نشـــریه بین المللـــی مهنـــدسی صنـــایع و مدیـــریت تولیـــد

شمـــاره ۱، جلـــد ۲۳، خـــرداد ۱۳۹۱

صفحـــه ۱۲۸-۱۲۲

http://IJIEPM.iust.ac.ir/

ISSN: 2008-4870

نشریه بین المللی مهندسی صنایع و مدیریت تولید دانشگاه علم و صنعت ایران



یک روش ترکیبی خوشه بندی مبتنی بر الگوریتم ژنتیک با استفاده از عملگر های جدید تغییر

مسعود یقینی*، رویا سلطانیان و جواد نوری

کلمات کلیدی

خوشه بندی،
الگوریتم ژنتیک،
الگوریتم *K*-Means.

چکیده:

مساله خوشه بندی به منظور کمینه کردن مجموع مجذور انحراف، یک مساله غیر خطی و غیـر محـدب بوده و دارای تعداد زیادی نقاط بهینه محلی است. هدف از این مقاله، ارائه روشی ترکیبی بـا اسـتفاده از الگوریتم ژنتیک و *K*-Means برای خروج از نقاط بهینه محلی است.

استفاده از الگوریتم ژنتیک برای خروج از نقاط بهینه محلی، توسط محققین بسیاری انجام شده است. در این مقاله روش های جدیدی برای عملگرهای بازترکیبی و جهش ارائه شده است. منطق روش های پیشنهادی بر این امر استوار است که اگر عملگرهای تغییر به جای آنکه بطور تصادفی در کل فضای جواب اعمال گردند، در یک منطقه محدود از پیش تعریف شده، انجام شوند، به جواب های بهتری دست خواهیم یافت.

برای ارزیابی الگوریتم پیشنهادی، از سه نوع عملگر جهش و پنج نوع عملگر بازترکیبی بر روی مجموعه داده‌های استاندارد استفاده شده است. مقایسه نتایج بدست آمده با سایر روش ها، به ازای *K*های متفاوت، نشان می‌دهد می‌توان با استفاده از عملگر بازترکیبی ساده یک نقطه ای و عملگر جهش ارائه شده در این مقاله با نام "عملگر جهش منطقه ای خوشه ای"، به جواب های بهتری دست یافت.

۱. مقدمه

خوشه بندی در فضای n بعدی اقلیدسی، فرایند تقسیم بندی یک مجموعه داده یا نمونه به تعداد K گروه یا خوشه براساس شـباهت یا عدم شباهت آنها است. در واقـع نمونـه هـایی کـه در یـک زیـر مجموعه قرار می گیرند، بهم شـبیه انـد و بـا آنهـایی کـه در زیـر مجموعه دیگری قرار می گیرند متفاوت و غیـر شـبیه هسـتند. در بعضی از مسائل خوشه بندی، تعداد K یا خوشه ها از پیش تعیین شده است.

یکی از رایج ترین روش های افرازبندی[۲] برای خوشه بندی، الگوریتم *K-Means* است [۱]. مشکل *K-Means* این است که در نقاط بهینه محلی گیر می افتد. جواب های بدست آمده از این الگوریتم به نقاط یا مراکز خوشه اولیه انتخاب شده، وابسته است. یعنی اگر مراکز مناسبی برای خوشه های اولیه انتخاب شده باشد، خوشه بندی خوبی بدست می آید. در الگوریتم *K*-Means تعداد خوشه ها یا K از پیش تعریف شده است. در این الگوریتم ابتدا تعداد K نمونه[۳] بعنوان مراکز خوشه[۴] بصورت تصادفی انتخاب می شود. سپس سایر نمونه ها براساس کمترین فاصله (معمولا فاصله

تاریخ وصول: ۸۹/۷/۹

تاریخ تصویب: ۹۰/۳/۲۱

*نویسنده مسئول مقاله: دکتر مسعود یقینی، استادیار دانشکده مهندسی راه آهن، دانشگاه علم و صنعت ایران، yaghini@iust.ac.ir

رویا سلطانیان. دانشجوی کارشناسی ارشد مدیریت اجرایی، دانشگاه علم و صنعت ایران، roya_soltanian@yahoo.com

جواد نوری دانشجوی مهندسی کامپیوتر، دانشگاه علم و صنعت ایران، javad.nuri@gmail.com

[2] . *Partitioning*
[3] . *Object*
[3] . *Mean value*

Archive of SID

www.SID.ir

اقلیدسی) با مرکز خوشه ها در یکی از خوشه ها قرار می گیرند. به این ترتیب K خوشه خواهیم داشت که هر یک حاوی تعدادی نمونه است. سپس میانگین نمونه های هر خوشه را محاسبه کرده و بعنوان مراکز جدید آن خوشه در نظر می گیریم و براساس آن مرکز خوشه جدید، مجدداً نمونه ها خوشه بندی می شوند، بدین ترتیب خوشه های جدید با نمونه های جدید خواهیم داشت. عموماً برای خاتمه الگوریتم از رابطه (۱-۱) استفاده می شود. تلاش الگوریتم برآنست که میزان این انحراف را کمینه کند. در صورتی که در مقدار SSE[1] بهبودی حاصل نشود، الگوریتم متوقف می شود.

$$SSE = \sum_{i=1}^{k} \sum_{p \in C_i} |P - m_i|^2 \qquad (۱)$$

در رابطه (۱-۱) P، نقطه ای است که نشان دهنده یک نمونه است ، m_i، مرکز خوشه c_i و SSE مجموع مجذور انحراف برای همه نمونه ها است.

