

دسترسی در سایت <http://jnrm.srbiau.ac.ir>

سال یازدهم، شماره پنجاه و چهارم، خرداد و تیر ۱۴۰۴

شماره شاپا: ۵۸۸۸-۲۵۸۸



پژوهش‌های نوین در ریاضی



دانشگاه آزاد اسلامی، واحد علوم و تحقیقات

الگوریتمی برای تقریب جواب بهینه سراسری مسائل بهینه‌سازی براساس تکنیک خوشه‌بندی در داده‌کاوی

محمد علی نژاد مفرد^{۱*}، مهتاب حدادپور^۲، محمد دهقان نیری^۳

(^{۱و۲و۳}) گروه ریاضی، دانشکده علوم پایه، دانشگاه بجنورد، بجنورد، ایران

تاریخ ارسال مقاله: ۱۴۰۱/۱۱/۱۲ تاریخ پذیرش مقاله: ۱۴۰۲/۹/۲۱

چکیده

در برخی از مسائل بهینه‌سازی، یافتن جواب بهینه سراسری اهمیت فراوان دارد. اگر فضای جستجوی مسئله را بتوان به‌طور مداوم به‌گونه‌ای کوچک کرد که به احتمال زیاد شامل جواب بهینه سراسری باشد آنگاه می‌توان تقریب مناسبی برای جواب بهینه مسئله به دست آورد. در این مقاله برای کاهش فضای جستجو از تکنیک خوشه‌بندی در داده‌کاوی استفاده می‌شود. با استفاده تکراری از این تکنیک به همراه طرح یکنواخت از نوع نقاط مشبکه‌ای خوب برای تولید توزیع مناسبی از نقاط روی مرزها و درون فضای شدنی در هر تکرار، الگوریتمی تصادفی پیشنهاد می‌شود که قادر است تقریب مناسبی از جواب بهینه سراسری را بیابد. مقایسه نتایج حاصل از این الگوریتم با الگوریتم ژنتیک برای سه مسئله بهینه‌سازی نامقید غیرخطی نشان می‌دهد که الگوریتم پیشنهادی می‌تواند تقریب‌ها را با تعداد تکرار و زمان محاسباتی کمتری به دست دهد؛ همچنین اگر جواب بهینه سراسری روی مرزها واقع شده باشد، سرعت همگرایی آن در قیاس با الگوریتم ژنتیک به‌طور قابل توجهی بیشتر است.

واژه‌های کلیدی: بهینه‌سازی سراسری، جواب بهینه سراسری، داده‌کاوی، خوشه‌بندی، فضای جستجوی کاهش‌یافته.

۱- مقدمه

بهینه‌سازی سراسری، حوزه ارزشمندی در مباحث بهینه‌سازی است که کاربردهای متنوعی نیز دارد [۱]. از این رو دسته الگوریتم‌های مختلفی برای حل آن ارائه شده است. یکی از این دسته‌ها، الگوریتم‌های تصادفی هستند که در حالت کلی می‌توانند تقریب‌هایی از جواب بهینه سراسری بیابند [۲]. دسته دیگر، الگوریتم‌های قطعی‌اند که معمولاً جواب‌های بهینه محلی در فضای جستجو را به دست می‌دهند [۳]. با این حال هیچ کدام از این الگوریتم‌ها، در حالت کلی، موفق نیستند و در مواجهه با مسائل با چندین بهینه محلی^۲ ممکن است با گیرافتادن در یکی از آنها تقریب‌های خوب و حتی قابل قبول ارائه ندهند. درواقع تقریب جواب بهینه‌سازی سراسری، در حالت کلی، یک مسئله حل‌ناپذیر است؛ یعنی الگوریتمی که بتواند آن را در زمان متناهی حل کند وجود ندارد [۴]. با این وجود، به خاطر کاربردها، یافتن الگوریتم‌هایی که تقریب‌های قابل قبول از جواب مسائل بهینه‌سازی سراسری ارائه دهند ضرورت دارد. از این رو الگوریتم‌های ترکیبی که از مزیت‌های الگوریتم‌های قطعی و تصادفی به‌طور هم‌زمان استفاده می‌کنند ارائه شده‌اند تا کیفیت تقریب جواب بهینه سراسری را بهبود بخشند [۲، ۳]. اخیراً استفاده از تکنیک‌های داده‌کاوی نیز به دسته الگوریتم‌های ترکیبی اضافه شده که نتایج مطلوبی به دنبال داشته است [۵، ۶، ۷].

داده‌کاوی^۳ علم تجزیه و تحلیل داده‌ها و استخراج الگوها و اطلاعات از آنها است. در سال‌های اخیر با استفاده روزافزون از رایانه‌ها به‌طور چشمگیری، داده‌های انبوه در فایل‌های بسیار بزرگ، ذخیره‌سازی و ثبت شده است. از این رو، داده‌کاوی برای استخراج اطلاعات ارزشمند از این داده‌ها توسعه یافته است؛ به‌طوری که اکنون در بخش‌های تجاری، صنایع تولیدی، موسسات تحقیقاتی و بسیاری از بخش‌های دیگر، بدون استناد به پژوهش‌های داده‌محور و تجزیه و تحلیل آنها، تصمیم‌گیری نمی‌شود [۸].

داده‌کاوی تکنیک‌هایی را به کار می‌گیرد که طبقه‌بندی^۴، انجمنی^۵ و خوشه‌بندی^۶ از آن جمله‌اند. این تکنیک‌ها، داده‌ها را تجزیه و تحلیل کرده و قوانین ارزنده یا الگوهای پنهان در آنها را کشف می‌کنند. برای نمونه، تکنیک خوشه‌بندی، داده‌ها را به چند خوشه تقسیم می‌کند که در هر خوشه، داده‌ها رفتار مشابهی دارند [۸]. برای به‌دست آوردن الگوها و استخراج اطلاعات از داده‌ها، نرم‌افزارهایی برای پیاده‌سازی تکنیک‌های داده‌کاوی مورد نیاز است. یکی از نرم‌افزارهایی که در این زمینه نیز توسعه یافته است، نرم‌افزار محبوب و لفرام متمتیکا^۷ است که امکان بهره‌برداری از تکنیک‌های داده‌کاوی را فراهم می‌کند.

