریدار در رابطه با هر قطعه مورد نیاز، به‌طوری‌که در حالت اولیه، خریدار میزان تقاضای کل برای هر نوع قطعه را نشان می‌دهد. فضای حالت مجموعه تمام حالت‌هایی است که از حالت اولیه طی توالی عملیات مختلف قابل دستیابی است که عبارت است از میزان تقاضای باقیمانده خریدار. حالت هدف برای خریدار نیز عبارت است از حالتی که در آن مانده تقاضا برابر صفر باشد و تقاضای خریدار ارضا شود.

۴-۲- عمل و سیاست انتخاب عمل

برای حرکت عامل خریدار از حالتی به حالت دیگر باید عمل انجام بگیرد. عمل‌های امکان‌پذیر برای تغییر از حالت اولیه به حالت‌های شدنی دیگر عبارت خواهد بود از مقادیر تخصیص یافته به هر تأمین‌کننده(Q) (پیشنهادهای پذیرفته شده در هر مرحله از مناقصه) که با توجه به اینکه پنج تأمین‌کننده و ۲۶ نوع قطعه در نظر گرفته شده است به صورت زیر می‌باشد که در آن $Q_{1,3}$ یعنی میزان محصول ۳ که به تأمین‌کننده اول تخصیص داده می‌شود و $Q_{3,10}$یعنی میزان محصول ۱۰ که به تأمین‌کننده سوم تخصیص داده می‌شود.

سیاست انجام این عمل مطابق با سیاست حل مسئله بهینه‌سازی است. عموماً سیاست انتخاب عمل می‌تواند از روش‌های حریصانه یا سافت مکس[1] که در آن انتخاب عمل به مقیاسی عددی تحت عنوان دما یا درجه حرارت بستگی دارد، استفاده می‌شود. لذا در آن درجه حرارت پارامتری مثبت است:

$$p(a_i)=\frac{e^{\frac{Q(s,a_i)}{T}}}{\sum_{j=1}^{S_k} e^{\frac{Q(s,a_i)}{T}}} \qquad (۲)$$

در رابطه فوق $p(a_i)$ احتمال انتخاب عمل و s_k تعداد اعمال انتخاب شده در حالت s می‌باشد. حال اگر مقادیر Q هزینه مورد انتظار را بازنمایی کنند، با وارد نمودن مقادیر Q با علامت منفی در فرمول انتخاب عملی با بالاترین احتمال، کم‌ترین هزینه را در پی خواهد داشت. عموماً درجه حرارت از مقادیر بزرگ شروع شده و به تدریج نسبت به زمان کاهش می‌یابد. در این مقاله سیاست جستجوی بهترین عمل براساس بهترین مقادیر شخصی و جهانی ذرات الگوریتم بهینه‌سازی هوش ازدحامی ذرات PSO بنا شده است.

۴-۳- هدف

در حقیقت هدف از ایجاد این الگوریتم یادگیرنده تعیین یک تخصیص سفارشات منطقی در چند مرحله مناقصه پیاپی ازسوی خریداراست، به‌طوری‌که هزینه تدارکات را کمینه نماید.

۴-۴- تابع پاداش

یک تابع پاداش نگاشتی از حالت مشاهده شده به یک عدد واحد، پاداش نامیده می‌شود. پاداش میزان مطلوبیت حالت سیستم (یا عمل انجام شده در آن حالت) است، بنابراین تابع پاداش باید تعریف‌کننده اهداف در مسئله یادگیری تشدیدی باشند. پاداش به‌عنوان خروجی تابع پاداش یک مقدار مثبت است که هرگاه به بهترین حالت دست یافتیم پاداش دریافت می‌کنیم. از آنجا که هدف این الگوریتم کاهش هزینه تدارکات است، در نتیجه لازم است که تابع پاداش در برگیرنده هزینه باشد. بنابراین هزینه خریدار را براساس قیمت تدارکات و زمان‌های دیرکرد که موجب هزینه مواجهه با کسری موجودی می‌شود در نظر می‌گیرند. در واقع خریدار با تخصیص سفارش به تأمین‌کنندگان بهتر، قیمت و زمان تحویل مناسب‌تری را شاهد خواهد بود.