در این مقاله سعی شده با کمک الگوریتم ژنتیک به یک خوشه بندی بهینه دست یابیم. بدین ترتیب که با عملگر ساده بازترکیبی یک نقطه ایی و با استفاده از یک عملگر جهش جدید ارائه شده بنام "جهش منطقه ای خوشه ای " به جواب های بهتری دست یافتیم. در این روش، مرکز خوشه مورد نظر برای اعمال جهش، بجای آنکه بطور تصادفی با یک نقطه در کل فضای جواب جابجا شود، با نقطه ای در یک فضای محدود انتخاب شده، جابجا می‌شود. برای ارزیابی عملکرد الگوریتم پیشنهادی از مجموعه داده‌های استاندارد استفاده شده است. مقایسه نتایج بدست آمده با سایر روش ها، به ازای K های متفاوت، نشان می دهد با استفاده از روش پیشنهادی برای عملگر جهش، می توان به جواب‌های بهتری دست یافت.

در بخش دوم مروری بر مطالعات انجام شده‌ی مرتبط با موضوع مقاله ارائه می شود. در بخش سوم الگوریتم پیشنهادی مورد بررسی قرار گرفته و در بخش چهارم نتایج بدست آمده از الگوریتم پیشنهادی برروی داده های استاندارد ارائه می گردد و در بخش آخر به جمع بندی و نتیجه گیری پرداخته می شود.

۲. تحقیقات انجام شده مرتبط

در دو دهه اخیر تحقیقات بسیاری بر مبنای الگوریتم ژنتیک انجام شده است تا بتوان احتمال پیدا کردن نقاط بهینه[2] سراسری را بالا برد. در سال ۱۹۹۳ بابو و مرتی[3] [۲] الگوریتم ژنتیکی برای انتخاب مراکز خوشه اولیه در الگوریتم *K-Means* ارائه کردند که در آن از نمایش رشته بیتی[4] استفاده می شد. در این الگوریتم از یک عملگر ساده بازترکیبی برای جابجایی مراکز خوشه والدین استفاده می‌شد، بطوریکه این جابجایی بصورت کاملاً تصادفی در فضای مجموعه داده‌ها انجام می شد.

درسال ۱۹۹۹ کریشنا و مورتی[5] [۳] یک الگوریتم ژنتیک ترکیبی ابداع کردند که جواب های نزدیک به بهینه سراسری را برای مساله خوشه بندی با تعداد خوشه های مشخص بدست می داد. در این شکل ابداعی از الگوریتم ژنتیک، از الگوریتم کلاسیک شیب نزولی[6] در فرآیند خوشه بندی استفاده شده بود. در سال ۲۰۰۰ ماولیک و همکاران[7] [۴] یک روش خوشه بندی مبتنی بر بر الگوریتم ژنتیک پیشنهاد کردند که در آن از قابلیت جستجوی الگوریتم ژنتیک به منظور تعیین K مرکز خوشه در فضای R^N بهره گرفته می شد.

در این روش مقدار K از پیش معلوم در نظر گرفته شده است و یکی از مقادیر ورودی مساله است، در واقع تعداد مراکز خوشه تعداد نمونه ها در پایگاه داده است. معیاری که برای سنجش فاصله بین نقاط از آن استفاده می شود، فاصله اقلیدسی نقاط با مرکز خوشه متناظرشان است. کروموزوم ها که به صورت رشته هایی از اعداد حقیقی هستند، مراکز خوشه ها را در خود ذخیره می کنند. در سال ۲۰۰۱ هانسن و ملادوونیک[8] الگوریتم ژنتیکی بنام *J-Means* [۵] ارائه دادند که درآن روش جدیدی برای جستجوی منطقه‌ای ارائه شده بود.

در این روش پس از انتخاب یک مرکز خوشه، یکایک مراکز خوشه اطراف آن به ترتیب با مرکز خوشه انتخابی جابجا می شوند تا مرکز خوشه ی بهتر انتخاب شود. در سال ۲۰۰۲ ماولیک و همکاران الگوریتم ژنتیکی [۶] ارائه دادند که در آن به جای آنکه هر کروموزوم شامل اعضای هر خوشه باشد تنها مراکز خوشه نگهداری می شوند.

این امر کمک می کند زمان فرایند الگوریتم کوتاه شود و الگوریتم برای پایگاه های داده بزرگ بهتر جوابگو باشد. در سال ۲۰۰۶ لازلو و موهاراجه[9] [۷] الگوریتم ژنتیکی برای انتخاب مراکز خوشه خوشه ارائه دادند. این مراکز خوشه به کمک درخت هایپر کواد[10] که یک زیر بخش[11] فضایی منطقه ای است، تعیین می شوند.

