روشی برای خوشه بندی ترکیبی با استفاده از الگوریتم ازدحام ذرات بهینه شده

نفیسه رامندی[1]، عادل قاضی خانی[2*]

۱- کارشناسی ارشد ،دانشگاه بین المللی امام رضا (ع)،ایران،
n_ramandi@yahoo.com

۲- استادیار، دانشگاه بین المللی امام رضا (ع) ،ایران،
adel.ghazi@gmail.com

چکیده

خوشه بندی یکی از شاخه های یادگیری نظارت نشده است که طی فرایندی نمونه ها از یکدیگر جدا و در گروه های شبیه به هم قرار می‌گیرند. به دلیل بدون ناظر بودن مسئله خوشه بندی، انتخاب الگوریتمی خاص جهت خوشه بندی یک مجموعه ناشناس امری پرخطر و معمولا شکست خورده می‌باشد. به دلیل پیچیدگی مسئله و ضعف روش‌های خوشه‌بندی پایه، امروزه اکثر پژوهش‌ها به سمت روش‌های خوشه بندی ترکیبی هدایت شده است و همچنین برای بهبود کیفیت و استحکام خوشه‌ها از خوشه‌بندی ترکیبی استفاده می‌شود. پراکندگی و کیفیت نتایج اولیه از جمله عواملی است که در کیفیت نتایج حاصل از ترکیب موثر است. در این مقاله راهکاری برای بهبود دقت خوشه بندی و افزایش اطمینان پذیری سیستم، پیشنهاد شده است که مبتنی بر ازدحام ذرات می‌باشد. در این مقاله، از خوشه بندی ازدحام ذرات استفاده شده است که دو کار را انجام می دهد: ایجاد خوشه بند های پایه و تابع ترکیب، که پارامترهای اولیه الگوریتم خوشه بندی ازدحام ذرات مانند سرعت و وزن با استفاده از الگوریتم بهینه سازی ازدحام ذرات بهینه می‌شوند. این الگوریتم بر روی مجموعه داده‌هایی مورد ارزیابی قرار گرفت و ملاحظه شد که تغییرات دقت نسبت به روش مورد مقایسه، کمتر و میانگین دقت راهکار پیشنهادی در مجموعه داده‌های Iris، Bupa، Ionosphere و Wine به ترتیب ۲/۰۲٪، ۵/۷۵٪، ۱/۹۲٪ و ۲/۰۳٪ نسبت به روش مورد مقایسه بهبود داشته است.

کلمات کلیدی: خوشه‌بندی، خوشه‌بندی ترکیبی، ازدحام ذرات

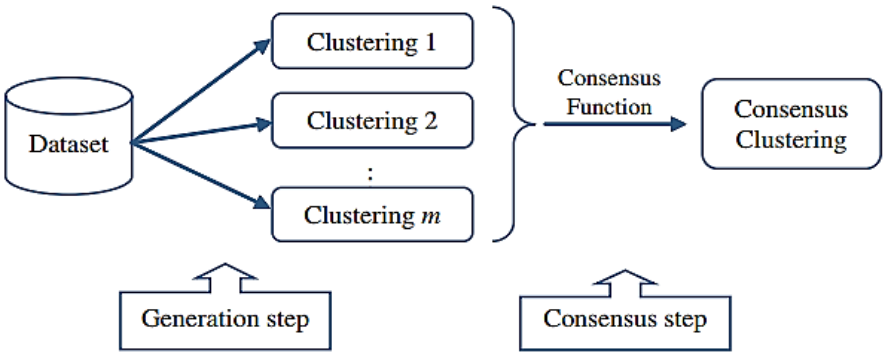
۱– مقدمه

خوشه‌بندی به دسته‌بندی بدون ناظر الگوها (مانند: مشاهدات، اقلام داده‌ها ، بردارهای ویژگی و...) به گروه‌های مشابهی به نام خوشه گفته می‌شود. به این معنی که نمونه‌های شبیه به هم باید در یک گروه قرار بگیرند و با نمونه‌های دسته‌های دیگر بیشترین تفاوت را داشته باشند. به عبارت دیگر خوشه‌بندی داده‌ها ابزاری ضروری برای پیدا کردن دسته‌ها در داده‌های بدون برچسب است [۱]. خوشه‌بندی در زمینه‌های بسیاری از قبیل مهندسی (یادگیری ماشین، هوش مصنوعی، تشخیص الگو، مهندسی مکانیک و الکترونیک) ، علوم کامپیوتر (کاوش وب ، تحلیل پایگاه داده فضایی، جمع آوری مستندات متنی، تقسیم بندی تصویر) ، علوم پزشکی (ژنتیک، زیست شناسی، میکروب شناسی، فسیل شناسی، روانشناسی، آسیب شناسی) ، علوم زمین شناسی (جغرافیا، زمین شناسی، نقشه برداری از زمین) ، علوم اجتماعی (جامعه شناسی، روانشناسی، تاریخ، آموزش و پرورش و اقتصاد (بازاریابی، تجارت) کاربرد دارد [۲, ۳].