کاهش فضای جستجو^۸ به کمک تکنیک‌های داده‌کاوی و سپس ترکیب آن با سایر الگوریتم‌های قطعی و یا تصادفی، رویکرد جدیدی است که با جستجو در یک ناحیه کوچک‌تر، می‌تواند احتمال پیدا کردن جواب بهینه سراسری را بالا ببرد. در پژوهش [۵]، یک مسئله بهینه‌سازی غیرخطی با الگوریتم ژنتیک^۹ در ترکیب با سه تکنیک داده‌کاوی روی فضاها کاهش‌یافته حل شده است. نتایج نشان می‌دهند که این روش ترکیبی، تقریب‌های دقیق‌تری برای جواب بهینه سراسری مسئله نسبت به الگوریتم ژنتیک به تنهایی روی کل فضای شدنی به‌دست می‌دهد. در پژوهشی

^۲ Multimodal problems^۳ Data mining^۴ Classification^۵ Association^۶ Clustering^۷ Wolfram Mathematica^۸ Reducing search space^۹ Genetic Algorithm (GA)

دیگر، ترکیبی از سه تکنیک داده‌کاوی با روش برنامه‌ریزی درجه دو متوالی^{۱۰} (SQP) برای یافتن جواب بهینه سراسری یا جواب‌های بهینه محلی بهتر، استفاده شده است. نتایج عددی نشان داده‌اند که: جواب‌های بهینه‌ای که در فضاهای کاهش‌یافته با SQP به دست می‌آیند در مقایسه با جواب‌های حاصل از SQP به تنهایی روی کل فضای شدنی، بسیار بهترند؛ هزینه محاسباتی کاهش فضای جستجو با تکنیک‌های داده‌کاوی قابل چشم‌پوشی است و در عوض می‌تواند نتایج بسیار بهتری تولید کند؛ همچنین تکنیک خوشه‌بندی ساده‌ترین پیاده‌سازی و نسبتاً بهترین نتایج را در مقایسه با دو تکنیک طبقه‌بندی و انجمنی دارد [۶]. در تحقیق [۷] یک الگوریتم بهینه‌سازی ترکیبی پیشنهاد شده که در آن تکنیک‌های داده‌کاوی با استراتژی تکامل^{۱۱} و برنامه‌ریزی درجه دو متوالی^{۱۲} ترکیب شده است. نتایج این پژوهش نشان می‌دهد حتی زمانی که ترکیب یک الگوریتم تکاملی با الگوریتمی بر پایه گرادینان^{۱۳} - که اولی تصادفی و دومی قطعی است- برای حل مسائل با فضای جستجوی بزرگ به کار می‌رود، یافتن جواب بهینه سراسری هنوز می‌تواند مشکل باشد؛ اما ترکیب این دو الگوریتم با تکنیک‌های داده‌کاوی نتایج بهتری به دست می‌دهد.

در این پژوهش، الگوریتمی نوین ارائه می‌کنیم که از تکنیک خوشه‌بندی در داده‌کاوی به‌طور تکراری برای کاهش مرزها یا فضای جستجو استفاده می‌کند. برای انجام این کار لازم است که تعدادی نقاط نمونه برای مسئله به‌گونه‌ای تولید کنیم که توزیع آنها با کیفیت مناسبی روی مرزها و درون ناحیه شدنی گسترده شده باشند. روش‌های متنوعی برای تولید این دست از داده‌ها وجود دارد که طرح‌های یکنواخت^{۱۴} در طرح آزمایش‌ها^{۱۵} در علم آمار از مناسب‌ترین آنهاست؛ در اینجا طرح یکنواخت از نوع نقاط مشبکه‌ای خوب^{۱۶} (GLP) را برای تولید این نقاط استفاده می‌کنیم [۹-۱۲]؛ به این ترتیب که در هر تکرار، یکی از انتخاب‌های مختلف برای تولید نقاط نمونه از نوع GLP را به تصادف برمی‌گزینیم و به کمک آن، داده‌های مورد نیاز را تولید می‌کنیم. سپس به‌وسیله خوشه‌بندی مقادیر تابع هدف متناظر با نقاط نمونه تولیدشده در آن تکرار، فضای جستجو را کاهش می‌دهیم. نتایج حاصل از کاربرد این الگوریتم پیشنهادی روی سه مسئله بهینه‌سازی نامقید غیرخطی نشان می‌دهد که می‌توان تقریب‌های خوبی برای جواب بهینه سراسری این مسائل به دست آورد. شایان ذکر است که الگوریتم پیشنهادی همانند بسیاری از روش‌ها، از جمله الگوریتم ژنتیک، تضمینی برای همگرایی به جواب بهینه سراسری در تعداد تکرار متناهی ندارد، مگر آنکه از راه‌کارهایی برای حل‌پذیر شدن مسئله (تقریب جواب بهین سراسری در زمان متناهی) بهره‌برد. این راه‌کارها، به‌طور کلی، از اطلاعات سراسری درباره مسئله یا تابع هدف، مانند کران (بالا یا پایین برای) جواب مسئله، مقدار ثابت لیپ‌شیتس تابع هدف یا هنگ پیوستگی^{۱۶} تابع هدف، به منظور دستیابی به جستجویی که تمام ناحیه شدنی را پوشش دهد، استفاده می‌کنند؛ که این اطلاعات فقط برای رده‌های خاصی از مسائل فراهم است و در بسیاری از موارد هزینه محاسباتی اجرا آنقدر زیاد است که کاربرد آنها را توجیه‌ناپذیر می‌کند [۴].