---

1. *Sum of Squared Euclidean distance*
2. *Global Optimal*
3. *Babu and Murty*
4. *Bit- string representation*
5. *K.Krishna and M.Narasimha Murty*
6. *Gradient Descent Algorithm*
7. *S.Bandyopadhyay, U.Maulik*
8. *P. Hansen and N. Mladenovic*
9. *M .Laszlo and S. Mukherjee*
10. *Hyper-Quad Tree*
11. *Subdivision*

www.SID.ir
Archive of SID

اشکال این روش آن بود که تنها برای پایگاه های داده کوچک با تعداد K کوچک جوابگو بود. به همین دلیل در سال ۲۰۰۷ آنها الگوریتم ژنتیک دیگری [۸] برای انتخاب مراکز خوشه ارائه دادند که برای پایگاه های داده بزرگ نیز جوابگو بود. در این روش از نوعی عملگر بازترکیبی جدید استفاده می شود که به آن عملگر بازترکیبی منطقه ای[1] می گویند.

این عملگر بازترکیبی ابداعی، مراکز خوشه هایی را جابجا می کند که در یک منطقه از فضای پایگاه داده واقع شده باشند و یا بعبارت دیگر مراکز خوشه هایی را جابجا می کند که در همسایگی هم قرار دارند. در این روش باور بر آنست که اگر بتوانیم مجموعه مراکز خوشه اشغال شده در یک فضای یکسان یا نزدیک به هم را بصورت جوابهای با کیفیت بالا درآوریم، تاحدی مشکل مراکز خوشه حل می شود. نتایج، نشان داده است که این روش علاوه بر آنکه به جواب های بهتری منجر می شود، زمان اجرای الگوریتم را نیز کاهش می دهد. در سال ۲۰۰۸ چانگ زانگ و زنگ[2] [۹] الگوریتم ژنتیکی ارائه دادند که در آن از نوع خاصی عملگر بازترکیبی استفاده می شد.

عملگر بازترکیبی مورد استفاده، عملگر مبتنی بر مسیر[3] است که که در آن بین ژن های والدین، مسیری مشخص می شود و ژن هایی که روی این مسیر قرار دارند با ژن های والدین جابجا می شوند تا فرزندان یا نسل جدید حاصل شوند. همچنین در سال ۲۰۰۸ الگوریتم خوشه بندی تحت عنوان K-Means توسط زالیک[4] [۱۰] ارائه شد. ایده اصلی در این مقاله کمینه کردن تابع تابع هزینه است.

در سال ۲۰۱۰ جینگ و همکارانش[5] [۱۱] برای برطرف کردن مشکلات الگوریتم *K-Means* الگوریتم ژنتیکی کوانتومی[6] ارائه دادند. همچنین در سال ۲۰۱۰ یو و وانگ[7] [۱۲] الگوریتمی تحت تحت عنوان QBCA[8] ارائه نمودند که در آن به کمک توابع ریاضی، داده ها به تعدادی هیستوگرام تخصیص داده می شوند.

بررسی های انجام شده بر روی مطالعاتی که تاکنون انجام شده است، نشان می دهد اکثر این مطالعات بر روی ارائه عملگرهای بازترکیبی جدید برای پیدا کردن نقاط اولیه در الگوریتم K-Means تمرکز دارند. اما در مقاله پیش رو نشان داده می شود که با استفاده از عملگرهای جهش مناسب نیز، می توان به جواب های بهتر دست یافت.

[1]. *Regional Crossover*
[2]. *D-X.Chang, X-D .Zhang, C-W .Zheng*
[3]. *Path-based Crossover*
[4]. *Krista Rizman Zalik*
[5]. *Jing Xiao, YuPing Yan, Jun Zhang and Yong Tang*
[6]. *The quantum-inspired genetic algorithm*
[7]. *Zhiwen Yu and Hau-San Wong*
[8]. *Quantization -Based Clustering Algorithm (QBCA)*

۳. الگوریتم پیشنهادی

در روش پیشنهادی، از الگوریتم ژنتیک [۱۳]، بعنوان یک الگوریتم فراابتکاری برای پیدا کردن مراکز خوشه ها استفاده شده است. در این بخش به تشریح اجزای الگوریتم پیشنهادی می پردازیم.

۳-۱. جمعیت اولیه

الگوریتم پیشنهادی با مجموعه ای از کروموزوم ها که هر کدام نشان دهندهٔ یک جواب برای مساله مفروض هستند، تحت عنوان جمعیت اولیه شروع به کار می کند. نکته مهم در تولید جمعیت اولیه در روش پیشنهادی این است که هر کروموزوم بصورت یک رشته بطول m خواهد بود که در آن m برابر با تعداد مراکز خوشه است. بدین ترتیب طول کروموزومها حتی در پایگاه‌های داده بزرگ خیلی بزرگ نخواهد بود و زمان اجرای الگوریتم[9] وقت گیر و طولانی نمی شود.

