ششمین همایش بین‌المللی افق های نویـن در
مهندسی برق، کامپیوتـر و مکانیک
6th International Conference on the New Horizons in
Electrical Engineering, Computer and Mechanical
www.mhconf.ir

خوشه‌بندی داده‌ها با استفاده از الگوریتم بهینه‌سازی تکاملی تجمع پرندگان

دکتر حسین مرادی[۱]، دکتر محمدرضا مصلحی[۲]، مهندس زهرا چناری[۳]

[۱]دانشگاه آزاد اسلامی واحد خمینی شهر، اصفهان Moradyhsnm@yahoo.com
[۲]موسسه آموزش عالی جهاد دانشگاهی استان اصفهان، اصفهان moslehi@acecr.ac.ir
[۳]موسسه آموزش عالی جهاد دانشگاهی استان اصفهان، اصفهان Zahra.chenari60@yahoo.com

چکیده

خوشه‌بندی داده‌ها در علوم مختلف کاربردهای فراوان دارد. مثلا در علوم مهندسی، علوم کامپیوتر، علوم پزشکی و زندگی، علوم زمین و ستاره شناسی، علوم اجتماعی، اقتصاد و غیره. یک دسته از این روشها الگوریتم‌های خوشه‌بندی تکاملی هستند. یکی از روشهای خوشهبندی داده ها استفاده ازالگوریتم های بهینه سازی تکاملی است.

به این صورت که یک تابع هدف که اهداف خوشه بندی را ارضا میکند تعیین میشود و سپس با استفاده از یک الگوریتم بهینه سازی تکاملی آن تابع بهینه میشود و خوشه ها را مشخص میکند. هر تابع هدف در خوشه بندی دو ویژگی را باید داشته باشد اول اینکه فاصله درون خوشه ای را کم کند و دوم اینکه فاصله بین خوشهای را زیاد کند. هدف از این تحقیق ارایه یک الگوریتم برای خوشه بندی داده ها است.

در این الگوریتم به مسئله خوشه بندی با دید یک مسئله بهینه سازی نگاه می شود. به این صورت که از یک یا چند شاخص خوشه بندی استفاده کرده و به عنوان تابع هدف در نظر گرفته می شود و با الگوریتم بهینه ساز ی تجمع پرندگان و تنظیم مناسب پارامترهای آن، این توابع هدف را بهینه می کنیم و خوشه های مناسب از الگوریتم استخراج می شوند.

سپس این الگوریتم با استفاده از شاخص های ارزیابی خوشه بندی ارزیابی شده و با دیگر الگوریتم های خوشه بندی مقایسه می شود .

این الگوریتم با شش الگوریتم معتبر خوشه بندی مقایسه می شود. لازم به ذکر است که مسئله خوشه بندی داده ها تا کنون با الگوریتم تجمع پرندگان انجام نشده است و این جنبه نوآوری تحقیق می باشد.

خوشه‌بندی یکی از شاخه‌های حائز اهمیت در علم داده کاوی می‌باشد که به بررسی شباهت و عدم شباهت داده‌ها و قسمت‌بندی آن‌ها می‌پردازد و دارای کاربردهای زیادی می‌باشد. الگوریتم‌های متعددی برای خوشه‌بندی داده‌ها وجود دارد. در این تحقیق خوشه‌بندی یک مسئله بهینه‌سازی در نظر گرفته می‌شود و با استفاده از یک تابع هدف مناسب به کمک یک الگوریتم بهینه‌سازی تکاملی سریع و کارا این مسئله را حل می‌کند. ایده این تحقیق در این است که تاکنون مسئله خوشه‌بندی با الگوریتم بهینه‌سازی تکاملی تجمع پرندگان حل نشده و نتایج این الگوریتم می‌تواند خوشه‌بندی دقیق و سریع باشد. استفاده از یک شاخص خوشه‌بندی مناسب که شرایط خوشه‌بندی را ارضا کند نیز دارای اهمیت است. همچنین استفاده از دیتابیس‌های مناسب و استاندارد نیز لازم است تا به صورت درست الگوریتم با استفاده یک معیار ارزیابی مناسب ارزیابی و مقایسه شود.

در حالت کلی یادگیری را می‌توان به دوگروه اصلی تقسیم کرد: یادگیری با نظارت و یادگیری بدون نظارت. در یادگیری با نظارت از ابتدا دسته‌ها مشخص هستند و هر یک از داده‌های آموزشی به دسته‌ای خاص نسبت داده شده است. اصطلاحاً گفته می‌شود که ناظری وجود دارد که در هنگام آموزش اطلاعاتی علاوه بر داده‌های آموزش در اختیار یادگیرنده قرار می‌دهد. ولی در یادگیری بدون نظارت هیچ اطلاعاتی به‌جز داده‌های آموزشی در اختیار یادگیرنده قرار ندارد و این یادگیرنده است که باید درون داده‌ها به دنبال ساختاری خاص بگردد.

الگوریتم خوشه‌بندی یک الگوریتم بهینه‌سازی می‌باشد که دو عمل بهینه‌سازی باید در آن اعمال شود. اول اینکه واریانس داخل خوشه کم بشود یا به اینکه فاصله درون خوشه‌ای کمینه شود و دوم اینکه فاصله بین خوشه‌ای بیشینه شود. این کمینه و بیشینه شدن یک مساله بهینه‌سازی است، بنابراین ما می‌توانیم این دو تابع را به وسیله حاصل جمع با ضرایب متفاوت به یک تابع یا شاخص بهینه‌سازی تبدیل کنیم و با استفاده از یک الگوریتم بهینه‌سازی که ما در این تحقیق از الگوریتم بهینه‌سازی تجمع پرندگان استفاده کرده‌ایم شاخص را کمینه یا بیشینه بکنیم و خوشه‌های بهینه را استخراج کنیم. انتخاب شاخص مناسب و الگوریتم بهینه‌سازی کار ما را در بهتر نتیجه گرفتن کمک می-کند.