در ادامه این نوشتار، نخست در بخش دو، مسئله بهینه‌سازی سراسری و مفاهیم مرتبط با آن را معرفی می‌کنیم. در بخش سه، تولید داده‌ها و نحوه کاهش فضای جستجو با تکنیک خوشه‌بندی را با یک مثال شرح می‌دهیم. سپس در بخش چهار، الگوریتم پیشنهادی را که تصادفی است و بر اساس کاربرد تکراری تکنیک خوشه‌بندی بنا شده است

^{۱۰} Sequential Quadratic Programming (SQP)

^{۱۱} Evolution Strategy

^{۱۲} Sequential Quadratic Programming

^{۱۳} Uniform designs

^{۱۴} Design of Experiments

^{۱۵} Good Lattice Points (GLP)

^{۱۶} Modulus of continuity

معرفی می‌کنیم. در بخش پنج، سه مسئله نامقید غیرخطی را معرفی کرده و الگوریتم پیشنهادی و الگوریتم ژنتیک را روی آنها پیاده‌سازی می‌کنیم تا نتایج حاصل را تجزیه و تحلیل کنیم. در پایان در بخش شش به نتیجه‌گیری و بحث خواهیم پرداخت.

۲- مسئله بهینه‌سازی سراسری

مسئله بهینه‌سازی نامقید را می‌توان به صورت زیر بیان کرد:

$$\min f(\mathbf{x}) \quad (1)$$

subject to $\mathbf{x} \in \Omega$,

که در آن $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$ بردار n -بعدی از متغیرهای مجهول، f یک تابع حقیقی مقدار از $\mathbf{x}$ و Ω زیرمجموعه‌ای از فضای n -بعدی E^n است. f را تابع هدف مسئله و هر نقطه $\mathbf{x}$ که عضو Ω باشد را شدنی می‌نامند [۱۳].

تعریف ۱-۲: نقطه شدنی $\mathbf{x}^* \in \Omega$ را کمینه محلی f روی Ω می‌نامند اگر برای تمامی $\mathbf{x} \in \Omega$ در یک ϵ -همسایگی از $\mathbf{x}^*$ داشته باشیم $f(\mathbf{x}) \geq f(\mathbf{x}^*)$.

تعریف ۲-۲: نقطه شدنی $\mathbf{x}^* \in \Omega$ را کمینه سراسری f روی Ω می‌نامند اگر برای تمامی $\mathbf{x} \in \Omega$ داشته باشیم $f(\mathbf{x}) \geq f(\mathbf{x}^*)$.

تعریف ۳-۲: هرگاه در مسئله (۱)، تابع هدف خطی باشد مسئله را خطی و در غیر این صورت غیرخطی می‌نامند. مسئله (۱) را مسئله بهینه‌سازی سراسری می‌نامند اگر هدف از حل آن، یافتن کمینه سراسری باشد. در عمل، چه از نظر تئوری و چه از نظر محاسباتی، یافتن جواب کمینه سراسری دشوار است [۱۳].

۳- کاهش فضای شدنی با تکنیک خوشه‌بندی

در این بخش نخست به چگونگی تولید داده‌های مورد نیاز برای تکنیک خوشه‌بندی می‌پردازیم. سپس با معرفی یک مسئله کمینه‌سازی، فضای شدنی آن را با خوشه‌بندی کوچک می‌کنیم.

۳-۱ تولید داده با طرح‌های یکنواخت

برای اینکه تکنیک خوشه‌بندی در داده‌کاوی برای کاهش فضای شدنی مسئله (۱) به کار رود، نخست باید مجموعه‌ای از داده‌های نمونه تولید شود. این داده‌های نمونه باید در فضای شدنی مسئله توزیع شده باشند. در اینجا از طرح یکنواختی استفاده می‌کنیم که داده‌های نمونه به دست آمده از آن، نقاط مشبکه‌ای خوب هستند. نقاط مشبکه‌ای خوب، یکنواختی^{۱۷} را برآورده می‌کنند (یعنی با افزایش تعداد، توزیع آنها به توزیع یکنواخت می‌گراید)؛ توزیع‌شان با توزیع یکنواخت ناجوری^{۱۸} کمی دارد (یعنی با تعداد کم نیز توزیع‌شان با توزیع یکنواخت مغایرت کمی دارد)؛ و نیز نظم هندسی دارند؛ افزون‌براین، می‌تواند پراکندگی‌شان^{۱۹} به گونه‌ای باشد که فضا پرکن^{۲۰} نیز باشند [۹، ۱۱، ۱۲].

^{۱۷} Uniformity

^{۱۸} Discrepancy

^{۱۹} Dispersion

^{۲۰} Space-Filling

برای ساختن طرح یکنواخت از نوع GLP، نخست باید آرایه‌ای یکنواخت^{۲۱} تعریف گردد. فرض کنید که می‌خواهیم ۱۰ نقطه نمونه درون ناحیه $[0,1] \times [0,1]$ از فضای شدنی یک مسئله بهینه‌سازی با تابع هدف دومتغیره تولید کنیم. از این‌رو اصطلاحاً یک آزمایش^{۲۲} با ۱۰ اجرا^{۲۳} و چهار فاکتور^{۲۴} در ناحیه $[0,1] \times [0,1]$ خواهیم داشت [۱۲]. آرایه یکنواخت حاصل با نماد U_4 نمایش داده می‌شود که عبارت است از

$$U_4 = \begin{bmatrix} 1 & 3 & 7 & 9 \\ 2 & 6 & 4 & 8 \\ 3 & 9 & 1 & 7 \\ 4 & 2 & 8 & 6 \\ 5 & 5 & 5 & 5 \\ 6 & 8 & 2 & 4 \\ 7 & 1 & 9 & 3 \\ 8 & 4 & 6 & 2 \\ 9 & 7 & 3 & 1 \\ 10 & 10 & 10 & 10 \end{bmatrix}$$