از آنجایی که اکثر روش‌های خوشه‌بندی پایه روی جنبه‌های خاصی از داده‌ها تاکید می‌کنند، بنابراین روی مجموعه داده‌های خاصی کارآمد می‌باشند. به همین دلیل، نیازمند روش‌هایی هستیم که بتوانند با استفاده از ترکیب این الگوریتم‌ها و

گرفتن نقاط قوت هر یک، نتایج بهتری را تولید کنند. در واقع هدف اصلی خوشه‌بندی ترکیبی[1]جستجوی نتایج بهتر، با کیفیت و مستحکم‌تر ، با استفاده از ترکیب اطلاعات و نتایج حاصل از چندین خوشه‌بندی اولیه است. تحقیقات اخیر در این زمینه نشان داده‌اند که خوشه‌بندی داده‌ها می‌تواند به طور چشم گیری از ترکیب چندین افراز داده سود ببرد. خوشه‌بندی ترکیبی می‌تواند جواب‌های بهتری از لحاظ استحکام، نو بودن، پایداری و انعطاف‌پذیری نسبت به روش‌های پایه ارائه دهد[۴].

به‌طور کلی الگوریتم‌های خوشه‌بندی ترکیبی شامل دو مرحله اصلی تولید و ترکیب می‌باشند. مطابق شکل ۱



شکل ۱- مراحل خوشه بندی ترکیبی [۵]

اولین گام در خوشه‌بندی ترکیبی، تولید می‌باشد. بطوریکه در این مرحله مجموعه‌ای از خوشه‌ها یا افرازهای اولیه با استفاده از الگوریتم‌های خوشه‌بندی تولید و نتایج آن برای تولید افرار نهایی ذخیره می‌شود. فرایند تولید مناسب در مسئله بسیار مهم است زیرا نتیجه نهایی وابسته به خوشه‌های بدست آمده در این مرحله می‌باشد [۶]. به منظور یکپارچه سازی خوشه‌بندی ترکیبی قوی و پایدار نیاز به تنوع اجزای افرازها می‌باشد، اولین و ساده‌ترین روش برای ایجاد نتایج متنوع و پراکنده از یک مجموعه داده، استفاده از الگوریتم‌های مختلف خوشه‌بندی است [۷]. روش دیگر برای ایجاد پراکندگی، به دست آوردن نتایج متنوع از یک الگوریتم خوشه‌بندی پایه با استفاده از یکی از روش‌های زیر می‌باشد:

- تغییر مقادیر اولیه الگوریتم خوشه‌بندی انتخاب شده [۴]
- تغییر پارامترهای الگوریتم خوشه‌بندی انتخاب شده [۸]
- استفاده از زیرمجموعه‌های مختلف از ویژگی‌ها [۹]
- نگاشت داده‌ها به فضاهای ویژگی دیگر [۱۰]
- تقسیم بندی داده‌های اصلی به زیر مجموعه‌هایی متفاوت و مجزا [۱۱, ۱۲]

پس از اینکه نتایج اولیه تولید شد، بوسیله یک تابع ترکیب‌کننده این نتایج ترکیب می‌شوند و افراز نهایی بدست می‌آید. این مرحله همان تابع ترکیب‌کننده[2] می‌باشد که مهمترین گام در الگوریتم‌های خوشه‌بندی ترکیبی است. به طور کلی توابع ترکیب مختلفی وجود دارد که از جمله می‌توان به روش‌های مبتنی بر ابرگراف[۱۳, ۱۴] روش‌های مبتنی بر رای‌گیری [۱۰, ۱۵, ۱۶] روش‌های مبتنی بر تئوری اطلاعات[۱۳] روش‌های مبتنی بر ماتریس همبستگی [۴] اشاره کرد.