۳-۲. عملگر انتخاب

به منظور گزینش کروموزومهای والدین و وارد نمودن آنها به مرحله تولید مثل[10] و تولید کروموزومهای جدید از عملگر انتخاب کمک گرفته می شود. در این روش دو کروموزوم انتخاب می شود که این انتخاب برای ورود به مرحله تولید مثل به میزان شایستگی[11] کروموزوم ها بستگی دارد. شایستگی یک کروموزوم رابطه مستقیم با مقدار SSE آن دارد. اگر از SSE به مجموع مجذور فواصل اقلیدسی هر نقطه تا مرکز خوشه تعبیرنماییم، هرچه این فاصله کمتر یا کوچکتر باشد، شایستگی آن کروموزوم یا احتمال انتخاب آن کروموزوم بیشتر می شود.

بنابراین در هر نسل، کروموزومی که شایستگی بیشتری دارد شانس بیشتری برای ورود به مرحله تولید مثل خواهد داشت. بعبارتی دیگر در هر نسل با توجه به احتمال انتخاب هر کروموزوم، یک جفت کروموزوم انتخاب می شوند و با ورود به مرحله تولید مثل تحت عملگرهای باز ترکیبی و جهش قرارمی گیرند. در این مقاله عملگر انتخاب به کمک روش چرخ رولت[12] انجام شده است. برای برطرف کردن مشکلات این روش ابتدا شایستگی ها مقیاس بندی[13] شدند. این مقیاس بندی نیز به دو روش خطی و سیگما انجام شد تا از صحت مقیاس بندی اطمینان خاطر حاصل شود.

[9]. *Run time*
[10]. *Mating Pool*
[11]. *Fitness Value*
[12]. *Wheel Roulette*
[13]. *Scaling*

www.SID.ir
Archive of SID

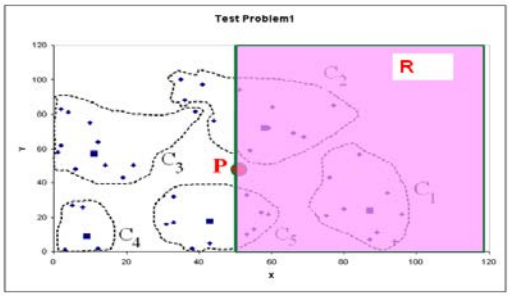
۳-۳. عملگر بازترکیبی

یکی از مهمترین عملگرهایی که در الگوریتم ژنتیک مورد استفاده قرار می گیرد، عملگر بازترکیبی است که در یافتن و بهبود جواب‌های حاصل شده بسیار موثر است. معمولاً برای اعمال عملگر بازترکیبی احتمالی مانند P_c یا نرخ بازترکیبی در نظر گرفته می شود که با آن کروموزوم انتخابی تحت عملگر بازترکیبی قرار می گیرد. میزان این احتمال ۰/۸ در نظر گرفته شده است. روش های متفاوتی برای عملگر بازترکیبی در این مقاله مورد استفاده قرار گرفته است.

این روش ها شامل سه روش شناخته شده بازترکیبی یک نقطه‌ایی [۱۳]، بازترکیبی چند نقطه ایی [۱۳]، بازترکیبی حسابی ساده [۱۳]، یک روش ابداعی توسط لازلو و مهاراجه بنام "بازترکیبی منطقه ای"[1] [۸] و یک روش جدید ارائه شده در این مقاله بنام روش بازترکیبی حسابی منطقه ای می باشد. از توضیح سه روش استاندارد صرف نظر می شود و تنها دو روش بازترکیبی منطقه ای و بازترکیبی حسابی منطقه ای شرح داده می شود.

عملگر بازترکیبی منطقه ای روش ابداعی ارائه شده توسط لازلو و موهاراجه[2] [۸] است که در آن مراکز خوشه هایی را جابجا می کند که در یک منطقه از فضای پایگاه داده مانند R واقع شده باشند و یا بعبارت دیگر مراکز خوشه هایی را جابجا می کند که در همسایگی هم قرار دارند. در این روش باور بر آنست که اگر بتوانیم مجموعه مراکز خوشه اشغال شده در یک فضای یکسان یا نزدیک بهم را بصورت جواب های باکیفیت بالا در آوریم، تاحدودی مشکل مراکز خوشه اولیه حل می شود. نتایج نشان داده است که این روش علاوه بر آنکه به نتایج بهتری منجر می شود، زمان الگوریتم را نیز کاهش می دهد.

نکته مهم و اساسی در این روش تعیین منطقه R است چرا که ابتدا بایستی منطقه R مشخص شود و سپس آن مراکزی از دو کروموزم احتمالی مانند C_1 و C_2 را که عضو R هستند جابجا کنیم. منطقه R (شکل ۱)یک فضای d بعدی است که بصورت تصادفی انتخاب می شود. برای هر عملگر بازترکیبی بطور تصادفی یک نقطه مانند P انتخاب می کنیم و بردار تصادفی n را به آن عمود می کنیم. حال بخشی از فضا که از تلاقی نقطه P و بردار n بدست آمده و بسمت مثبت بی نهایت میل می کند را بعنوان فضای R در نظر می گیریم. این فضای کاملاً تصادفی انتخاب شده، ژن های کرموزوم ها را به دو دسته تقسیم می کند. درست شبیه آنچه در روش استاندارد بازترکیبی یک نقطه ای رخ می داد. حال آن دسته از مراکز خوشه ای را که در این منطقه قرار دارند باهم جابجا می نماییم.