این آرایه ده سطر دارد که هر سطر معادل با یک اجرا است؛ همچنین چهار ستون دارد که هر ستون معادل با یک فاکتور است و ستون دوم تا چهارم درواقع جایگشت‌هایی از ستون اول هستند. پس از تعریف آرایه یکنواخت، از رابطه $X_{i,j} = (2u_{i,j} - 1) / (2N)$ می‌توان درایه‌های ماتریس طرح^{۲۵} متناظر با U_4 را محاسبه کرد که در آن $u_{i,j}$ درایه واقع در سطر i ام و ستون j ام آرایه یکنواخت و همچنین N تعداد اجرا است. بنابراین برای مثال بالا ماتریس طرح X_4 به دست می‌آید. چون تابع هدف دومتغیره است پس انتخاب دو فاکتور از چهار فاکتور (دو ستون از چهار ستون X_4) برای تولید نقاط نمونه از نوع GLP روی فضای شدنی دوبعدی کفایت می‌کند. توجه کنید که ستون اول حتماً باید انتخاب شود [۹]؛ از این‌رو یکی از سه ستون باقی‌مانده برای تولید ۱۰ نقطه استفاده می‌شود به عبارت دیگر با توجه به ماتریس X_4 سه انتخاب مختلف برای تولید نقاط نمونه از نوع GLP امکان‌پذیر است. توزیع نقاط نمونه متناظر با انتخاب ستون اول و دوم X_4 روی فضای شدنی $[0,1] \times [0,1]$ را در شکل (۱) مشاهده می‌کنید.

^{۲۱} Uniform array

^{۲۲} Experiment

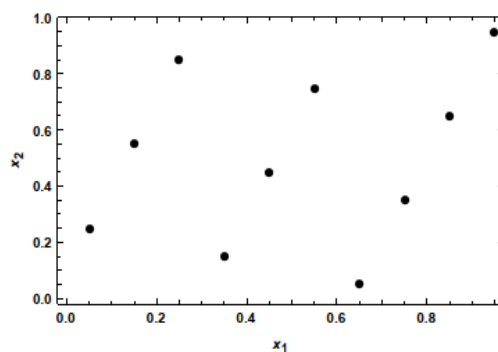
^{۲۳} Run

^{۲۴} Factor

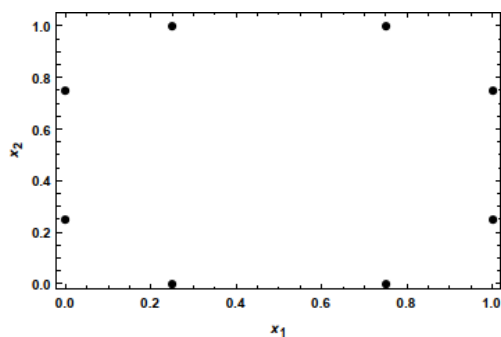
^{۲۵} Design Matrix

$$X_4 = \begin{bmatrix} 0.05 & 0.25 & 0.65 & 0.85 \\ 0.15 & 0.55 & 0.35 & 0.75 \\ 0.25 & 0.85 & 0.05 & 0.65 \\ 0.35 & 0.15 & 0.75 & 0.55 \\ 0.45 & 0.45 & 0.45 & 0.45 \\ 0.55 & 0.75 & 0.15 & 0.35 \\ 0.65 & 0.05 & 0.85 & 0.25 \\ 0.75 & 0.35 & 0.55 & 0.15 \\ 0.85 & 0.65 & 0.25 & 0.05 \\ 0.95 & 0.95 & 0.95 & 0.95 \end{bmatrix}$$

به همین ترتیب اگر بخواهیم روی مرزهای این ناحیه، نمونه‌ای از نوع GLP با تعداد هشت نقطه تولید کنیم، بعد از تعریف آرایه یکنواخت و محاسبه ماتریس طرح متناظر با آن، در این حالت نیز سه انتخاب متفاوت به دست می‌آید که توزیع نقاط نمونه متناظر با یکی از آنها در شکل (۲) به تصویر کشیده شده است.



شکل (۱). توزیع نقاط نمونه متناظر با انتخاب ستون اول و دوم X_4 درون ناحیه شدنی



شکل (۲). توزیع نقاط نمونه متناظر با انتخاب ستون اول و دوم X_4 روی مرز ناحیه شدنی

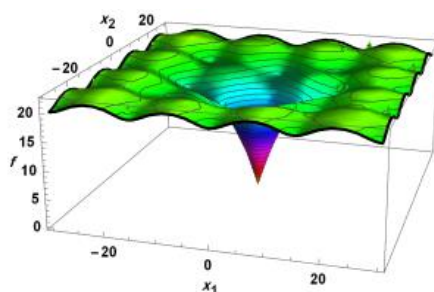
پس از اینکه نقاط نمونه به دست آمد، کافی است که مقادیر تابع هدف به ازای هر یک از این نقاط محاسبه شوند. سپس تمامی مقادیر تابع هدف به عنوان داده ورودی به نرم‌افزار متمتیکا سپرده می‌شوند تا با استفاده از تکنیک خوشه‌بندی، قوانین یا الگوهای دلخواه تعیین گردند. در پایان به کمک این الگوها، فضای شدنی یا مرزها کوچک می‌شوند.

۳-۲ خوشه‌بندی داده‌ها و کاهش فضای شدنی

در تکنیک خوشه‌بندی، خوشه‌بندی داده‌ها تنها با استفاده از مقادیر تابع هدف در نقاط نمونه انجام می‌شود. به این ترتیب که فاصله بین این مقادیر محاسبه می‌شود و بر اساس آن، خوشه‌های مورد نظر شکل می‌گیرند. پس از آن می‌توان تعداد اعضای هر خوشه، میانگین هر خوشه و هرگونه اطلاعات مورد نیاز از خوشه‌ها را به دست آورد. چون مسئله، کمینه‌سازی است از این رو بهترین خوشه، آن خوشه‌ای است که میانگین مقادیر تابع هدف کمتری دارد. مسئله زیر را در نظر بگیرید:

$$\begin{aligned} \min \quad & f(x_1, x_2) = -20 e^{-0.2\sqrt{0.5(x_1^2 + x_2^2)}} - e^{0.5(\cos(2\pi x_1) + \cos(2\pi x_2))} + 20 + e \\ \text{s. to} \quad & -30 \leq x_1 \leq 30 \\ & -30 \leq x_2 \leq 30 \end{aligned}$$