در این مقاله از تکنیک های خوشه بندی ازدحام ذرات و هرس اجماع با استفاده از عملگرهای انتخاب تکاملی در طراحی تابع ترکیب استفاده شده است. در الگوریتم بهینه سازی ازدحام ذرات[3] هر ذره یک افراز را کد گذاری می‌کند و ذرات در طول فرایند بهینه سازی حالت‌های خود را با تعامل با همسایگانشان بروزرسانی می‌کنند. اخیرا الگوریتم خوشه‌بندی ازحام

[1] Clustering ensemble
[2] Consensus functions
[3] Particle swarm optimization (PSO)

ذرات[1] برای خوشه‌بندی ظهور کرده است [۱۷]. در PSC هر ذره یک مرکز خوشه را نشان می‌دهد و ذرات خود را با استفاده از اصطلاحات شناختی، اجتماعی و خود سازمان دهی بروزرسانی می کند. برای جزئیات بیشتر به بخش ۲ مراجعه کنید.

در بخش ۲ الگوریتم خوشه‌بندی ازدحام ذرات بیان می‌شود، در بخش ۳ پیشینه تحقیق درباره خوشه‌بندی ترکیبی بیان می‌شود. جزئیات بیشتر الگوریتم پیشنهادی در بخش ۴ بیان می‌شود. تجزیه و تحلیل و آزمایشات در بخش ۵ بررسی می‌شود و در نهایت نتیجه گیری بیان می‌گردد.

۲-خوشه بندی ازدحام ذرات

الگوریتم PSO در سال ۱۹۹۵برای اولین بار توسط Kennedy و Eberhart مطرح شد [۱۸]. PSO یک الگوریتم جستجوی مبتنی بر جمعیت می‌باشد که از روی رفتار اجتماعی دسته‌های پرندگانی که به دنبال غذا می‌باشند، مدل شده‌است. ایده کلی آن است که هر عضو از ازدحام می‌تواند از تجارب قبلی اعضای دیگر در طول جستجوی راه حل استفاده کند. هر ذره می‌تواند با یک بردار n بعدی نشان داده شود که این بردار یک راه حل کاندید را برای مسئله بهینه‌سازی را نشان می‌دهد. همسایه‌ها هم با نگاشتی از ذرات جمعیت، تعریف می‌شوند که مدل‌های متفاوت همسایگی برای نشان دادن بهترین شخصی و بهترین سراسری وجود دارد که معروف‌ترین آن مانند مدل ون نیومن می‌باشد. این مدل‌ها با یک گراف نشان داده می‌شوند که رئوس آن ذرات می‌باشد. و دو ذره‌ای که با یالی به یکدیگر متصل باشند را همسایه هم می‌نامند. PSO روشی کلی برای بهینه‌سازی هدف می‌باشد که از سوی بسیاری از محققان به طور گسترده مورد استفاده قرار گرفته است.

الگوریتم PSC [۱۷] همانند PSO براساس ازدحام جمعیت است با این حال هر ذره یک خوشه واحد (مرکز خوشه) را نشان می‌دهد. بنابراین تمام جمعیت لازم است تا یک راه‌حل بالقوه را نشان دهد. این راه‌حل کاملا بدون نظارت تولید می‌شود و ذرات با هر مشاهده در هر مجموعه داده ظاهر می‌شوند و ذره برنده بروز رسانی می‌شود (یعنی نزدیکترین به مشاهده). به عبارت دیگر در الگوریتم PSO، هر ذره تمام راه‌حل های یک مسئله را شامل می‌شود ولی در الگوریتم PSC، هر ذره بخشی از راه‌حل های یک مسئله را شامل می‌شود. به عنوان مثال در خوشه‌بندی هر ذره مطابق با یک مرکز خوشه است. در این الگوریتم هیچ تابع ارزیابی لازم نمی‌باشد. ویژگی دیگر این است که علاوه بر وجود مولفه های شناختی و اجتماعی که در PSO نیز است، PSC دارای مولفه خود سازماندهی می‌باشد. به طور کلی هر ذره متشکل از سه حالت زیر است:

۱- سرعت فعلی ذره از طریق فضای وزن $s \in R^d$

۲- موقعیت فعلی ذره $p \in R^d$

۳- بهترین موقعیت ذره $b \in R^d$

در هر تکرار t، مشاهده ای به طور تصادفی انتخاب می‌شود که x_α نامیده می‌شود که از مجموعه داده نمونه برداری می‌شود: $X = \{x_1, \ldots, x_\alpha, \ldots, x_n\} \subset R^d$ که در آن n، کاردینالیتی و d، بعد مجموعه داده می‌باشد. در هر تکرار t، ذره برنده (نزدیک ترین ذره به مشاهده ورودی x_α) ذره i (i=1,…,c) با توجه به رابطه (۱) بروزرسانی می‌شود. که c برابر با تعداد ذرات (تعداد خوشه ها) و پارامترهای $\phi(t)$, Φ_1, Φ_2, Φ_3 برای کنترل رفتار جنبشی ذرات مورد استفاده قرار می‌گیرند. که به طور خاص $\phi(t)$، برابر با وزن اینرسی که در این مقاله برابر با: $\phi(t=0) = 0.95\phi(t)$, $\phi(t+1)$ و Φ_1, Φ_2, Φ_3 به ترتیب بردارهای مثبت برای بروزرسانی مقادیر شناختی، اجتماعی و خود سازماندهی می‌باشند، مطابق رابطه ۱:

$$
\begin{aligned}
s_i(t+1) &= \varphi(t) s_i(t) \\
&\quad + \phi_1 \otimes (b_i^{\alpha}(t) - p_i(t)) \\
&\quad + \phi_2 \otimes (g^{\alpha}(t) - p_i(t)) \\
&\quad + \phi_3 \otimes (x_\alpha - p_i(t)) \\
p_i(t+1) &= p_i(t) + s_i(t)
\end{aligned}
\qquad (۱)
$$

[1] Particle swarm clustering (PSC)

سرعت همانند PSO، در محدوده $\pm S_{max}$ می‌باشد که مانع از انحراف ذرات در طی اجرا می‌شود.

در این مقاله برای ایجاد تنوع در خوشه‌بندی سه پارامتر تعداد ذرات C، وزن اینرسی و سرعت ماکزیمم به صورت دستی تنظیم شده است.

۳–پیشینه

روش‌های خوشه‌بندی ترکیبی سعی می‌کنند با ترکیب افرازهای مختلف تولید شده از روش‌های خوشه‌بندی پایه یک افراز مستحکم از داده‌ها را تولید کنند. خوشه‌بندی ترکیبی جایگزین خوبی برای تجزیه و تحلیل‌های خوشه‌بندی می‌باشد. هدف این ترکیب، استفاده از روش‌هایی برای بهبود کیفیت داده‌ها می‌باشد. که در سال‌های اخیر نتایج امیدوار کننده‌ای در این رابطه بدست آمده است. در زیر تحقیقات مشابه در این زمینه بیان می‌شود که می‌توان در قالب دو قسمت، روش‌های خوشه‌بندی بدون استفاده از الگوریتم‌های فراابتکاری و روش‌های خوشه‌بندی با استفاده الگوریتم‌های فراابتکاری، بیان کرد.

۳–۱–خوشه‌بندی ترکیبی بدون استفاده از الگوریتم‌های فراابتکاری

نویسنده در [۱۹] روشی برای خوشه‌بندی ترکیبی ارائه داد، که ایده کلی روش این است که به جای استفاده از همه خوشه‌ها، تنها از زیرمجموعه‌ای از خوشه‌های پایه استفاده شود ابتدا در این روش از الگوریتم K-means به عنوان خوشه‌بند پایه برای تولید نتایج اولیه استفاده می‌شود. پراکندگی لازم در نتایج اولیه برای الگوریتم K-means نیز، با انتخاب تصادفی نقاط اولیه مراکز خوشه‌ها و همچنین با نمونه برداری بدست آمده است. سپس در مرحله بعد کیفیت خوشه‌های به دست آمده مورد ارزیابی قرار می‌گیرند. در مرحله بعد، خوشه‌ها با توجه به مقدار پایداری خوشه انتخاب می‌شوند. در گام آخر خوشه‌های انتخاب شده با هم ترکیب شده و خوشه نهایی از آن‌ها به دست می‌آید. استرل و گاش [۱۴] معیاری برای انتخاب از بین ترکیب‌های ممکن ارائه دادند که مبتنی بر کیفیت کلی یک خوشه‌بندی بود. آن ها میزان ثبات بین افرازهای ترکیبی و افرازهای پایه را در نظر گرفتند و با استفاده از یک قائده ترکیبی ثابت، معیار شباهت دو به دو را روی فضای ویژگی d بعدی به کار بردند. نویسندگان در [۲۰] یک روش مبتنی بر بازنمونه برداری را برای بررسی اعتبارسنجی نتایج خوشه بندی فازی ارایه کرده اند. در چند سال اخیر پایداری خوشه به عنوان یک معیار ارزیابی خوشه مورد توجه زیادی قرار گرفته است. نویسندگان در [۲۱] و [۲۲]، نیز روشی مبتنی بر بازنمونه برداری برای اعتبارسنجی خوشه ارایه کرده اند .عنصر اصلی در این روش، پایداری خوشه است . معیار پایداری، میزان همبستگی افرازهای به دست آمده از دو نمونه برداری مستقل از مجموعه داده را اندازه گیری می‌کند. هر چه میزان پایداری برای یک خوشه‌بندی بیشتر باشد، به این معنی است که اگر الگوریتم خوشه بندی چندین مرتبه دیگر روی آن نمونه ها به کار رود، نتایج مشابهی حاصل می‌شود.