واژه‌های کلیدی

بهینه سازی (Optimization)، خوشه بندی (Clustering)، الگوریتم تکاملی (Evolutionary algorithm)، الگوریتم تجمع پرندگان (Bird clustering algorithm)

مقدمه

خوشه‌بندی داده‌ها یک ابزار یادگیری بدون نظارت است که برای قسمت‌بندی یک مجموعه داده در گروه‌های همگن بر مبنای معیارهای مشابهت و عدم مشابهت استفاده می‌شود. بنابراین خوشه مجموعه ای از اشیاء می‌باشد که در آن اشیاء با یکدیگر مشابه بوده و با اشیاء موجود در خوشه‌های دیگر غیر مشابه می‌باشند. هدف توسعه یک الگوریتم اتوماتیک است که بتواند داده‌ها را در خوشه‌های مجزا با دقت بالا قرار دهد[1].

یکی از روش‌های خوشه‌بندی داده‌ها استفاده از الگوریتم‌های بهینه‌سازی تکاملی است، به این صورت که یک تابع هدف که اهداف خوشه-بندی را ارضا می‌کند تعیین می‌شود و سپس با استفاده از یک الگوریتم بهینه‌سازی تکاملی آن تابع بهینه می‌شود و خوشه‌ها را مشخص می‌کند. هر تابع هدف در خوشه‌بندی دو ویژگی را باید داشته باشد اول اینکه فاصله درون خوشه‌ای را کم کند و دوم اینکه فاصله بین خوشه‌ای را زیاد کند.

هدف از این تحقیق یافتن یک راه‌حل بهینه برای خوشه‌بندی داده‌ها از طریق الگوریتم‌های تکاملی می‌باشد. در این پژوهش قصد داریم روشی را پیدا کنیم که از لحاظ سرعت و دقت نسبت به الگوریتم‌های مشابه وضعیت بهتری داشته باشد و همچنین انعطاف‌پذیری بهتری در پارامترها داشته باشیم.

برای پیاده‌سازی الگوریتم تجمع پرندگان بر روی شاخص خوشه‌بندی و به‌دست آوردن خوشه‌های مناسب از نرم‌افزار متلب (Matlab) استفاده می‌شود. این نرم‌افزار ابزار مناسبی برای پیاده‌سازی الگوریتم موردنظر می‌باشد و همچنین ابزارهای زیاد و مناسبی برای مسائل داده کاوی در اختیار کاربر قرار می‌دهد. دیتابیس‌ها معمولا در نرم‌افزار اکسل (Excel) قرار دارند که با ارتباط با نرم افزار متلب از آن‌ها استفاده می‌کنیم. بعضی از دیتابیس‌ها باید به صورت موردی بررسی شوند، چون ویژگی‌های متفاوت و منحصر به‌فردی دارند، مانند منیفلد. در این تحقیق این مدل دیتابیس‌ها بررسی نمی‌شوند.

جدول ۱. دیتابیس‌هایی که برای ارزیابی تحقیق استفاده می‌شود

Database Name	K(number of segment)	Dimension
Jain	2	2
Flame	2	2
Thyoid	2	5
Wdbc	2	32
Pathbased	3	2
Spiral	3	2
Compound	6	2
Aggregation	7	2
Glass	7	7
Unbalance	8	2
Yeast	10	8
R 15	15	2
S 1	15	2
S 2	15	2
S 3	15	2
S 4	15	2

خوشه‌بندی داده‌ها یکی از الگوریتم‌های داده‌کاوی می‌باشد که در علوم مختلف کاربردهای فراوان دارد. مثلاً در علوم مهندسی (بینش محاسباتی، یادگیری ماشین، تشخیص الگو، مهندسی مکانیکی، مهندسی برق)، علوم کامپیوتر (کاربردهای خیلی زیادی از خوشه‌بندی را در وب‌کاوی، دسته بندی اسناد و یا دسته‌بندی مشتریان به سایت‌ها، تحلیل پایگاه داده فضایی، بازیابی اطلاعات، گردآوری مستندات متنی و پردازش تصویر، تقسیم‌بندی تصاویر پزشکی و یا ماهواره‌ای)، علوم پزشکی و زندگی (ژنتیک، زیست شناسی « دسته‌بندی حیوانات و گیاهان از روی ویژگی‌های آن‌ها»، میکروب شناسی، روانپزشکی، تکامل نژادی، آسیب شناسی)، علوم زمین و ستاره شناسی (جغرافیا، زمین شناسی)، علوم اجتماعی (جامعه شناسی، روان شناسی، انسان شناسی، کتابداری «دسته بندی کتاب ها»، آموزش و پرورش)، اقتصاد (بازاریابی، تجارت) و غیره. یک دسته از این روش‌ها الگوریتم‌های خوشه‌بندی تکاملی هستند.