شکل ۱. روش تعیین منطقه R

دومین عملگر بازترکیبی، عملگر بازترکیبی منطقه ای حسابی می باشد. این عملگر، نوعی بازترکیبی ابداعی در این مقاله است که با الهام از باز ترکیبی منطقه ای، تولید شده است. در این روش پس از تعیین منطقهR ، مراکزی(ژن ها) از دو کروموزوم جفت شده را که در منطقه R قرار دارند بطور تصادفی انتخاب کرده و پس از محاسبه میانگین حسابی آنها، میانگین ها را در کروموزوم های فرزند قرار می دهیم و بقیه ژن ها که جفت نشده اند بدون تغییر به کروموزوم های فرزند منتقل می شوند.

۴-۳. عملگر جهش

منطق استفاده از عملگر جهش در الگوریتم ژنتیک این است که الگوریتم در خلال فرایند جستجو در نقاط بهینه محلی به دام نیفتد و مناطق بیشتری از فضای جواب را به جستجو بپردازد. در این مسائل بطور معمول از عملگرهایی تحت عنوان جابجایی نقطه‌ای و یا جابجایی رشته ای استفاده می شود. در این مقاله برای عملگر جهش از یک روش رایج بنام جهش تصادفی[۸] [۷] و دو روش ابداعی جدید بنام های "جهش منطقه ای" و "جهش منطقه ای خوشه ای" استفاده شده است. جهش منطقه ای نوعی روش ابداعی است که از آن می توان به جهش منطقه ای با انتخاب n نقطه در منطقه R نیز یاد نمود. در این روش، منطقه ای مانند R بطور تصادفی انتخاب می گردد (همانطور که در عملگر بازترکیبی منطقه ای توضیح داده شد). n پارامتر ورودی الگوریتم است که بر اساس حجم پایگاه داده تعیین می شود. فاصله این n نقطه با مرکز خوشه ای که قرار است روی آن جهش صورت گیرد محاسبه می شود. نزدیکترین نقطه به مرکز خوشه انتخاب شده و جهش یا جابه جایی نقطه ای رخ می دهد. از آنجا که جهش بین نقاطی با فاصله کمتر و یا در همسایگی انجام می شود امید است که به نتایج بهتر نسبت به جهش تصادفی منجر شود. دومین عملگر جهش، "عملگر جهش منطقه ای خوشه ای" می باشد. این عملگر، نوعی جهش ابداعی در این مقاله است. در این روش ابداعی که از آن می توان به جهش منطقه ایی با انتخاب n نقطه از خوشه مربوط به ژن مورد جهش

[1]. *Regional Crossover*
[2]. *Michael Laszlo& Sumitra Mukherjee*

Archive of SID
www.SID.ir

نیز یاد نمود، ابتدا بطور تصادفی منطقه ای مانند R انتخاب می شود، سپس n نقطه از خوشه ژن مورد جهش انتخاب شده (n بسته به حجم پایگاه داده تعیین می شود) و فاصله این n نقطه با مرکز خوشه ای که قرار است روی آن جهش صورت گیرد، محاسبه می شود. نزدیکترین نقطه به مرکز خوشه انتخاب شده و جهش یا جابجایی نقطه ای رخ می دهد.

البته باید توجه داشت که در این روش در اثر عملگر بازترکیبی، خوشه بندی بهم خورده و باید یکباردیگر الگوریتم K-Means، اجرا شود تا خوشه های جدید بدست آیند و بعد از آن که خوشه های جدید مشخص شدند، به انتخاب n نقطه و ادامه الگوریتم پرداخته می شود. لازم به ذکر است بطور معمول برای عملگر جهش یک مقدار احتمالی ثابت وکوچک در نظر می گیرند. در این مطالعه P_m، ۰/۰۰۱ در نظر گرفته شده است.

۵-۳. جایگزینی

آخرین مرحله در هر تکرار از الگوریتم ژنتیک با انجام عملگر جایگزینی پایان می پذیرد. هدف از این مرحله حفظ بهترین کروموزوم های تولید شده در خلال تکرارهای قبلی است. بنابراین در هر نسل بازیابی صورت گرفته، بهترین کروموزوم های فعلی را نگه داشته و بهترین نسلِ کروموزوم های قبل بطور تصادفی با یکی از کروموزومهای فعلی (بجزء بهترین فعلی) جایگزین می شوند.

۶-۳. شرط خاتمه الگوریتم

در هر مرحله پس از ساخت و پیاده سازی مدل های پیشنهادی، جهت ارزیابی کیفیت خوشه بندی و مقایسه دقت مدل با نتایج سایر الگوریتم های موجود، از معیار انحراف درون خوشه ها استفاده شده است. کیفیت خوشه بندی بستگی به بیشترین شباهت اعضاء درون یک خوشه و کمترین شباهت اعضای یک خوشه از اعضاء سایر خوشه ها دارد. الگوریتم زمانی خاتمه می‌یابد که تمامی خوشه ها بررسی گردند و بهبودی در جواب حاصل نگردند.