در [۲۳] یک روش خوشه‌بندی ترکیبی ارایه شده است که در آن با استفاده از معیار پایداری خوشه شباهت دو به دو آموزش داده می‌شود. در این روش، به جای استفاده از معیارهای ارزیابی مبتنی بر افراز نهایی، افرازهای حاصل از الگوریتم‌های پایه در نواحی مختلف از فضای ویژگی d- بعدی مورد ارزیابی قرار گرفته‌است.

۳–۲–خوشه‌بندی ترکیبی با استفاده از الگوریتم‌های فراابتکاری

در [۲۴] برای خوشه‌بندی ترکیبی دو گام در نظر گرفته شده است. در گام اول برای خوشه‌بندی پایه از سه الگوریتم کلونی مورچه استفاده شده است. و سپس در گام دوم خوشه‌های بدست آمده از گام قبل با استفاده از روش ابرگراف با هم ترکیب می‌شوند. در [۲۵] از ازدحام ذرات در تابع ترکیب استفاده کرده است. نویسندگان در این روش خوشه‌بندی ترکیبی وزن دار مبتنی بر الگوریتم ازدحام ذرات را ارائه دادند. ساده‌ترین روش برای ترکیب خوشه‌ها، رای اکثریت می‌باشد. روش رای‌گیری بر اساس وزن‌ها نوع دیگری از روش رای گیری می‌باشد که نیازمند محاسبه وزن‌ها با بررسی اطلاعات خوشه‌ها می‌باشد. که تخمین وزن‌ها بر اساس الگوریتم PSO با تابع برازندگی متکی بر نرخ خطای مجموعه تصدیق، صورت گرفته شده است، و

وزن‌های بهینه با جستجو در یک فضای k بعدی بدست آمده اند. و افرازهای تولید شده بوسیله خوشه‌بندهای پایه، قبل از ترکیب شدن، همتراز می‌شوند. ولی نتایج آزمایشات روی چند مجموعه داده نشان داد، این روش بر روی مجموعه داده‌هایی که نمونه‌های کمتری دارند، به دلیل مشکل over fitting، روش مناسبی نیست.

الگوریتم [۲۶] برای بهبود کارایی خوشه‌بندی ترکیبی پیشنهاد شده‌است که مبتنی بر استفاده از زیرمجموعه‌ای از خوشه‌های اولیه می‌باشند. ایده اصلی در این روش برای انتخاب زیرمجموعه‌ای از خوشه‌ها، استفاده از خوشه‌های پایدار با الگوریتم‌های جستجوی هوشمند می‌باشد. از k-means با k های متفاوت برای ایجاد خوشه‌های اولیه پراکنده استفاده شده‌است. برای ارزیابی خوشه‌ها، از معیار پایداری مبتنی بر اطلاعات متقابل استفاده شده‌است. در گام بعد عمل انتخاب خوشه‌ها توسط الگوریتم ژنتیک انجام می‌شود، در آخر نیز خوشه‌های انتخاب شده به کمک چندین روش ترکیب نهایی با هم جمع می‌شوند. در [۲۷] خوشه‌بندی ترکیبی به عنوان یک مسئله بهینه سازی چند هدفه در نظر گرفته شده‌است که یک الگوریتم ژنتیک برای آن پیشنهاد داده‌شد. که در آن دو معیار بهینه سازی می‌شود. اول اینکه شباهت را با استفاده از محاسبه شاخص Adjusted Rand به حداکثر می‌رساند. دوم اینکه انحراف معیار به حداقل برسد. در این روش جمعیت اولیه را خوشه‌های ورودی در نظر می‌گیریم. در اینجا روش انتخاب مسابقه استفاده شده است و از تکنیک مسابقه باینری فراگیر برای انتخاب کروموزم ها برای عمل تقاطع استفاده شده‌است. و از روش برچسب زنی مبتنی بر گراف دوبخشی برای هدف تقاطع استفاده شده است. گراف دو بخشی براساس محاسبه عدم شباهت بین خوشه‌ها ایجاد می‌شود.