هدف از این تحقیق ارائه یک الگوریتم خوشه بندی با استفاده از الگوریتم فراابتکاری تجمع پرندگان و طراحی یک شاخص خوشه بندی مناسب است. خوشه بندی داده ها یک مسأله بهینه‌سازی می باشد و استفاده از الگوریتم بهینه‌سازی تجمع پرندگان می تواند دقت و کارایی خوشه بندی را بالا ببرد. الگوریتم های تکاملی به خاطر ویژگی جمعیت گرایی و پوشش دادن بیشتر فضای مسله نسبت به روش های کلاسیک جواب‌های بهتری را می دهند. الگوریتم تجمع پرندگان نیز یک الگوریتم بهینه سازی تکاملی است که سرعت همگرایی آن بالا می باشد و می‌تواند در خوشه بندی داده ها موثر واقع شود. همچنین طبیعت مسله خوشه بندی داده ها یک مسله بهینه‌سازی است. نام الگوریتم CPSOA(Clustering Based Particle Swarm optimization Algorithm) است که یک روش خوشه بندی بر مبنای الگوریتم بهینه‌سازی تجمع پرندگان می‌باشد.)

هر الگوریتم خوشه‌بندی تکاملی موارد زیر را دنبال خواهد کرد:

الف- انتخاب یک کدگذاری کروموزم به عنوان نماینده جواب خوشه‌بندی و ایجاد جمعیت اولیه کروموزوم‌ها

ب- انتخاب یک توابع هدف مناسب از شاخص‌های صحت خوشه‌بندی

ج- طراحی اپراتورهای تکاملی مناسب شبیه انتخاب، ترکیب و جهش برای دستکاری کروموزم‌هایی که شامل جواب خوشه‌بندی هستند.

د- توسعه یک تکنیک برای انتخاب جواب خوشه‌بندی

نحوه کدگذاری کروموزم به عنوان نماینده جواب خوشه‌بندی

در این روش، برای یک دیتابیس با n داده، d بعد، و k خوشه، کروموزوم بصورت یک بردار با اعداد حقیقی به ابعاد $(k_{max}+T_{kmax})*d$ خواهد بود که k مراکز خوشه‌ها را در خود قرار می‌دهد و T بیشترین تعداد خوشه را دربر می‌گیرد. مقدار T بین صفر و یک می‌باشد و اگر در جریان الگوریتم مقدار T بیشتر از ۰,۵ شود به عنوان خوشه فعال شناخته خواهد شد و اگر مقدار آن کمتر از T باشد به عنوان خوشه غیر فعال خواهد بود و در نظر گرفته نمی‌شود. شکل ۱ بردار کروموزوم را نمایش می‌دهد. اگر i امین بردار کروموزوم باشد. $\vec{m}_{i,j}$ها مراکز خوشه‌ها می‌باشند که i شماره کروموزوم و j شماره خوشه می‌باشد. و همچنین $T_{i,j}$ فعال یا غیر فعال بودن خوشه را مشخص می‌کند.

IF *Ti,j* > 0.5, **THEN** the *j*th cluster center *mi,j* is **ACTIVE**
ELSE *_mi,j* is **INACTIVE**

$\vec{V}_i(t)=$

$T_{i,1}$	$T_{i,2}$	⋯	$T_{i,K_{max}}$	$\vec{m}_{i,1}$	$\vec{m}_{i,2}$	$\vec{m}_{i,3}$	⋯	$\vec{m}_{i,k_{max}}$

شکل ۱: نمایش کروموزوم استفاده شده در الگوریتم پیشنهادی

توابع هدف

شاخص‌های اعتبار خوشه بندی توابع ریاضی آماری هستند که برای ارزیابی نتایج یک الگوریتم خوشه‌بندی برمبنای کمی استفاده می‌شوند. به طور کلی یک شاخص اعتبار خوشه‌بندی می‌تواند دو هدف تعدادخوشه ها و بهترین خوشه‌بندی را دنبال کند. یک روش سنتی برای تعیین تعداد خوشه اجرای الگوریتم با تعداد خوشه‌های متنوع است و اینکه کدوم تعداد خوشه می‌تواند بهترین ارزیابی را در شاخص خوشه‌بندی داشته باشد. در حالت ایده آل، یک شاخص ارزیابی باید از دو جنبه خوشه‌ها را مراقبت کند. الف) همبستگی(cohesion): الگوها در یک خوشه باید تا حد ممکن مشابه باشند. شاخص واریانس الگوها در یک خوشه می‌تواند این همبستگی را نشان دهد. ب) تفکیک(separation): خوشه‌ها باید به خوبی ازهم تفکیک شوند. فاصله بین مراکز خوشه‌ها می‌تواند شاخص خوبی برای تفکیک باشد. در این الگوریتم از دو شاخص DI و CS استفاده شده است.

شاخص DB[2] : این معیار یک تابع از نسبت مجموع پراکندگی داخل خوشه به تفکیک بین خوشه‌ای است و این معیار هم از هر دو مفهوم همبستگی و تفکیک استفاده می‌کند. در ابتدا پراکندگی داخل خوشه ای iام را تعریف می‌کنیم و سپس فاصله بین خوشه‌ای خوشه iام و jام را تعریف می‌کنیم.

$$S_{i,q} = \left(\frac{1}{N_i}\sum_{\vec{X}\in C_i}\left\|\vec{X} - \overrightarrow{m_i}\right\|_2^q\right)^{1/q} \quad (۱)$$

$$d_{ij,t} = \left(\sum_{p=1}^{d}\left|m_{i,p} - m_{j,p}\right|^t\right)^{1/t} = \left\|\overrightarrow{m_i} - \overrightarrow{m_j}\right\|_t \quad (۲)$$

در جاییکه m_i هست iامین مرکز خوشه و $t \geq 1$ و q یک عدد صحیح می‌باشند و مستقلاً هر مقداری می‌توانند داشته باشند. N_i تعداد عناصر خوشه C_i می‌باشد. در ادامه $R_{i,qt}$ به صورت زیر تعریف می‌شود:

$$R_{i,qt} = \max_{j\in K, j\neq i}\left\{\frac{S_{i,q}+S_{j,q}}{d_{ij,t}}\right\} \quad (۳)$$

سرانجام معیار DB به صورت زیر تعریف می‌شود:

$$DB(K) = \frac{1}{K}\sum_{i=1}^{K} R_{i,qt} \quad (۴)$$

هرچه مقدار شاخص $DB(K)$ کمتر باشد خوشه های بهینه مطلوب‌تری خواهیم داشت.