۴. ارزیابی الگوریتم پیشنهادی

برای ارزیابی الگوریتم پیشنهادی از چهار مجموعه داده GTD، BPZ، IRIS و ۹۰۰-۲-۹ استفاده شده است. از ترکیب پنج عملگر بازترکیبی و سه عملگر جهش، ۱۵ مدل مورد بررسی قرار گرفت. برای انتخاب مدل پیشنهادی، ابتدا ۱۵ مدل با مجموعه داده های استاندارد انتخاب شده، آزمایش شدند. آنگاه نتایج حاصل از الگوریتم پیشنهادی با نتایج تحقیقات پیشین بر روی مجموعه داده های مختلف مقایسه گردید. از آنجا که بهترین جوابها با استفاده از عملگر بازترکیبی یک نقطه ای و عملگر جهش منطقه ای خوشه ای بدست آمده است، لذا این مدل بعنوان الگوریتم پیشنهادی معرفی گردیده و از ارائه نتایج حاصل از ۱۴ مدل دیگر صرفه نظر شده است. برای تنظیم پارامترهای الگوریتم پیشنهادی، ابتدا از طریق مطالعه تطبیقی، پارامترهای پیشنهاد شده در سایر مطالعات مرتبط جمع آوری گردید.

سپس از طریق روش آزمون و خطا مقادیر نهایی پارامترهای الگوریتم پیشنهادی تعیین گردیدند. برای تنظیم پارامترها از مجموعه داده های شبیه سازی شده استفاده شده است. پارامترهای مورد استفاده در نتایج بدست آمده شامل : نرخ بازترکیبی، P_c = ۰/۸ ، نرخ جهش، P_m= ۰/۰۰۱ ، اندازه جمعیت، ۱۰۰کروموزوم، تعداد نسل در هر اجرا ، ۱۰۰ بار و تعداد نمونه انتخاب شده برای جهش n= ۱۰ است. مجموعه داده های مورد استفاده به شرح جدول (۱) است.

جدول (۱). معرفی مجموعه داده های مورد استفاده

نام مجموعه داده	تعداد نمونه	تعداد ابعاد	منبع	مرجع
GTD	۵۹	۲	Spath (۱۹۸۰) [۱۴]	[۳] [۲] [۷] [۸]
BPZ	۸۹	۳	Spath (۱۹۸۰) [۱۴]	[۷] [۵]
IRIS	۱۵۰	۴	UCI [۱۵]	[۶] [۴]
۹-۲-۹۰۰	۹۰۰	۲	S.Bandyopadhyay & U.Maulik(2002)[۶]	[۴] [۶]

در ادامه نتایج اجرای مدل پیشنهادی بر روی چهار سری مجموعه داده مشروح در جدول (۱)، ارائه می شود. پیاده سازی الگوریتم جدید با زبان برنامه نویسی Java بر روی کامپیوتری با ۲ گیگا بایت حافظه اصلی و پردازنده ۱/۶ گیگا هرتز انجام شده است. مدل پیشنهادی برای هر مجموعه داده ده بار اجرا شده است. ضمنا تعداد نسل ها صد نسل در نظرگرفته شده که در هر نسل بهترین، بدترین و میانگین نتایج مطابق با معیار انحراف کم داده‌های درون یک خوشه محاسبه گردیده است. برای یک یک مجموعه داده ها نتایج با بهترین نتایج حاصل از مطالعات قبلی مقایسه شده است .

۱-۴. نتایج بدست آمده برای مجموعه داده GTD

جدول (۲) نتایج بدست آمده از الگوریتم پیشنهادی و همچنین بهترین جوابهای بدست آمده قبلی به ازای K های مختلف را برای مجموعه داده GTD نشان می دهد. این مجموعه داده دارای ۵۹ نمونه و دو بعد است.

Archive of SID

www.SID.ir

جدول ۲. مقایسه نتایج مدل پیشنهادی با الگوریتم های مورد استفاده در تحقیقات پیشین برای مجموعه داده GTD

تعداد K	بهترین جواب بدست آمده توسط[۲]،[۳]، [۷]	الگوریتم پیشنهادی	درصد اختلاف
۴	۴۹۶۰۰/۵۹	۴۹۶۰۰/۵۹	۰/۰۰
۵	۳۸۷۱۶/۰۲	۳۸۷۱۶/۰۲	۰/۰۰
۶	۳۰۵۳۵/۳۹	۳۰۵۳۵/۳۹	۰/۰۰
۷	۲۴۴۳۲/۵۷	۲۴۴۳۲/۵۷	۰/۰۰
۸	۲۱۴۸۳/۰۲	۲۱۴۸۳/۰۲	۰/۰۰
۹	۱۸۵۵۰/۴۴	۱۸۵۵۰/۴۴	۰/۰۰
۱۰	۱۶۳۰۷/۹۷	۱۶۳۰۷/۹۷	۰/۰۰

همانطور که مشاهده می شود نتایج بدست آمده از الگوریتم پیشنهادی با بهترین جواب های بدست آمده برای مجموعه داده GTD به ازای K های مختلف یکسان بوده و درصد اختلاف با آنها صفر است.

۴-۲. نتایج بدست آمده برای مجموعه داده BPZ

جدول (۳) نتایج بدست آمده از الگوریتم پیشنهادی و همچنین بهترین جوابهای بدست آمده قبلی به ازای K های مختلف را برای مجموعه داده BPZ نشان می دهد. این مجموعه داده دارای ۸۹ نمونه و سه بعد است.