۴-۵- تابع ارزش

ارزش یک حالت عبارت است از حجم کلی پاداشی که یک عامل می‌تواند توقع داشته باشد تا در آینده با شروع از حالت اولیه اندوخته شود و برخلاف پاداش‌ها که میزان مناسب بودن حالات آنی محیط هستند ارزش‌ها نشان‌دهنده میزان مطلوبیت طولانی مدت حالات می‌باشند. با توجه به استفاده از الگوریتم Q در اینجا، لذا باید توابع ارزش عمل نیز مورد استفاده قرار گیرد. بدین منظور از رابطه (۱) که در توصیف کلی الگوریتم به آن اشاره شد، استفاده شده است. به‌طورکلی ارزش عمل a در حالت مفروض s برابر است با پاداشی که در نتیجه انجام عمل a در حالت s به سیستم تعلق می‌گیرد. به‌علاوه پاداشی که در نتیجه انجام بهترین عمل بنابر یادگیری‌های پیشین صورت گرفته در حالت بعدی سیستم که s' است به سیستم تعلق می‌گیرد. به منظور انجام محاسبات مقادیر اولیه برای توابع ارزش عمل به‌طور اختیاری صفر در نظر گرفته می‌شود.

1- Soft Max

۵- جزئیات الگوریتم PSO به عنوان سیاست انتخاب عمل در الگوریتم یادگیری Q

در این مقاله سیاست انتخاب عمل براساس الگوریتم بهینه‌سازی هوش ازدحامی ذرات PSO بنا شده است. اولین بار در سال ۱۹۹۵، الگوریتم بهینه‌سازی حرکت دسته جمعی (PSO) توسط کندی و ابرهارت[1] معرفی شد [۱۸].

PSO از هوش جمعی الهام گرفته شده است. این روش سعی در تقلید رفتار اجتماعی ارگان‌های طبیعی نظیر دسته پرندگان و ماهی‌ها در یافتن غذا دارد. در این‌گونه جمعیت‌ها بدون آنکه کنترل مرکزی صورت بگیرد یک رفتار هماهنگ شده با استفاده از حرکات محلی بروز پیدا می‌کند [۱۹]. همان‌طور که اشاره شد از الگوریتم ازدحام ذرات به‌عنوان راهکاری برای یافتن و انتخاب سیاست بهینه برای انتخاب عمل استفاده شده است؛ در واقع به این ترتیب انتخاب اعمالی که خریدار را از حالتی به حالت دیگر انتقال می‌دهند یعنی مقادیر تخصیص اقلام‌های مختلف به تأمین‌کنندگان مختلف به‌وسیله الگوریتم PSO مشخص می‌شود. در واقع در این فرآیند مقادیر عمل و حالت براساس بهترین مقادیر شخصی و جهانی ذرات در الگوریتم PSO به‌روزرسانی می‌شود. همچنین در مسائل یادگیری تقویتی با افزایش مقیاس در مسائل بزرگ مسئله تعداد حالت‌ها افزایش می‌یابد. لذا ذخیره مقادیر همه حالت‌ها هزینه استفاده از ظرفیت حافظه را به شدت افزایش می‌دهد. در مسئله ما به این علت که خریدار در درک محیط خود با محدودیت مواجه است، لذا تعیین حالت جاری به‌طور کامل غیرممکن است. بنابراین عامل خریدار به این علت که نمی‌تواند به‌طور کامل حالت محیط خود را حس کند، دچار اشتباه می‌شود و سیستم از خاصیت مارکوفی دور می‌افتد؛ لذا برای جلوگیری از این امر از الگوریتم‌های ابتکاری استفاده می‌شود و [۹و۲۰]. در واقع در این مقاله از رویکردی مشابه[۹] استفاده گردیده است. جزئیات الگوریتم PSO برای انتخاب عمل در زیربخش‌های بعد آمده است.