شاخص *CS[51]*: در این شاخص مرکز خوشه به صورت میانگین بردارهای داده که متعلق به همان خوشه هستند به شکل زیر محاسبه می‌شود:

$$\overrightarrow{m_i} = \frac{1}{N_i}\sum_{x_j\in C_i}\overrightarrow{x_j} \quad (۵)$$

یک معیار فاصله بین دو نقطه داده $\overrightarrow{X_i}$ و $\overrightarrow{X_j}$ به صورت $d(\overrightarrow{X_i}, \overrightarrow{X_j})$ نشان داده می‌شود. سپس می‌توانیم معیار CS را به شکل زیر تعریف کنیم:

$$CS(K) = \frac{\sum_{i=1}^{K}\left(\frac{1}{N_i}\sum_{\overrightarrow{X_i}\in C_i}\max_{\overrightarrow{X_q}\in C_i}\left\{d(\overrightarrow{X_i},\overrightarrow{X_q})\right\}\right)}{\sum_{i=1}^{K}\left(\min_{j\in K, i\neq j}\left\{d(\overrightarrow{m_i},\overrightarrow{m_q})\right\}\right)} \quad (۶)$$

همانطوری‌که می‌توان به راحتی درک کرد این شاخص تابعی است که نسبت حاصل جمع پراکندگی درون خوشه ای به تفکیک بین خوشه‌ای را محاسبه می‌کند. هر چه مقدار این شاخص کمتر باشد خوشه‌های مطلوب‌تری را خواهیم داشت.

استفاده از الگوریتم تجمع پرندگان برای کمینه کردن تابع هدف

الگوریتم تجمع پرندگان، جستجو را از میان یک سری عوامل انجام می‌دهد.

این عوامل که به ذرات معروفند، مسیرهایی را به صورت آماری یا دقیق تعیین شده اند طی می‌کنند. هر ذره متاثر از بهترین موقعیت خود و بهترین موقعیت کل گروه است. اما باید به صورت تصادفی حرکت کند. هر ذره i، بوسیله ی بردار موقعیت خود x_i و بردار سرعت خود v_i تعریف می‌شود.

در هر بار تکرار این الگوریتم، هر ذره موقعیت مکانی جدیدی را بر اساس بردار سرعت جدید خود بدست می‌آورد.

ویژگی های الگوریتم PSO:

در عین سادگی قدرت بالایی را داراست.

عنوان الگوریتم تکاملی طبقه بندی می‌شود.

ذره‌ای که به جواب نزدیک‌تر باشد، شایستگی (Fitness) بیشتری دارد.

الگوریتمی پیوسته است.

اعضای جمعیت (ذره ها) یا جواب ها، به صورت مستقیم با هم ارتباط دارند.

مراحل الگوریتم PSO:

ایجاد جمعیت اولیه و ارزیابی آن

تعیین بهترین خاطرات شخصی و بهترین خاطره جمعی

بروزرسانی سرعت و موقعیت

در صورت برآورده نشدن شرایط توقف به مرحله ۲ می رویم.

پایان

شرایط توقف:

رسیدن به حد قابل قبولی از پاسخ

سپری شدن تعداد تکرار یا زمان مشخص

سپری شدن تعداد تکرار یا زمان مشخص بدون مشاهده بهبود خاص در نتیجه

بررسی تعداد مشخصی از پاسخ ها

عملکرد الگوریتم تجمع پرندگان به صورت مرحله به مرحله

سیستم با یک جمعیت تصادفی از جواب‌های محتمل شروع به کار می‌کند. هر عضو محتمل راه حل یک سرعت تصادفی دارد و به آن ذره می‌گوییم. این ذره در فضای پاسخ دارای یک موقعیت نیز می‌باشد. این ذرات در فضای پاسخ شروع به حرکت می‌کنند.

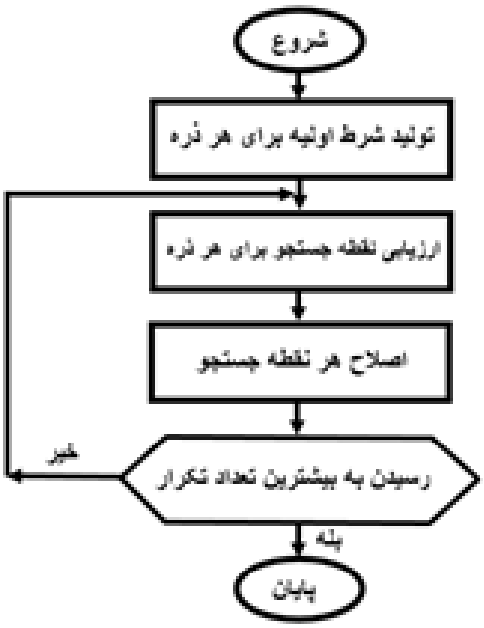
ارزیابی تناسب (fitness) هر ذره در بین گروه.