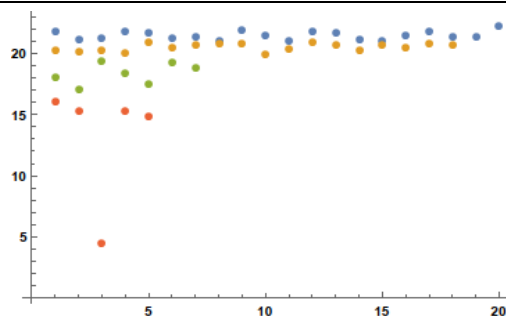
تابع هدف این مسئله تابع اکلی^{۲۶} است که روی فضای شدنی در شکل (۳) رسم شده است. آشکار است که مسئله تعداد بسیاری کمینه محلی در فضای شدنی دارد؛ همچنین تنها یک کمینه سراسری در نقطه $\mathbf{x}^* = (0, 0)$ وجود دارد که مقدار تابع هدف در این نقطه صفر است.



شکل (۳). تابع اکلی

با استفاده از همان روشی که در بخش ۱-۳ بیان شد، اکنون یک نمونه از نوع GLP با ۵۰ نقطه روی فضای شدنی $[-30, 30] \times [-30, 30]$ با انتخاب ستون اول و سوم ماتریس طرح تولید کرده و مقادیر تابع اکلی را در این نقاط محاسبه می‌کنیم. سپس از نرم‌افزار متمتیکا برای خوشه‌بندی مقادیر تابع هدف مسئله استفاده می‌کنیم. خروجی خوشه‌بندی در شکل (۴) نمایش داده شده است. چهار خوشه در این شکل مشخص شده‌اند که رنگ‌های متفاوتی دارند. بهترین خوشه به رنگ قرمز و شامل پنج عضو است که میانگین مقادیر تابع هدف در آن برابر 13.2266 است. حال از پنج عضو این خوشه برای تعیین کران‌های فضای کاهش‌یافته استفاده می‌کنیم.

^{۲۶} Ackley

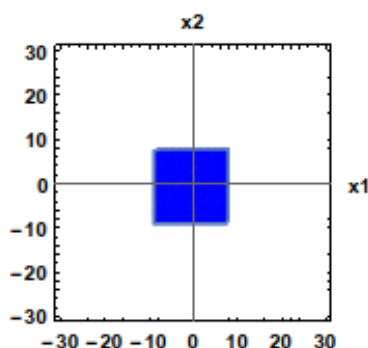


شکل (۴). خوشه‌های به دست آمده از اولین تکرار خوشه‌بندی برای ۵۰ داده نمونه از مسئله اکلی

چون کمترین و بیشترین مقدار x_1 در این خوشه برای پنج عضو آن به ترتیب برابر ۹- و ۷.۴ است پس $-9 \leq x_1 \leq 7.8$. به همین روی برای x_2 نیز داریم $-9 \leq x_2 \leq 7.8$. بنابراین مساحت فضای کاهش یافته در تکرار اول از تکنیک خوشه‌بندی حدود هشت درصد از کل فضای شدنی اولیه است. این فضای کاهش یافته نسبت به فضای شدنی مسئله در شکل (۵) نمایش داده شده است. همچنین از بررسی بهترین خوشه چنین برمی آید که کمینه مقادیر تابع هدف در این خوشه برابر ۴.۵۳۴۵۸ است و در نقطه $(-0.6, -0.6)$ از این خوشه اتفاق می افتد. اکنون برای تکرار دوم، کافی است که یک بار دیگر، تکنیک خوشه‌بندی را روی فضای کاهش یافته به دست آمده از تکرار اول با نقاط نمونه تولیدی جدید پیاده سازی کنیم.

۴- الگوریتم خوشه‌بندی تکراری^{۲۷} (ICA)

اکنون الگوریتمی را پیشنهاد می کنیم که تکنیک خوشه‌بندی در داده کاوی را به صورت تکراری به کار می گیرد تا جواب بهینه سراسری با دقت مطلوب به دست آید. معیار توقف الگوریتم، $|f_{k+1} - f_k| < \varepsilon(1 + |f_k|)$ است که در آن f_k مقدار کمینه تابع هدف در k امین تکرار است.



شکل (۵). فضای کاهش یافته حاصل از اولین تکرار خوشه‌بندی

^{۲۷} Iterative Clustering Algorithm (ICA)

گام مقداردهی اولیه. مقدار ϵ را انتخاب کن. تعداد ۴۸ نقطه نمونه با توزیع یکنواخت روی مرز ناحیه شدنی به‌طور تصادفی تولید و کمینه مقادیر تابع هدف ($\bar{f}$) را در این نقاط محاسبه کن. تعداد ۵۰ نقطه نمونه با توزیع یکنواخت درون ناحیه شدنی به‌طور تصادفی تولید و کمینه مقادیر تابع هدف ($\bar{f}^0$) را در این نقاط محاسبه کن. $k = 1$ و به گام ۱ برو.

گام ۱. اگر $\bar{f} \leq \bar{f}^0$ ، آنگاه $f_k = \bar{f}$ ؛ قسمتی از مرز را که شامل کمینه است انتخاب کن و به گام ۲ برو. درغیراین‌صورت $\bar{f}_k = \bar{f}^0$ ؛ تکنیک خوشه‌بندی را روی مقادیر تابع هدف متناظر با ۵۰ نقطه درون ناحیه شدنی از گام مقداردهی اولیه به کار بگیر و فضای شدنی را کاهش بده و به گام ۳ برو.