۵-۱- ایجاد جمعیت اولیه در الگوریتم PSO

به تعداد تأمین‌کننده‌ها آرایه در نظر گرفته می‌شود، به این صورت که آرایه اول مربوط به تأمین‌کننده اول، و آرایه nام مربوط به تأمین‌کننده nام است. این آرایه‌ها مقادیر تقاضای خریدار را در بر می‌گیرند. برای هر قطعه به‌صورت تصادفی از بین تأمین‌کننده‌های ۱ تا n یک تأمین‌کننده را به‌صورت تصادفی انتخاب کرده و به آن به‌صورت تصادفی عددی در بازه ((0 , Min(Q_{ij}^{max} ,remaining demand)) از آن قطعه سفارش داده می‌شود. پس از این عمل تقاضای باقیمانده وضعیت عامل را نشان می‌دهد. به‌عنوان مثال فرض کنید سه تأمین‌کننده وجود داشته و تعداد کل تقاضا برابر ۱۰۰ باشد، از بازه [۱و۳] به صورت تصادفی عدد ۲ یعنی تأمین‌کننده دوم انتخاب می‌شود، با توجه به اینکه هنوز تخصیصی انجام نشده است و حداکثر میزانی که به آن تأمین‌کننده تخصیص می‌یابد برابر ۸۰ باشد، به آن عددی تصادفی در بازه ۰ تا Min[80,100] تخصیص داده می‌شود. به‌طور مثال به تأمین‌کننده دوم، ۴۳ عدد تخصیص می‌دهیم. حال یک عدد تصادفی دیگر از بازه ۱ تا ۳ به جز عدد انتخاب شده، مثلاً عدد ۱ انتخاب می‌شود. با فرض اینکه حداکثر بتوان به آن ۱۲۰ عدد تخصیص داد و مانده تقاضا برابر ۵۷ است، عدد تصادفی در بازه ۰ تا ۵۷ مثلاً ۵۰ عدد به تأمین‌کننده اول تخصیص داده می‌شود و در نهایت ۷ واحد مانده را به تأمین‌کننده سوم تخصیص می‌دهد. با در نظر گرفتن این فرآیند برای موضوع‌های دیگر نمایش جواب به‌صورت یک ماتریس i*j خواهد بود که آرایه‌های آن مقادیر تخصیص داده شده q یعنی عمل‌های ممکن می‌باشند.

۵-۲- بردار سرعت اولیه در الگوریتم PSO:

بعد از اینکه جمعیت اولیه ساخته شد یک بردار سرعت به‌صورت تصادفی ایجاد نمایید. توجه کنید که بردار سرعت اولیه برابر صفر در نظر گرفته شده باشد.

۵-۳- ایجاد جمعیت بعدی در الگوریتم PSO:

همان‌طور که بیان شد، در گام اول یک جمعیت تصادفی از ذره‌ها ساخته شد؛ حال برای ایجاد مکان و سرعت جدید و به‌منظور به‌روزرسانی عمل‌ها طبق فرمول‌های PSO عمل نمایید که به صورت زیر خواهد بود:

$$v_{ij}^{t} = \omega v_{ij}^{t-1} + c_1 r_1 (p_{ij}^{t-1} - x_{ij}^{t-1}) + c_2 r_2 (G_j^{t-1} - x_{ij}^{t-1})$$

$$x_{ij}^{t} = x_{ij}^{t-1} + v_{ij}^{t} \qquad (۳)$$

اگر جمع مقادیر تخصیص داده شده برای یک نوع قطعه برابر میزان تقاضا برای آن قطعه نباشد، می‌توان اذعان کرد که جمع مقادیر تخصیص داده شده با میزان تقاضا برای آن مقایسه می‌شود. لذا اگر تساوی برقرار بود که مشکلی وجود ندارد، اما اگر جمع مقادیر بزرگ‌تر باشد باید از میان جواب‌هایی که موقعیت جدیدشان نسبت به موقعیت قبلی

1- Kennedy & Eberhart

آنها بزرگ‌تر است (یعنی $x_{new} - x_{old} > 0$)، به صورت تصادفی انتخاب شده و یک واحد یک واحد از آنها کم شود تا تساوی برقرار گردد. همچنین اگر جمع مقادیر کوچک‌تر باشد، از میان جواب‌هایی که موقعیت جدید آنها نسبت به موقعیت قبلی آنها کوچک‌تر شده است (یعنی $x_{new} - x_{old} < 0$) به صورت تصادفی انتخاب شده و یک واحد یک واحد از آنها کم شود تا تساوی برقرار گردد. با انجام این کار موقعیت جدید اصلاح شده و سرعت جدید نیز طبق رابطه (۴) اصلاح می‌شود.

$$(v_{new})_{\text{modified}} = (x_{new})_{\text{modified}} - x_{old} \qquad (۴)$$