در هر بار تکرار این الگوریتم، تناسب ذره در آن تکرار با بهترین تناسب قبلی آن P_{best} مقایسه می‌گردد. اگر مقدار فعلی بهتر از P_{best} باشد، بنابراین P_{best} را برابر با مقدار فعلی قرار می‌دهیم. و موقعیت P_{best} را نیز به همین ترتیب.

مقایسه P_{best} ذرات با یکدیگر و بروزرسانی بهترین موقعیت کل گروه با بزرگترین تناسب g_{best}

سرعت هر ذره متناسب با P_{best} و g_{best} اش تغییر می‌کند (شتاب می‌گیرد). این شتاب با مقدار تصادفی وزن می‌گیرد. یک موقعیت جدید در فضای پاسخ برای هر ذره محاسبه می‌گردد. این کار بوسیله‌ی اضافه کردن سرعت جدید به هر جزء از بردار موقعیت ذره انجام می‌پذیرد.

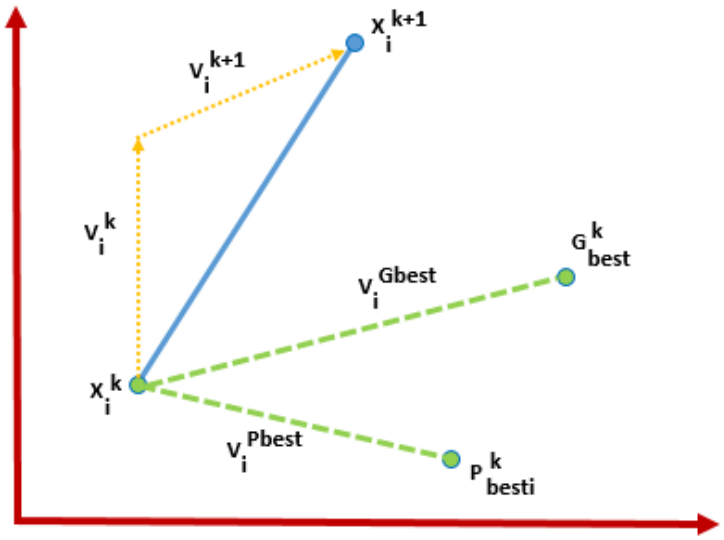
شکل ۲. عملکرد الگوریتم تجمع پرندگان



برتری الگوریتم تجمع پرندگان PSO نسبت به سایر الگوریتم‌ها

یکی از نکات مثبت این الگوریتم نسب به دیگر روش‌ها، کاهش تعداد پارامترهای مورد نیاز برای تنظیم و رعایت محدودیت‌هاست.

شکل زیر نمایشگر دو بعدی ذره‌ی i است.



شکل ۳. نمایشگر دو بعدی ذره i در الگوریتم تجمع پرندگان

حرکت ذره‌ی i در فضای پاسخ. در تکرار k و k+1 تکامل حرکت ذره متاثر از بهترین موقعیت ذره P_{best} و بهترین موقعیت کل گروه G_{best}

این شکل به خوبی نشان می‌دهد که چگونه بهترین موقعیت ذره P_{best} و بهترین موقعیت گروه G_{best} بر روی سرعت ذره در تکرار بعدی تاثیر گذار هستند. با این حال، خواص آماری الگوریتم این امکان را فراهم می سازد که کل فضای پاسخ جستجو گردد.

الگوریتم PSO در بهینه سازی مسائل چندهدفه:

در مسائل بهینه سازی چندهدفه، اهداف چندگانه نیاز به بهینه شدن به طور همزمان دارند. در اغلب موارد، جواب بهینه تکی (مجرد) معمولاً نمی‌تواند یافت شود تا تمام توابع هدف را بهینه سازد. در عوض یک گروه از جوابها وجود دارد که به عنوان مجموعه بهینه پارتو شناخته می‌شوند. راه حل ها در این گروه در غیاب برتری در میان اهداف، متعادل (برابر) هستند. مساله تصمیم‌گیری چندهدفه (MODM) از پرکاربردترین حوزه‌های الگوریتم PSO شده‌اند.

شبه کد الگوریتم PSO گام به گام در زیر نشان داده شده است:

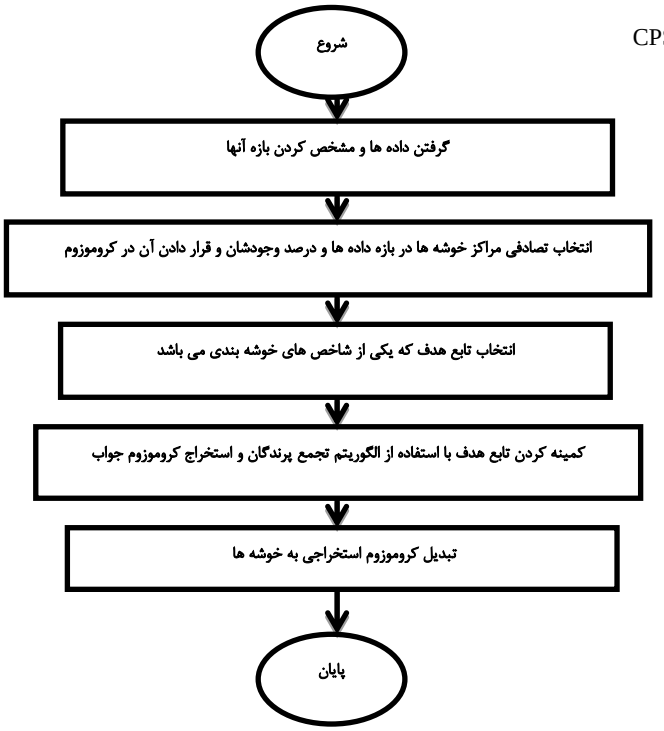
Algorithm gbest PSO
(Initialize)
gbet = X_0
for i = 0 to N do
pbest $_i$ = X_i (initialize randomly)
fitness$_i$= f (X_i)
if fitness$_i$ < f (gbest) then
gbest X_i
end if
end for