گام ۲. تعداد ۵۰ نقطه نمونه با توزیع یکنواخت روی مرز به‌طور تصادفی تولید و کمینه مقادیر تابع هدف (f_{k+1}) را در این نقاط محاسبه کن. اگر معیار توقف برآورده شد آنگاه پایان؛ درغیراین‌صورت تکنیک خوشه‌بندی را روی مقادیر تابع هدف شامل f_{k+1} به کار بگیر و مرز را کاهش بده سپس $k = k + 1$ و گام ۲ را با مرز کاهش‌یافته تکرار کن.

گام ۳. تعداد ۵۰ نقطه نمونه با توزیع یکنواخت درون ناحیه شدنی کاهش یافته به‌طور تصادفی تولید و کمینه مقادیر تابع هدف (f_{k+1}) را در این نقاط محاسبه کن. اگر معیار توقف برآورده شد آنگاه پایان؛ درغیراین‌صورت تکنیک خوشه‌بندی را درون ناحیه شدنی کاهش‌یافته شامل f_{k+1} اعمال کن و آن را کاهش بده سپس $k = k + 1$ و گام ۳ را تکرار کن.

بیان چند نکته درباره الگوریتم خوشه‌بندی تکراری: الف) به دلیل اهمیت مرزها در مسائل بهینه‌سازی و همچنین در نظر گرفتن مسائلی با ناحیه شدنی مستطیلی در این مقاله، باید مقادیر تابع هدف روی مرزها را نیز بررسی کنیم. از این‌رو باید روی هر یک از چهار پاره‌خط ناحیه شدنی، نمونه‌ای از نوع GLP با ۱۲ نقطه تولید شود تا به تعداد ۵۰ نقطه درون ناحیه نزدیک باشد و بنابراین روی مرز روی هم‌رفته ۴۸ نقطه تولید می‌کنیم (هرچند از این نظر، تولید ۵۲ نقطه نیز روی مرزها مناسب است). اما درون ناحیه شدنی در هر تکرار می‌توان تعداد ۵۰ نقطه تولید کرد. به هر روی، نقاط تولید شده در واقع از انتخاب ستون اول و انتخاب تصادفی یکی از سایر ستون‌های ماتریس طرح به دست می‌آید. ب) چون نتایج الگوریتم پیشنهادی با الگوریتم ژنتیک مقایسه خواهد شد از این‌رو در تکرارهای متوالی همچنان که فضای حاصل از تکنیک خوشه‌بندی کوچکتر می‌شود در انتخاب یکی از سایر ستون‌های ماتریس طرح به‌طور تصادفی عمل می‌کنیم تا به نوعی تنوع در نمونه‌ها نیز اتفاق بیفتد. پ) تعداد نقاط در هر تکرار عدد ۵۰ انتخاب شده است تا نتایج حاصل از الگوریتم پیشنهادی با نتایج حاصل از الگوریتم ژنتیک در نرم‌افزار متلب^{۲۸} که به طور پیش‌فرض ۵۰ نقطه را به عنوان جمعیت اولیه جواب در نظر می‌گیرد مقایسه شود. همچنین معیار توقف الگوریتم نیز همان معیار توقف پیش‌فرض الگوریتم ژنتیک است.

۵- نتایج پیاده‌سازی الگوریتم‌ها روی چند مسئله

در این بخش سه مسئله نامقید غیرخطی معرفی و در هر مورد نتایج حاصل از الگوریتم پیشنهادی با نتایج حاصل از الگوریتم ژنتیک هم از نظر دقت نتایج و هم از نظر تعداد تکرار و زمان محاسبات مقایسه می‌شود. برای پیاده‌سازی الگوریتم ژنتیک از نرم‌افزار متلب نسخه ۲۰۱۴a و برای تولید نقاط نمونه و پیاده‌سازی الگوریتم پیشنهادی از نرم‌افزار ممتیکا نسخه ۱۰ استفاده شده است. همچنین مشخصات سخت‌افزاری رایانه استفاده شده برای اجرای الگوریتم‌ها عبارت است از:

Intel(R) Pentium(R) CPU G۸۴۰ @ ۲,۸۰ GHz ۲,۸۰ GHz with ۴ GB of RAM.

برای معرفی سه مسئله نامقید غیرخطی از پژوهش [۱۴] کمک می‌گیریم. در این پژوهش، ۲۰ تابع تست^{۲۹} برای تابع هدف، ارائه و به سه دسته دارای یک بهینه محلی^{۳۰}، دارای تعداد کمی بهینه محلی^{۳۱} و دارای تعداد زیادی بهینه محلی^{۳۲} طبقه‌بندی شده‌اند؛ از این رو در ادامه، تابع هدف مسائل را از دو دسته آخر که پیچیده‌ترند برمی‌گزینیم:

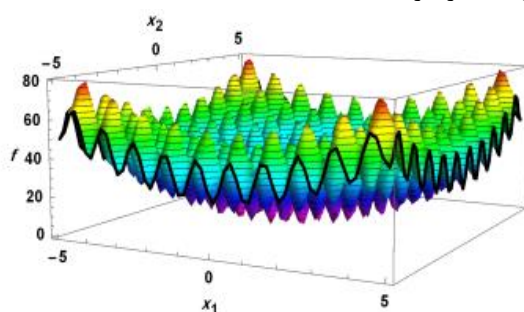
الف) مسئله کمینه‌سازی تابع هدف بسیار نویزی راستریگین^{۳۳}:

$$\min f(x_1, x_2) = 20 + x_1^2 + x_2^2 - 10 (\cos(2\pi x_1) + \cos(2\pi x_2))$$

$$s. to \quad -5 \leq x_1 \leq 5$$

$$-5 \leq x_2 \leq 5$$

نمودار این تابع روی فضای شدنی را در شکل (۶) مشاهده می‌کنید. این مسئله تنها یک کمینه سراسری در نقطه $\mathbf{x}^* = (0, 0)$ با مقدار تابع هدف $f(\mathbf{x}^*) = 0$ و تعداد بسیار زیادی کمینه محلی در فضای شدنی دارد که برخی از آنها بسیار نزدیک به جواب بهینه سراسری مسئله هستند.