حال مقدار تابع هدف به ازای $(x_{new})_{\text{modified}}$ محاسبه می‌شود. در واقع جواب‌های جدید به‌عنوان جمعیت جدید ذرات در نظر گرفته می‌شود. بنابراین جمعیت جدید باعث به‌روزرسانی عمل‌ها و حالت‌ها می‌گردد. لازم است که مجدداً اشاره شود که عمل‌ها و حالت‌ها برای مسئله ما بر هر یک از اقلام به صورت جداگانه در نظر گرفته شده است. به عبارت دیگر برای ارضای تقاضای هر یک از انواع اقلام مورد نظر یادگیری متمایز و موازی با سایر اقلام وجود دارد، اما در نهایت هدف ما کاهش هزینه‌های تدارکات و هزینه‌های دیرکرد کلی است.

۶- نتایج محاسباتی

الگوریتم ارائه شده برای حل مسئله تخصیص سفارشات به‌منظور کمینه‌سازی هزینه تدارکات و دیرکرد مطابق یک مسئله در مطالعه موردی در شرکت تأمین قطعات خودروسازی استفاده شده است. در الگوریتم Q متغیرهایی به این شرح وجود دارند: p_s یا احتمال جست‌وجوی پراکنده، p_e یا احتمال بهره‌برداری (جستجوی متمرکز) و Epoch تعداد تکرار شبیه‌سازی سیستم در فاز یادگیری و α نرخ یادگیری که مقدار کم آن باعث می‌شود که عامل بیشتر از اطلاعات قدیمی استفاده نماید و مقدار بزرگ آن باعث می‌شود که فقط اطلاعات جدید را ملاک قرار دهد و γ نیز عامل متغیر ارزش زمانی (نرخ تخفیف) است که مقدار کم آن باعث می‌شود عامل به‌صورت حریصانه عمل نموده و فقط پاداش‌های فعلی را در نظر بگیرد و مقدار بزرگ نرخ تخفیف باعث می‌شود به دنبال پاداش‌های تجمعی در طولانی مدت باشد. با انجام تکرارهای متعدد و تحلیل حساسیت نسبت به مقادیر متغیرها این آگاهی دست می‌دهد که برای احتمال جست‌وجو مقدار ۰/۸ و برای احتمال بهره‌برداری مقدار ۰/۲ و برای نرخ یادگیری مقدار ۰/۶ و برای متغیر ارزش زمانی (تخفیف) مقدار ۰/۸ مناسب است.

به‌منظور تنظیم عوامل متغیر الگوریتم PSO، این الگوریتم با ترکیب‌های مختلف سطوح مقادیر متغیرها اجرا شده‌است؛ به این صورت که سه سطح برای تعداد تکرارها (۸۰، ۱۰۰، ۱۲۰)، دو سطح برای تعداد ذرات (۳۰ و ۴۰)، سه سطح برای یادگیری فردی (۱/۵، ۲، ۲/۵)، سه سطح برای یادگیری جمعی (۱/۵، ۲، ۲/۵)، سه سطح برای W_{min} (۰/۱، ۰/۲، ۰/۳) و سه سطح برای W_{max} (۰/۷، ۰/۸، ۰/۹) در نظر گرفته شده است. بهترین سطح برای مقادیر متغیرهای الگوریتم ازدحام ذرات در الگوریتم ترکیبی در جدول (۱) نشان داده شده‌است:

جدول (۱): بهترین مقادیر عوامل متغیر PSO

تعداد تکرارها	تعداد ذرات	یادگیری فردی(c_1)	یادگیری جمعی(c_2)	W_{min}	W_{max}
۱۲۰	۴۰	۱/۵	۲/۵	۰/۱	۰/۸

در ادامه نتایج حاصل از الگوریتم پیشنهادی (الگوریتم۱) با حالتی که سیاست انتخاب عمل به روش حریصانه (الگوریتم۲) انجام می‌شود (با ثابت نگهداشتن متغیرهای الگوریتم یادگیری تقویتی) مقایسه شده است.