بدست آوردن جواب از کروموزوم خروجی الگوریتم تجمع پرندگان

کروموزوم نهایی جواب مسئله می‌باشد ولی جوابی که به ما می‌دهد فقط مراکز خوشه‌ها می‌باشد. برای بدست آوردن خوشه‌ها تمام نقاطی که به یک مرکز خوشه نزدیک‌تر از مراکز دیگر خوشه‌ها هست در یک خوشه قرار می‌گیرد و به این شکل می‌توانیم خوشه‌های خود را بشناسیم.

فلوچارت الگوریتم خوشه بندی داده‌ها توسط تجمع پرندگان

شکل ۴. فلوچارت الگوریتم پیشنهادیCPSOA



ارزیابی الگوریتم پیشنهادی

آزمایشات

کارایی الگوریتم پیشنهادی با الگوریتم‌های[4]MCPSO, [3] NC-Closers ,[5]single-linkage ,[6]DBSCAN ,[7]k-means مقایسه شده است. الگوریتم k-means داده‌ها را به نزدیک‌ترین مرکز خوشه خود نسبت می دهد و باید مقدار تعداد خوشه‌ها از قبل مشخص باشد. الگوریتم single-linkage داده‌ها را به صورت سلسله مراتبی خوشه بندی می‌کند و تعداد خوشه‌های نهایی باید مشخص باشد. الگوریتم DBSCAN داده‌ها را بر مبنای چگالی و فشردگی آن‌ها خوشه‌بندی می‌کند و الگوریتم NC-Closures داده‌ها را بر اساس مفهوم همسایگی خوشه‌بندی می‌کند.

معیار ارزیابی ARI

برای مقایسه کارایی این الگوریتم‌ها از شاخص ARI[8] استفاده شده است. این شاخص می‌تواند مقداری بین صفر تا یک داشته باشد. هرچه مقدار این شاخص به صفر نزدیک‌تر باشد، یعنی خوشه‌ها تصادفی ایجاد شده‌اند و هرچه مقدار این شاخص به یک نزدیک‌تر باشد، یعنی خوشه‌ها درست ایجاد شده‌اند. مقدار این شاخص از روی یک جدول درهمریختگی ایجاد می‌شود.

$X \backslash Y$	Y_1	Y_2	...	Y_s	Sums
X_1	n_{11}	n_{12}	...	n_{1s}	a_1
X_2	n_{21}	n_{22}	...	n_{2s}	a_2
⋮	⋮	⋮	⋱	⋮	⋮
X_r	n_{r1}	n_{r2}	...	n_{rs}	a_r
Sums	b_1	b_2	...	b_s	

(۷)

فرض کنید مجوعه S با n عضو داریم که به دو شکل X و Y تقسیم بندی شده اند. nij تعداد اشیایی است که در دو خوشه قرار دارند.

$X = \{ X_1 , X_2 , \dots , X_r\}$

$Y= \{Y_1, Y_2, \dots , Y_s\}$

بنابراین معیار ARI به صورت زیر تعریف می‌شود:

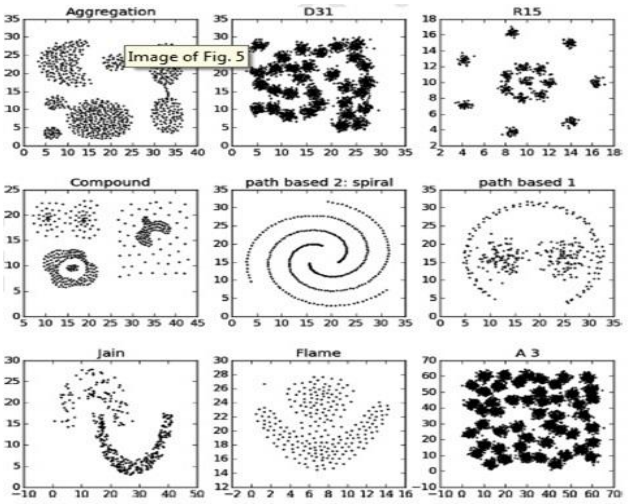
$$\overbrace{ARI}^{\text{Adjusted Index}} = \frac{\overbrace{\sum_{ij}\binom{n_{ij}}{2}}^{\text{Index}} - \overbrace{[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}]/\binom{n}{2}}^{\text{Expected Index}}}{\underbrace{\frac{1}{2}[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}]}_{\text{Max Index}} - \underbrace{[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}]/\binom{n}{2}}_{\text{Expected Index}}} \quad (۸)$$

دیتابیس‌های مورد استفاده

دیتابیس‌های مورد استفاده برای ارزیابی الگوریتم ارائه شده از سایت UCI Machine Learning Repository[9] و سایت Clustering Repository of the Speech and Image Processing Unit[10] می‌باشد که تعداد آنها شانزده است و نام‌های آن‌ها همراه با تعداد خوشه‌ها و ابعاد داده‌ها در جدول ۲ ذکر شده است. سعی شده است داده‌ها از نظر پراکندگی تنوع داشته باشند تا الگوریتم با داده‌های متنوع از نظر پراکندگی ارزیابی شود.