شکل (۶). تابع راستریگین

ب) مسئله کمینه‌سازی با دامنه نامتقارن:

$$\min f(x_1, x_2) = (x_2 - \frac{5.1}{4\pi^2} x_1^2 + \frac{5}{\pi} x_1 - 6)^2 + 10(1 - \frac{1}{8\pi} \cos(x_1)) + 10$$

$$s. to \quad -5 \leq x_1 \leq 10$$

$$0 \leq x_2 \leq 15$$

^{۲۹} Test Function

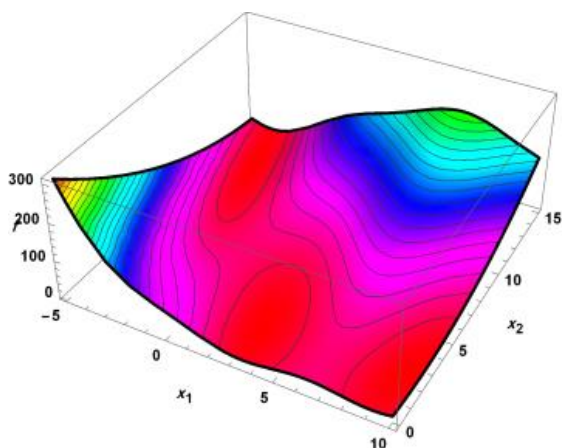
^{۳۰} Unimodal functions

^{۳۱} Multimodal functions with a few local optima

^{۳۲} Multimodal functions with many local optima

^{۳۳} Rastrigin function

نمودار این تابع روی فضای شدنی در شکل (۷) رسم شده است. این مسئله تنها یک کمینه سراسری در نقطه $\mathbf{x}^* = (3.14159, 2.275)$ با مقدار تابع هدف $f(\mathbf{x}^*) = 0.397887$ و تعدادی کمینه محلی در فضای شدنی دارد.



شکل (۷). تابع هدف مسئله ب

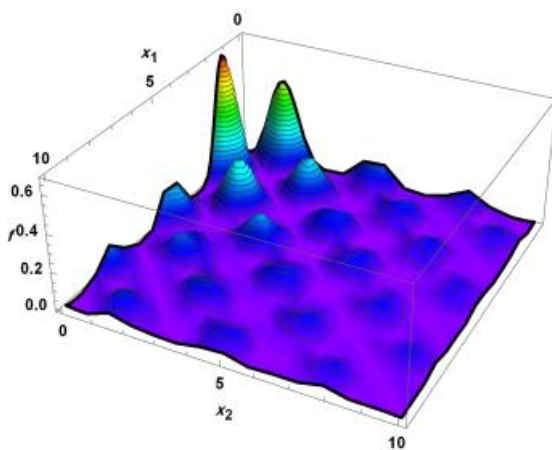
پ) مسئله بیشینه‌سازی تابع هدف بامپی^{۳۴}:

$$\max f(x_1, x_2) = |(\cos^4(x_1) + \cos^4(x_2) - 2\cos^2(x_1)\cos^2(x_2)) / \sqrt{x_1^2 + 2x_2^2}|$$

$$s. to \quad 0 \leq x_1 \leq 10$$

$$0 \leq x_2 \leq 10$$

نمودار این تابع روی فضای شدنی در شکل (۸) نمایش داده شده است. این مسئله تنها یک بیشینه سراسری در نقطه $\mathbf{x}^* = (1.39325, 0)$ با مقدار تابع هدف $f(\mathbf{x}^*) = 0.673668$ و تعداد بسیاری بیشینه محلی در فضای شدنی دارد که برخی از آنها بسیار نزدیک به جواب بهینه سراسری مسئله هستند.



شکل (۸). تابع بامپی

^{۳۴} Bumpy

اکنون با استفاده از الگوریتم پیشنهادی (ICA) و الگوریتم ژنتیک (GA) این سه مسئله را حل می‌کنیم. مقدار ε در معیار توقف الگوریتم را $\varepsilon = 10^{-6}$ در نظر می‌گیریم. بهترین نتیجه حاصل از ده مرتبه اجرای دو الگوریتم روی مسائل الف)، ب) و پ) در جدول (۱) گردآوری شده است.

جدول ۱: نتایج محاسباتی حاصل از پیاده‌سازی GA و ICA برای سه مسئله بهینه‌سازی سراسری

مسئله	$\mathbf{x}^*$	$f(\mathbf{x}^*)$	الگوریتم	$\mathbf{x}^k$	$f(\mathbf{x}^k)$	تکرار (k)	زمان محاسبات (s)
الف	(۰,۰)	۰	GA	$(۲,۲۲,۱,۵۳) \times 10^{-6}$	$۱,۴۴ \times 10^{-9}$	۷۳	۰,۳۴
			ICA	$(-۱,۶۱,-۹,۰۶) \times 10^{-6}$	$۱,۶۸ \times 10^{-8}$	۲۶	۰,۳۱
ب	(۳,۱۴۱۵۹,۲,۲۷۵)	۰,۳۹۷۸۸۷	GA	(۳,۱۴۱۵۸,۲,۲۷۴۹۹)	۰,۳۹۷۸۸۷	۷۱	۰,۳۳
			ICA	(۳,۱۴۱۵۱,۲,۲۷۵۱۲)	۰,۳۹۷۸۸۷	۱۶	۰,۲۲
پ	(۱,۳۹۳۲۵,۰)	۰,۶۷۳۶۶۸	GA	(۱,۳۹۳۲۳,۰)	۰,۶۷۳۶۶۸	۶۰	۰,۴۰
			ICA	(۱,۳۹۳۲۱,۰)	۰,۶۷۳۶۶۶	۶	۰,۰۵

آشکار است که در تمامی مسائل، تعداد تکرارها و زمان محاسباتی الگوریتم خوشه‌بندی تصادفی تکراری نسبت به الگوریتم ژنتیک کمتر است اما الگوریتم ژنتیک، مقادیر نسبتاً دقیق‌تری را به دست داده است؛ همچنین با در نظر گرفتن مسئله پ) می‌توان دریافت که اگر جواب بهینه سراسری روی مرزها واقع شده باشد، الگوریتم پیشنهادی می‌تواند بسیار سریع‌تر به جواب همگرا شود.