جدول ۳. مقایسه نتایج مدل پیشنهادی با الگوریتم های مورد استفاده در تحقیقات پیشین برای مجموعه داده BPZ

تعداد K	بهترین جواب بدست آمده توسط[۷]	الگوریتم پیشنهادی	درصد اختلاف
۲	۶/۰۲۵۴۷	۶/۰۲۵۴۷	۰/۰۰
۳	۲/۹۴۵۰۷	۲/۹۴۵۰۷	۰/۰۰
۴	۱/۰۴۴۷۵	۱/۰۴۴۷۵	۰/۰۰
۵	۵/۹۷۶۱۵	۵/۹۷۶۱۵	۰/۰۰
۶	۳/۵۹۰۸۵	۳/۵۹۰۸۵	۰/۰۰
۷	۲/۱۹۸۳۲	۲/۱۹۸۳۲	۰/۰۰
۸	۱/۳۳۸۵۴	۱/۳۳۸۵۴	۰/۰۰
۹	۸/۴۲۳۷۴	۸/۴۲۳۷۴	۰/۰۰
۱۰	۶/۴۴۶۴۷	۶/۴۴۶۴۷	۰/۰۰
۱۱	۵/۱۹۷۹۸	۵/۱۹۷۹۸	۰/۰۰
۱۲	۳/۹۵۰۵۲	۳/۹۵۰۵۲	۰/۰۰
۱۳	۲/۷۷۸۰۳	۲/۷۷۸۰۳	۰/۰۰
۱۴	۲/۱۱۵۵۴	۲/۱۱۵۵۴	۰/۰۰
۱۸	۹/۸۰۶۸۷	۹/۸۰۷۰۰	۰/۰۰
۲۲	۵/۴۲۱۱۳۷	۵/۴۲۱۱۳۸	۰/۰۰
۲۶	۲/۸۲۲۳۵	۲/۸۲۲۳۷	۰/۰۰
۳۰	۱/۷۱۳۸۱	۱/۷۱۳۸۲	۰/۰۰

همانطور که مشاهده می شود نتایج بدست آمده از الگوریتم پیشنهادی با بهترین جواب های بدست آمده برای مجموعه داده BPZ به ازای K های مختلف یکسان بوده و درصد اختلاف با آنها صفر است.

۴-۳. نتایج بدست آمده برای مجموعه داده IRIS

جدول (۴) نتایج بدست آمده از الگوریتم پیشنهادی و همچنین بهترین جوابهای بدست آمده قبلی را برای مجموعه داده IRIS نشان می دهد. این مجموعه داده دارای ۱۵۰ نمونه، چهار بعد و سه خوشه (k=۳) است.

جدول ۴. مقایسه نتایج مدل پیشنهادی با الگوریتم های مورد استفاده در تحقیقات پیشین برای مجموعه داده IRIS

روش حل	بهترین جواب
GA & KGA Clustering]۴[]۶[	۹۷/۱۰۰۷۷
الگوریتم پیشنهادی	۷۸/۹۴۰۸۴
درصد بهبود	۲۳

همانطور که مشاهده می شود نتایج بدست آمده از الگوریتم پیشنهادی با بهترین جواب های بدست آمده برای مجموعه داده IRIS مقایسه شده اند. الگوریتم پیشنهادی بهبود ۲۳ درصدی را نسبت به تحقیقات پیشین نشان می دهد.

۴-۴- نتایج بدست آمده مجموعه داده ۹۰۰-۲-۹

جدول (۵) نتایج بدست آمده از الگوریتم پیشنهادی و همچنین بهترین جواب بدست آمده قبلی را برای مجموعه داده ۹۰۰-۲-۹ نشان می دهد. این مجموعه داده دارای ۹۰۰ نمونه و دو بعد و ۹ خوشه (k=۹) است.

جدول ۵. مقایسه نتایج مدل پیشنهادی با الگوریتم های مورد استفاده در تحقیقات پیشین برای مجموعه داده ۹۰۰-۲-۹

روش حل	بهترین جواب
GA Clustering]۴[	۹۶۶/۳۵۰۴۸۱
الگوریتم پیشنهادی	۴۶۹/۰۴۵۱۶۷
درصد بهبود	۱۰۶

همانطور که مشاهده می‌شود نتایج بدست آمده از الگوریتم پیشنهادی با بهترین جواب بدست آمده برای مجموعه داده ۹۰۰-۲-۹ مقایسه شده‌اند. الگوریتم پیشنهادی بهبود ۱۰۶ درصدی را نسبت به تحقیقات پیشین نشان می دهد.

Archive of SID

www.SID.ir

۵. نتیجه گیری

با توجه به اینکه در الگوریتم *K*-Means نمونه های اولیه به صورت تصادفی انتخاب می شوند، ممکن است الگوریتم در دام بهینه محلی قرار گرفته و جواب بهینه را تولید ننماید. لذا جهت خروج از وضعیت بهینه محلی، با ترکیب الگوریتم فوق الذکر با الگوریتم ژنتیک مدل خوشه بندی جدیدی ارایه شده است که سبب خروج از دام بهینه محلی و تولید جواب های بهتر می گردد.