ابزارهای استفاده شده و محدودیت های پروژه

برای پیاده سازی الگوریتم ذرات بر روی شاخص خوشه بندی و به دست آوردن خوشه‌های مناسب از نرم‌افزار متلب(Matlab) استفاده می‌شود. این نرم‌افزار ابزار مناسبی برای پیاده‌سازی الگوریتم مورد نظر می‌باشد و همچنین ابزارهای زیاد و مناسبی برای مسائل داده کاوی در اختیار کاربر قرار می‌دهد. دیتابیس‌ها معمولاً در نرم‌افزار اکسل (Excel) قرار دارند که با ارتباط با نرم‌افزار متلب از آن‌ها استفاده می-کنیم. بعضی از دیتابیس‌ها باید به صورت موردی بررسی شوند چون ویژگی های متفاوت و منحصربه فردی دارند، مانند منیفلد. در این تحقیق این مدل دیتابیس‌ها بررسی نمی‌شوند.



شکل ۵. نمایش دو بعدی بعضی دیتابیس ها

جدول ۲. دیتابیس‌هایی که برای ارزیابی تحقیق استفاده می‌شود.

Database Name	K(number of segment)	Dimension
Jain	2	2
Flame	2	2
Thyoid	2	5
Wdbc	2	32
Pathbased	3	2
Spiral	3	2
Compound	6	2
Aggregation	7	2
Glass	7	7
Unbalance	8	2
Yeast	10	8
R 15	15	2

	S 1	15	2
	S 2	15	2
	S 3	15	2
	S 4	15	2

جدول ۳. جدول ارزیابی ARI برای خوشه‌بندی‌های الگوریتم‌هایK-means, DBSCAN, Single linkage, NC-Clouser , و مقایسه آن با روش پیشنهادی

Dataset	**K**	**D**	**Mean and standard deviation of ARI measured on the outputs of cgpso (over 40 independent runs)**	**K-means**	**DBSCAN**	**Single linkage**	**NC-Closures**
Jain	2	2	0.75 ± 0.004	0.41 ± 0.010	0.69 ± 0.002	0.37 ± 0.002	0.50 ± 0.014
Flame	2	2	0.81 ± 0.029	0.43 ± 0.005	0.56 ± 0.031	0.47 ± 0.004	0.51 ± 0.022
Thyoid	2	5	0.82 ± 0.014	0.68 ± 0.007	0.60 ± 0.001	0.63 ± 0.014	0.66 ± 0.052
Wdbc	2	32	0.83 ± 0.001	0.67 ± 0.032	0.74 ± 0.023	0.81 ± 0.011	0.73 ± 0.014
Pathbased	3	2	0.85 ± 0.034	0.84 ± 0.005	0.78 ± 0.002	0.72 ± 0.024	0.76 ± 0.005
Spiral	3	2	0.88 ± 0.076	0.70 ± 0.012	0.86 ± 0.041	0.79 ± 0.003	0.74 ± 0.002
Compound	6	2	0.87 ± 0.023	0.75 ± 0.022	0.74 ± 0.016	0.68 ± 0.023	0.70 ± 0.009
Aggregation	7	2	0.85 ± 0.023	0.81 ± 0.011	0.76 ± 0.067	0.80 ± 0.010	0.72 ± 0.027
Glass	7	7	0.93 ± 0.004	0.90 ± 0.011	0.83 ± 0.002	0.80 ± 0.006	0.79 ± 0.063
Unbalance	8	2	0.8 ± 0.001	0.89 ± 0.023	0.86 ± 0.067	0.78 ± 0.011	0.84 ± 0.014
Yeast	10	8	0.9 ± 0.034	0.76 ± 0.005	0.61 ± 0.023	0.71 ± 0.089	0.65 ± 0.094
R 15	15	2	0.77 ± 0.078	0.85 ± 0.074	0.80 ± 0.027	0.77 ± 0.008	0.83 ± 0.006
S 1	15	2	0.75 ± 0.003	0.69 ± 0.028	0.49 ± 0.022	0.53 ± 0.003	0.67 ± 0.002
S 2	15	2	0.7 ± 0.057	0.53 ± 0.035	0.31 ± 0.016	0.39 ± 0.007	0.45 ± 0.032
S 3	15	2	0.71 ± 0.023	0.43 ± 0.023	0.22 ± 0.073	0.30 ± 0.041	0.40 ± 0.065

S 4	15	2	0.85 ± 0.001	0.39 ± 0.023	0.28 ± 0.012	0.25 ± 0.004	0.33 ± 0.031

جدول ۴. جدول ارزیابی متوسط تعداد خوشه‌ها برای خوشه‌بندی‌های الگوریتم‌هایK-means, DBSCAN, Single linkage, NC-Clouser , و مقایسه آن با روش پیشنهادی