۶- نتیجه‌گیری و بحث

با توجه به اهمیت یافتن جواب بهینه سراسری در بعضی از مسائل بهینه‌سازی، در این پژوهش الگوریتمی تصادفی پیشنهاد شد که قادر است تقریب‌های مناسبی از جواب بهینه سراسری را به دست دهد. این الگوریتم بر اساس استفاده تکراری از تکنیک خوشه‌بندی در داده‌کاوی ساخته شده است که برای تولید داده‌های مناسب روی مرزها و درون ناحیه شدنی از طرح یکنواخت از نوع GLP بهره می‌گیرد. نتایج نشان می‌دهند که این الگوریتم در قیاس با الگوریتم ژنتیک نتایجی با دقت مناسب و تعداد تکرار و زمان محاسباتی کمتری ارائه می‌دهد. همچنین برای مسائلی که کمینه آنها روی مرز اتفاق می‌افتد می‌تواند بسیار سریع‌تر به دقت مطلوب دست یابد.

اگر برای تولید نقاط نمونه، نخست بتوان تمامی انتخاب‌های GLP از طرح یکنواخت را بررسی کرد آنگاه می‌توان یکی از آنها را که نمونه‌اش، پراکندگی (فضا پُرکنی) بیشتری نسبت به بقیه انتخاب‌ها در ناحیه شدنی دارد برگزید و تمامی تکرارهای الگوریتم را با آن انتخاب به انجام رسانید؛ در این حالت، درجه تصادفی بودن الگوریتم کمتر خواهد بود. از آنجایی که تعداد نقاط نمونه در ناجوری و پراکندگی نمونه و بنابراین در کیفیت نمونه تاثیر دارد، اثر تعداد این نقاط در نتایج الگوریتم پیشنهادی می‌تواند قابل توجه باشد و باید مدنظر قرار گیرد. به نظر می‌رسد که می‌توان مسائل مقید را نیز با الگوریتم پیشنهادی تجزیه و تحلیل کرد؛ هرچند در این نوع مسائل، یافتن توزیع مناسبی از نقاط روی انواع مرزها همچنین درون ناحیه شدنی متناظر، خود مسئله دیگری است که باید برای آنها چاره‌ای اندیشید. همگرایی الگوریتم پیشنهادی نیز به عواملی همچون انتخاب تصادفی نقاط و خوشه‌های به دست آمده در هر تکرار وابسته

است. از این رو چون الگوریتم تصادفی است – همان گونه که در مقدمه بیان شد –، مانند سایر الگوریتم‌های تصادفی، تضمینی برای همگرایی و یافتن جواب بهینه سراسری دقیق در هر بار اجرای الگوریتم وجود ندارد. اگر به طریقی بتوان در هر تکرار الگوریتم، ناحیه شامل جواب بهینه سراسری مسئله را حفظ کرد در این صورت می‌توان احتمال یافتن جواب بهینه سراسری را باز هم افزایش داد. این موارد از اهداف پژوهشی آتی نویسندگان این مقاله است.

فهرست منابع

- [۱] R. Horst, and P. M. Pardalos. Handbook of Global Optimization. Springer (۱۹۹۵)
- [۲] A. Zhigljavsky, A. Zilinskas. Stochastic global optimization. Springer (۲۰۰۸)
- [۳] R. Horst, H. Tuy. Global Optimization (Deterministic Approaches). Springer (۱۹۹۶)
- [۴] A. Torn, A. Zilinskas. Global optimization. Springer (۱۹۸۹)
- [۵] M. Haddadpour, M. Alinejadmofrad, M. Dehghan Nayyeri. Reducing search space in optimization problems with data mining techniques. Towards Mathematical Sciences ۲ (۱): ۱۱۰-۱۲۱ (۲۰۲۲). Doi: ۱۰.۲۲۰۶۷/tmsj.۲۰۲۲.۴۲۸۲۴
- [۶] T. Y. Chen, J. H. Huang. An efficient and practical approach to obtain a better optimum solution for structural optimization. Engineering Optimization ۴۵ (۸): ۱۰۰۵-۲۶ (۲۰۱۳)
- [۷] T. Y. Chen, J. H. Huang. Application of data mining in a global optimization algorithm. Advances in Engineering Software ۶۶: ۲۴-۳۳ (۲۰۱۳)
- [۸] J. Han, M. Kamber, J. Pei. Data Mining: Concepts and Techniques, ۳rd ed.. Morgan Kaufmann (۲۰۱۱)
- [۹] K. T. Fang, Y. Wang. Number-Theoretic Methods in Statistics. Chapman and Hall. London (۱۹۹۴)
- [۱۰] K. T. Fang, Y. Wang, P. M. Bentler. Some applications of number-theoretic methods in statistics. Statistical Science ۹ (۳): ۴۱۶-۲۸ (۱۹۹۴)
- [۱۱] K. T. Fang, D. K. J. Lin. Uniform Experimental Designs and Their Application in Industry. Vol. ۲۲. Handbook of Statistics. Elsevier Science (۲۰۰۳)
- [۱۲] W. Hopkins. Uniform Space-Filling Designs and Their Applications in Designed Experiments. M.Sc. thesis, Department of Mathematical Sciences. Montana State University (۲۰۱۸)
- [۱۳] J. Nocedal, and S. Wright. Numerical Optimization, ۲nd.ed.. Springer (۲۰۰۶)
- [۱۴] Z. Tu, Y. Lu. A robust stochastic genetic algorithm (StGA) for global numerical optimization. IEEE Transactions on Evolutionary Computation ۸ (۵): ۴۵۶-۷۰ (۲۰۰۴)