در این مقاله سعی گردیده با ارائه عملگرهای جدید تغییر برای الگوریتم ژنتیک، مشکل به دام افتادن الگوریتم *K*-Means در نقاط بهینه محلی برطرف گردد. الگوریتم پیشنهادی با استفاده از پنج عملگر بازترکیبی و سه عملگر جهش که مجموعا ۱۵ حالت مختلف را تشکیل می دهند، آزمایش شده است.

نتایج بدست آمده نشان می دهد با استفاده از یک عملگر بازترکیبی ساده یک نقطه ای و عملگر جهش منطقه ای خوشه ای می توان نتایج بهتری نسبت به سایر روشها بدست آورد. در واقع در این روش جهش برروی مرکز خوشه مورد نظر بصورت تصادفی از کل فضای جواب صورت نمی گیرد، بلکه مرکز خوشه مورد نظر با یک نقطه از فضای تعریف شده در خوشه مورد نظر جابجا می‌شود.

مراجع

[1] Han, J., Kamber, M., " *Data Mining: Concepts and Techniques",* Second Edition, Morgan Kaufmann Publishers, 2006.

[2] Babu, G.P., Murty, M.N., *"A Near-Optimal Initial Seed Value Selection for K-Means Algorithm Using Genetic Algorithm"*, Pattern Recognition Letters, 14: 1993, pp. 763-769.

[3] Krishna, K., Narasimha Murty, M., *"Genetic K-Means Algorithm",* IEEE Transations on Systems, MAN, and Cybernetics – Part B: Cybernetics, 29: 1999, pp. 433-439.

[4] Bandyopadhyay, S., Maulik, U., *"Genetic Algorithm – based Clustering Technique*", Pattern Recognition Letters, 33, 2000, pp. 1455-1465.

[5] Hansen, P., Mladenovic, N., *"J-Means: A New Local Search Heuristic for Minimum Sum-of-Squares Clustering*", Pattern Recognition, 34, 2: 2001, pp. 405-413.

[6] Bandyopadhyay, S., Maulik, U., *"An Evolutionary Technique Based on K-Means Algorithm for Optimal Clustering in R^N"*, Pattern Recognition Letters, 146, 2002, pp. 221-237.

[7] Laszlo, M., Mukherjee, S., *"A Genetic Algorithm Using Hyper-Quadtrees for Low-Dimensional K-Means Clustering"*, Pattern Recognition Letters,28, 2006, pp. 533–543.

[8] Laszlo, M., Mukherjee, S., "*A Genetic Algorithm that exchanges neighboring centers for K-Means Clustering",* Patt ern Recognition Letters , 28, 2007, pp. 2359–2366.

[9] Chang, D.X., Zhang, X.D., Zheng, C.W., *"A Genetic Algorithm with Gene Rearrangement for K-Means Clustering*", Pattern Recognition, 42, 2009, pp. 1210 – 1222.

[10] Zalik, K.R., *"An Efficient K-Means Clustering Algorithm"*, Pattern Recognition Letters, 29, 2008, pp. 1385–1391.

[11] Xiao, J., Yan, Y., Zhang, J., Tang, Y., "*A Quantum-inspired Genetic Algorithm for K-Means Clustering*", Pattern Recognition Letters, Expert Systems with Applications, 37, 2010, pp. 4966–4973.

[12] Yu, Z., Wong, H.S., *"Quantization-based Clustering Algorithm",* Pattern Recognition 43, 2010, pp. 2698–2711.

[13] Eiben, AE., Smit, J.E., Introduction to Evolutionary Computing, Springer, 2003.

[14] Spath, H., *Clustering Analysis Algorithms*.Wiley, New York. 1980

[15] www.UCI.com.

Archive of SID

www.SID.ir

International Journal of Industrial Engineering & Production Management (2012)

June 2012, Volume 23, Number 1
pp. 121-128

http://IJIEPM.iust.ac.ir/



A Hybrid Clustering Method Using Genetic Algorithm with New Variation Operators

M.Yaghini*, R.Soltanian & J.Noori

Masood Yaghini, *Assistance Professor of Railway Eng-Iran University of Science and Technology*
Roya Soltanian*, MSc student of Master of Business Administration-Iran University of Science and Technology*
Javad Noori, *BSc student of Computer Eng-Iran University of Science and Technology*

Keywords

Clustering,
Genetic algorithm,
K-Means algorithm

ABSTRACT

The clustering problem under the criterion of minimum sum of squares is a non-convex and non-linear program, which possesses many locally optimal values, resulting that its solution often being stuck at locally optimal values and therefore cannot converge to global optima solution. In this paper, we introduce several new variation operators for the proposed hybrid genetic algorithm for the clustering problem. The novel mutation operator, called Clustering Regional Mutation, exchanges neighboring centers and a simple one-point crossover. The proposed algorithm identifies proper clustering. The experimental results are given to illustrate the effectiveness of the new genetic algorithm.

© 2012 IUST Publication, IJIEPM. Vol. 23, No. 1, All Rights Reserved

Archive of SID

* **Corresponding author. Masood Yaghini**
Email: *yaghini@iust.ac.ir*

www.SID.ir