Dataset	**K**	**D**	**Mean and standard deviation of the average number of clusters (Av. *K*) (over 40 independent runs) for cgpso**	**K-means**	**DBSCAN**	**Single linkage**	**NC-Closers**
Jain	2	2	**2.82 ± 0.013**	2.27±0.007	2.10±0.000	5.35±0.001	2.20±0.003
Flame	2	2	**2.24 ± 0.403**	3.14±0.006	2.41±0.005	4.21±0.011	2.26±0.002
Thyoid	2	5	**2.27 ± 0.002**	3.36±0.003	13.68±0.012	9.63±0.010	5.11±0.002
Wdbc	2	32	**2.39 ± 0.456**	10.42±0.005	4.89±0.009	2.47±0.001	3.65±0.007
Pathbased	3	2	**3.11 ± 0.025**	8.36±0.022	6.12±0.003	8.45±0.003	5.44±0.006
Spiral	3	2	**3.42 ± 0.056**	10.98±0.008	3.43±0.006	14.32±0.022	7.42±0.001
Compound	6	2	**6.54 ± 0.014**	12.63±0.003	35.12±0.006	9.32±0.005	24.62±0.004
Aggregation	7	2	**7.32 ± 0.004**	7.86±0.004	19.37±0.003	15.63±0.006	10.42±0.001
Glass	7	7	**7.62 ± 0.071**	7.37±0.008	9.63±0.001	15.74±0.023	17.49±0.006
Unbalance	8	2	**8.44 ± 0.002**	9.11±0.006	12.84±0.007	0.88±0.006	10.42±0.066
Yeast	10	8	**12.1 ± 0.736**	15.32±0.003	32.14±0.013	27.32±0.004	25.13±0.003
R 15	15	2	**15.72 ± 0.014**	17.68±0.004	24.32±0.004	35.24±0.017	18.34±0.002
S 1	15	2	**18.79 ± 0.531**	25.34±0.004	39.63±0.007	32.13±0.004	29.42±0.002
S 2	15	2	**23.44 ± 0.701**	31.46±0.007	49.13±0.007	44.12±0.004	34.34±0.004
S 3	15	2	**25.77 ± 0.129**	41.63±0.007	37.36±0.007	47.63±0.009	30.14±0.004
S 4	15	2	51.63±0.007	51.63±0.007	76.61±0.011	87.16±0.001	64.38±0.002

توضیحات

جدول ۲ جدول ارزیابی ARI برای خوشه‌بندی‌های الگوریتم‌هایK-means, DBSCAN, Single linkage, NC-Clouser , می‌باشد. معیار ARI هرچه به عدد صفر نزدیکتر باشد نشان می‌دهد که خوشه‌ها بیشتر تصادفی ایجاد شده‌اند، ولی هرچه به عدد یک نزدیکتر باشد نشان می‌دهد که خوشه‌ها درست‌تر ایجاد شده‌اند. این جدول معیار ARI را برای الگوریتم‌های ذکر شده برای ۴۰ انجام شده و میانگین و انحراف معیار نتایج در جدول نمایش داده شده است. جدول ۳ جدول ارزیابی متوسط تعداد خوشه‌ها برای خوشه‌بندی‌های الگوریتم‌های

K-means ,DBSCAN ,Single linkage, NC-Clouser را نمایش می‌دهد که برای ۴۰ بار انجام شده‌اند و میانگین و انحراف معیار آن نمایش داده شده است. جدول ۲ و ۳ جدول ارزیابی ARI و متوسط تعداد خوشه‌ها برای خوشه‌بندی الگوریتم پیشنهادی CGPSO را نمایش می‌دهد که نتایج نشان می‌دهد الگوریتم پیشنهادی در هر دو مورد بهتر از بقیه الگوریتم‌ها بوده است.

منابع

[1] A. Mukhopadhyay, U. Maulik, and S. Bandyopadhyay, "A Survey of Multiobjective Evolutionary Clustering," *ACM Computing Surveys (CSUR)*, vol. 47, p. 61, 2015
H. Frigui and R. Krishnapuram, "A robust competitive clustering

[2] J. C. Dunn, "Well-separated clusters and optimal fuzzy partitions," *Journal of cybernetics*, vol. 4, pp. 95-104, 2018.

[3] J. Handl and J. Knowles, "Multiobjective clustering around medoids," in *Evolutionary Computation, 2005. The 2005 IEEE Congress on*, 2005, pp. 632-639.

[4] J. Handl and J. Knowles, "An evolutionary approach to multiobjective clustering," *Evolutionary Computation, IEEE Transactions on*, vol. 11, pp. 56-76, 2019.

[5] Z. Qian, X. Zhang, L. Jiao, and W. Ma, "Unsupervised texture image segmentation using multiobjective evolutionary clustering ensemble algorithm," in *Evolutionary Computation, 2008. CEC 2008.(IEEE World Congress on Computational Intelligence). IEEE Congress on*, 2008, pp. 3561-3567.

[6] E. Chen and F. Wang, "Dynamic clustering using multi-objective evolutionary algorithm," in *Computational Intelligence and Security*, ed: Springer, 2020 ,pp. 73-80.

[7] K. S. N. Ripon, C.-H. Tsang, S. Kwong, and M.-K. Ip, "Multi-objective evolutionary clustering using variable-length real jumping genes genetic algorithm," in *Pattern Recognition, 2020. ICPR 2006. 18th International Conference on*, 2006, pp .1200 -1203.

[8] S. Bandyopadhyay, U. Maulik, and A. Mukhopadhyay, "Multiobjective genetic clustering for pixel classification in remote sensing imagery," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 45, pp. 1506-1511, 2019.

[9] A. Mukhopadhyay and U. Maulik, "Multiobjective approach to categorical data clustering," in *Evolutionary Computation, 2007. CEC 2017. IEEE Congress on*, 2007, pp. 1296-1303.

[10] J.-M. Won, S. Ullah, and F. Karray, "Data clustering using multi-objective hybrid evolutionary algorithm," in *Control, Automation and Systems, 2008. ICCAS 2019. International Conference on*, 2008, pp. 2298-2303.