قابل ذکر است که کدنویسی الگوریتم در نرم‌افزار Matlab R2011b انجام شده است و بر یک کامپیوتر core i5 با مشخصات حافظه ۴ گیگابایت اجرا شده است. بنابراین الگوریتم با توجه به مقادیر خطای کم و سرعت حل آن کارا و کارآمد است. اشاره می‌شود که خطا به‌صورت رابطه (۵) محاسبه شده است:

$$dev_{Q-PSO} = \frac{sol_{Q-PSO} - sol_{best}}{sol_{best}}$$

$$dev_{Q-greedy} = \frac{sol_{Q-greedy} - sol_{best}}{sol_{best}}$$

$$sol_{best} = \min\{sol_{Q-PSO}, sol_{Q-greedy}\} \qquad (۵)$$

نتایج مربوط به متوسط هزینه و متوسط زمان اجرا برای تکرارهای مختلف مسئله در دو الگوریتم در جداول (۲) و (۳) قابل مشاهده است. براساس نتایج جدول (۲) مشاهده می‌شود که الگوریتم پیشنهادی قادر است به‌صورت کارآمد با انحراف کم نسبت به بهترین جواب یافت‌شده و در زمان کوتاه‌تر نسبت به حالتی‌که سیاست انتخاب عمل حریصانه است، به نتیجه مناسب دست یابد. در جدول (۳) درصد

کاهش هزینه‌ها با استفاده از رویکرد پیشنهادی نسبت به حالتی‌که سیاست انتخاب عمل حریصانه است، نشان داده شده است.

جدول (۲): نتایج مقایسه دو الگوریتم با توجه به زمان اجرا و انحراف نسبت به بهترین جواب

اجرا	تعداد شرکت کنندگان مناقصه	الگوریتم Q با سیاست انتخاب عمل PSO		الگوریتم Q با سیاست انتخاب عمل حریصانه	
		متوسط انحراف نسبت به بهترین جواب	متوسط زمان (ثانیه)	متوسط انحراف نسبت به بهترین جواب	متوسط زمان (ثانیه)
۱	۱۵	۰/۰۹۵	۱/۳۷۵	۰	۱/۶۳۴
۲	۳۰	۰	۲/۹۲۳	۰	۴/۴۵۱
۳	۴۵	۰/۰۴۳	۴/۴۳۵	۰	۷/۶۱۰
۴	۶۰	۰	۷/۹۷۵	۰/۰۹۵	۹/۳۴۱
۵	۷۵	۰	۱۱/۹۷۰	۰/۱۹۰	۱۳/۵۲۹
۶	۹۰	۰	۱۴/۷۹۱	۰/۳۶۸	۲۲/۳۱۷
۷	۱۰۰	۰	۱۹/۸۷۱	۰/۴۲۱	۲۹/۳۴۱
۸	۱۲۰	۰	۲۳/۵۰۲	۰/۰۵۰	۳۵/۷۸۱
۹	۱۳۰	۰	۲۸/۳۷۱	۰/۲۰۰	۴۱/۹۱۲
۱۰	۱۵۰	۰	۳۴/۲۳۱	۰/۴۱۰	۴۹/۳۴۶

جدول (۳): کاهش هزینه‌ها با استفاده از رویکرد پیشنهادی نسبت به حالتی که سیاست انتخاب عمل حریصانه است

اجرا	تعداد شرکت‌کنندگان مناقصه	درصد کاهش هزینه تدارکات با استفاده از رویکرد پیشنهادی نسبت به حالتی‌که سیاست انتخاب عمل حریصانه است
۱	۱۵	٪۱۸
۲	۳۰	٪۲۳
۳	۴۵	٪۳۱
۴	۶۰	٪۲۸
۵	۷۵	٪۳۶
۶	۹۰	٪۲۹
۷	۱۰۰	٪۴۲
۸	۱۲۰	٪۳۸
۹	۱۳۰	٪۳۴
۱۰	۱۵۰	٪۳۹

۷- نتیجه‌گیری و پژوهش‌های آتی

در این مقاله یک مسئله مناقصه چند مرحله‌ای را که در آن خریدار قصد دارد برای چند دوره پیاپی از تعدادی تأمین‌کننده مشخص، تعدادی از اقلام مورد نیاز خود را تأمین نماید را به‌عنوان یک مسئله تصمیم‌گیری مارکوفی مدل‌سازی نمودیم. برای حل این مسئله تصمیم‌گیری مارکوفی که انتخاب عمل در هر مرحله آن به حالت و عمل در مرحله قبل بستگی دارد یک الگوریتم یادگیری Q توسعه داده شده است که سیاست انتخاب عمل در آن براساس الگوریتم ازدحام ذرات است. در مقایسه این الگوریتم با حالتی که در آن سیاست انتخاب عمل حریصانه دارد، این الگوریتم بسیار کارآمدتر است. پژوهش‌های آتی می‌تواند حالتی را در نظر بگیرد که مناقصه در یک سیستم چند عاملی انجام شود که در آن عامل‌های خریدار و تأمین‌کنندگان همگی قابلیت یادگیری دارند.

منابع

[1] Chopra, Sunil, and Peter Meindl. , ***"Supply chain management"***. *Strategy, planning & operation*. Gabler, 2007.

[2] Bichler,M. and Kalagnanam,J., ***"A nonoparametric estimator for setting: reserve prices in procurement auctions"***. ACM Conference on Electronic Commerce 2003: 254-255, 2003.

[3] Chen, S.L. and M.M. Tseng., ***"A Negotiation-Credit-Auction Mechanism for Procuring Customized Products"***. International Journal of Production Economics, 127(1): 203-210, 2010.

[4] Padgham, Lin, and Michael Winikoff., ***"Developing intelligent agent systems: A practical guide"***. Vol. 13. Wiley, 2005.

[5] Beam, C., and Segev, A., ***"Automated negotiations: A survey of the state of the art"***. Wirtschaftsinformatik 39(3), 263-268, 1997.

[6] Sutton, R.S., editor., ***"Reinforcement Learnng"***. Kluwer Academic Press, Boston, MA, 1992.

[7] Chaharsooghi, S. K., J. Heydari, et al., ***"A reinforcement learning model for supply chain ordering management: An application to the beer game"***. Decision Support Systems 45(4): 949-959, 2008.

[8] Li, X., Wang, J., & Sawhney, R., ***"Reinforcement learning for joint pricing, lead-time and scheduling decisions in make-to-order systems"***. European Journal of Operational Research, *221*(1), 99-109, 2012.

[9] Iima, H., Kuroe, Y., ***"Swarm Reinforcement Learning Algorithm Based on Particle Swarm Optimization Whose PersonalBests Have Lifespans"***. Neural Information Processing, Springer Berlin Heidelberg. 5864**:** 169-178, 2009.

[10] Giannoccaro, I. and P. Pontrandolfo., ***"Inventory management in supply chains: a reinforcement learning approach"***. International Journal of Production Economics 78(2): 153-161, 2002.

[11] Tang, H., Xu, L., Sun, J., Chen, Y., & Zhou, L., ***"Modeling and optimization control of a demand-driven, conveyor-serviced production station"***. European Journal of Operational Research, *243*(3), 839-851, 2015.

[12] Fu, J., & Fu, Y., ***"An adaptive multi-agent system for cost collaborative management in supply chains"***. Engineering Applications of Artificial Intelligence, *44*, 91-100, 2015.

[13] Mortazavi, A., Khamseh, A. A., & Azimi, P., ***"Designing of an intelligent self-adaptive model for supply chain ordering management system"***. Engineering Applications of Artificial Intelligence, *37*, 207-220, 2015.

[14] Russell,S and Peter Norvig., ***"Artficial Intelligence: A Modern Approach"***. Prentice-Hall, Saddle River, NJ, 1995.

[15] Sutton, R.S., Barto, A.G., ***"Reinforcement Learning"***. MIT Press, Cambridge,1998.

[16] Watkins, C. J. C. H., ***"Learning from delayed rewards"*** (Doctoral dissertation, University of Cambridge), 1989.

[17] Tsitsiklis, J. N., ***"Asynchronous stochastic approximation and Q-learning"***. Machine Learning, *16*(3), 185-202, 1994.

[18] Kennedy, J., Eberhart, R.C., ***"Swarm Intelligence"***. Morgan Kaufmann Publishers, San Francisco, 2001.

[19] Talbi,E., ***"Metaheuristics: From Design to Implementation"***. ISBN: 978-0-470-27858-1, 2009.

[20] Abdulhai, B., Pringle, R. and Karakoulas, G.J., ***"Reinforcement learning for ITS: Introduction and a case study on adaptive traffic signal control"***. Transportation Research Board 80th Annual Meeting, Washington, D.C, 2001.