در [۲۸] از ازدحام ذرات هم به عنوان الگوریتم خوشه‌بندی پایه و هم به عنوان تابع ترکیب استفاده می‌شود. این روش شامل دو مرحله می‌باشد: مرحله تنوع و مرحله تجمیع. در مرحله تنوع، چندین الگوریتم خوشه‌بندی بر اساس همکاری چندگانه ازدحام [۲۹]، راه‌حل‌هایی را برای داده‌های داده شده ایجاد می‌کند. سپس در مرحله تجمیع، خروجی‌های بدست آمده از مرحله قبل به عنوان یک فضای ویژگی جدید در نظر گرفته می‌شود، تا با استفاده از یک روش جدید همکاری چندگانه ازدحام خوشه‌بندی شوند.

جدیدترین روش خوشه‌بندی ترکیبی مبتنی بر PSO که خوشه‌بندی ترکیبی ازدحام ذرات[1] نام دارد [۳۰]. در این روش دو کار انجام گرفته شده است: ۱- خوشه‌بندی ازدحام ذرات. ۲- هرس اجماع[2]. در ابتدا از PSC به عنوان خوشه‌بند های پایه استفاده می‌شود (خوشه‌بندی ها یکسان ولی پارامترها مختلف) سپس PSC نقش تابع اجماع را بازی می‌کند که تابع اجماع پیشنهادی عمل هرس را انجام می‌دهد.در این روش PSC دو نقش مجزا دارد: اول اینکه PSC نقش یک خوشه‌بند پایه را دارد یعنی افرازهای مختلفی از الگوریتم PSC بدست می‌آید. دوم اینکه نقش یک تابع ترکیب را دارد که هرس اجماع را انجام می‌دهد، یعنی افرازهای پایه را قبل از مرحله نهایی (افراز نهایی) کاهش می‌دهد. به عبارت دیگر در الگوریتم هرس، به عنوان ورودی مرکز خوشه‌های بدست آمده از C تا الگوریتم خوشه‌بندی را دریافت می‌نماید سپس در گام هرس تکاملی[3] از بین این C تا الگوریتم خوشه‌بندی به تعداد C' تا از آنها را براساس یکی از عملگرهای انتخاب، مانند چرخ گردان انتخاب می‌نماید و سپس در مرحله بعدی مراکز این C' تا الگوریتم انتخابی را با الگوریتم PSC ترکیب می‌کند و مراکز نهایی خوشه‌ها را تعیین می‌نماید.

۴-روش پیشنهادی

طبق نتایج بدست آمده از روش PSCE، دریافتیم که تغییرات دقت اجرای الگوریتم در تکرارهای متعدد، نامنظم است. به عبارت دیگر دقت الگوریتم در اجراهای مختلف با تغییرات زیادی مواجه است. که این باعث می‌شود اطمینان‌پذیری سیستم کاهش پیدا کند. به همین دلیل راهکاری برای حل این مشکل و افزایش دقت الگوریتم ارائه دادیم. با بررسی پارامترهای اولیه PSC

[1] Particle swarm clustering ensemble (PSCE)
[2] Ensemble pruning
[3] Evolutionary pruning

مانند V_{max} و W دریافتیم که آن‌ها تاثیر بسیاری برای تعیین جایگاه مراکز خوشه دارند که این جایگاه مناسب مراکز خوشه تاثیر مستقیمی در تولید افرازهای اولیه و تغییرات دقت دارند، به همین دلیل درصدد بهبود این پارامترها برآمدیم. در روش PSCE از پارامترهایی برای خوشه‌بندی PSC استفاده شده است که عدم اطمینان‌پذیری و یا به عبارتی افزایش و کاهش نامنظم دقت را در پی داشت که این مسئله باعث کاهش میانگین دقت نیز شده است. در این مقاله، پارامترهای V_{max} و W در PSC با استفاده از الگوریتم PSO بهینه شدند که باعث شد خوشه بندی بهتر کند. در الگوریتم‌های تکاملی مانند PSO، تابع برازندگی رکن اساسی آن می‌باشد که اگر تابع برازندگی به خوبی تنظیم نشود و یا اصلا وجود نداشته باشد، سیستم به خوبی کار نمی‌کند. بنابراین در روش خوشه بندی ترکیبی ازدحام ذرات بهینه شده[1]، میانگین دقت در چند بار اجرای الگوریتم به عنوان تابع برازندگی برای PSO در نظر گرفته شده است. پس از بهینه شدن پارامترهای V_{max} و W، وارد مراحل تولید و ترکیب افرازها می‌شویم. در این مقاله ما علاقه‌مند به تشکیل اجماعی از خوشه‌بندهای PSC فردی از طریق یک PSC دیگر هستیم. رویکردهای اجماع هیچ نیازی به همبستگی خوشه‌بندی یا سازگاری برچسب‌های خوشه‌ای ندارند (عملی که از لحاظ محاسباتی گران می‌باشد و مهمتر از همه کیفیت نتیجه نهایی را تحت تاثیر قرار می‌دهد). روال کلی خوشه‌بندی ترکیبی به این صورت است که: در ابتدا مجموعه داده X تبدیل به $C^{(j)}$ خوشه‌بند پایه می‌شود $C^{(j)}= (j = 1, . . ., C)$ که C برابر با تعداد خوشه‌ها می‌باشد و سپس افرازها براساس $C^{(j)}$ تا خوشه‌بند پایه با بردار برچسب $\lambda^{(j)}$ تولید می‌شوند و در نهایت یک تابع Γ ترکیب استفاده می‌شود که افرازها را برای رسیدن به نتیجه نهایی λ، ترکیب می‌کند:

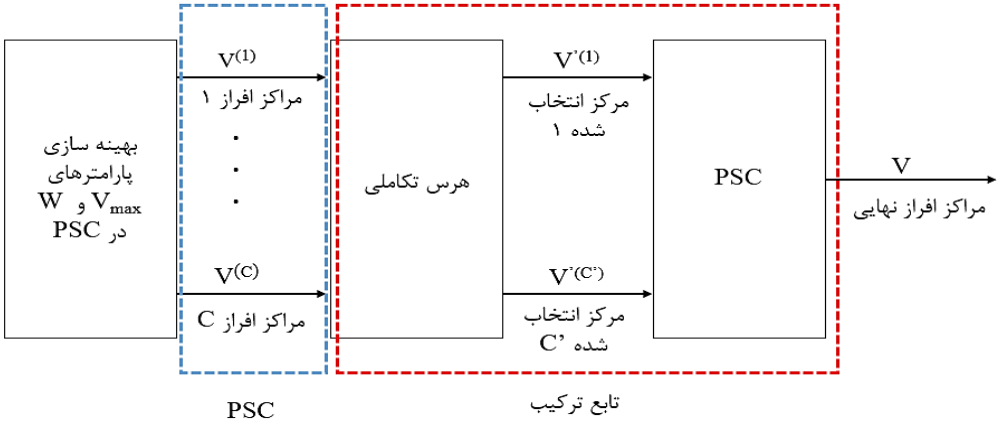
$\lambda= \Gamma(\lambda^{(1)}, . . ., \lambda^{(j)}, . . ., \lambda^{(C)})$

در گام اول که هرس تکاملی نام گرفته است، انتخاب افرازها $V = V(1)\cup . . . \cup V(C)$ انجام می‌گیرد که این انتخاب براساس برآورد کیفیت افرازها صورت می‌گیرد. شایستگی یک افراز به طور مستقیم در دسترس نیست بنابراین در این مقاله از شاخص ارزیابی داخلی مانند معیار Xie–Beni استفاده شده است و زیر مجموعه‌ای از خوشه‌ها با استفاده از عملگر انتخاب چرخ گردان انتخاب می‌شوند. رابطه (۱) فرمول محاسبه معیار Xie–Beni را نشان می‌دهد که در این معیار هدف، مینیمم شدن آن می‌باشد.

$$XB = \frac{\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^{m} \left\| x_j - v_i \right\|^2}{n \min_{i \neq j} \left\| v_i - v_j \right\|^2} \qquad (۱)$$

در گام دوم افرازهای انتخاب شده $V'^{(1)}\cup . . . \cup V'^{(C')}$ از مرحله قبل، به عنوان ورودی برای ترکیب نهایی استفاده می‌شوند. که این ترکیب با استفاده از الگوریتم PSC دیگری با همان پارامترهای قبل صورت می‌گیرد. و در نهایت افراز نهایی و مراکز خوشه‌های نهایی بدست می‌آید. در شکل ۲ فلوچارت الگوریتم OPSCE بیان شده است:

[1] Optimized particle swarm clustering ensemble(OPSCE)



شکل ۲- فلوچارت الگوریتم OPSCE

۵-نتایج ارزیابی

الگوریتم OPSCE بر روی سیستمی با مشخصات پردازنده intel core i7 ، حافظه 8GB و سیستم عامل 64bit و در محیط متلب R2013a پیاده‌سازی شده است. که برای ارزیابی آن از مجموعه داده‌هایی استفاده شده است که از سایت UCI انتخاب شده است. که در جدول ۱ اطلاعات مربوط به این مجموعه داده‌ها بیان شده است:

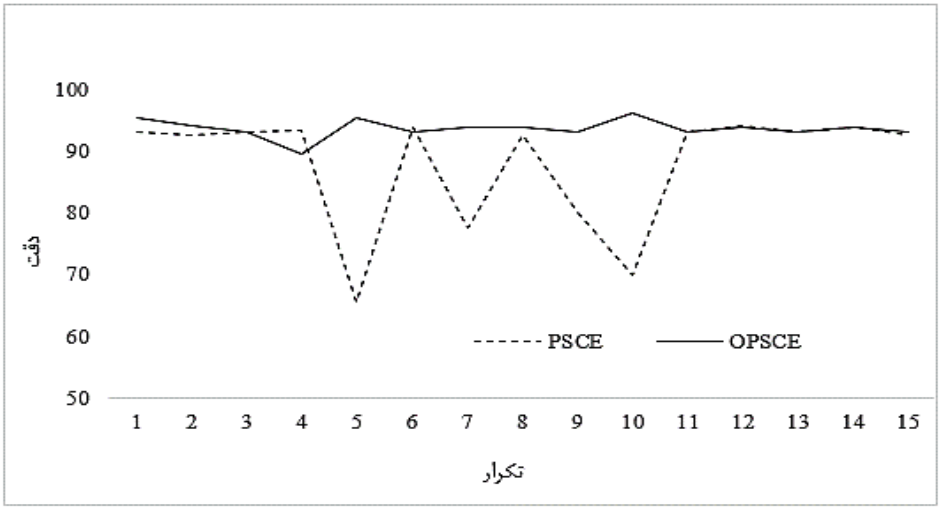
جدول ۱- مجموعه داده ها

مجموعه داده	تعداد نمونه	تعداد کلاس	تعداد ویژگی
Ionosphere	۳۵۱	۲	۳۴
Iris	۱۵۰	۳	۴
Wine	۱۷۸	۳	۱۳
Bupa	۳۴۵	۲	۷

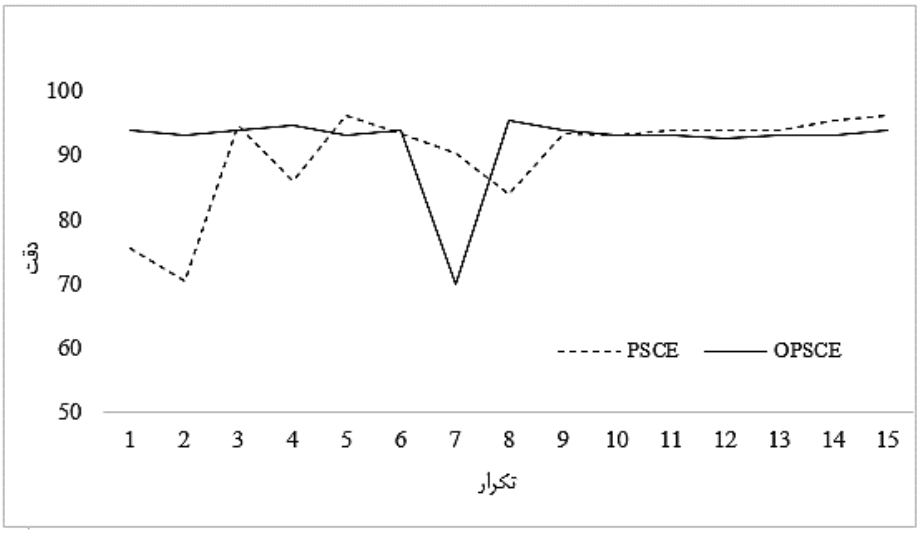
۵-۱-مقایسه دقت الگوریتم OPSCE و PSCE روی مجموعه داده‌های مختلف

میزان تغییرات دقت در الگوریتم OPSCE و PSCE بر روی مجموعه‌داده‌های Ionosphere، Bupa، Wine و Iris مورد ارزیابی قرار گرفت. همانطور که ملاحظه می‌شود، در هر چهار مجموعه داده (شکل‌های ۳-۶)، تغییرات دقت در ۱۵ مرتبه اجرا در روش پیشنهادی OPSCE بسیار کمتر از روش PSCE می‌باشد و روش PSCE در اجراهای مختلف تغییرات دقت بیشتری داشته است.

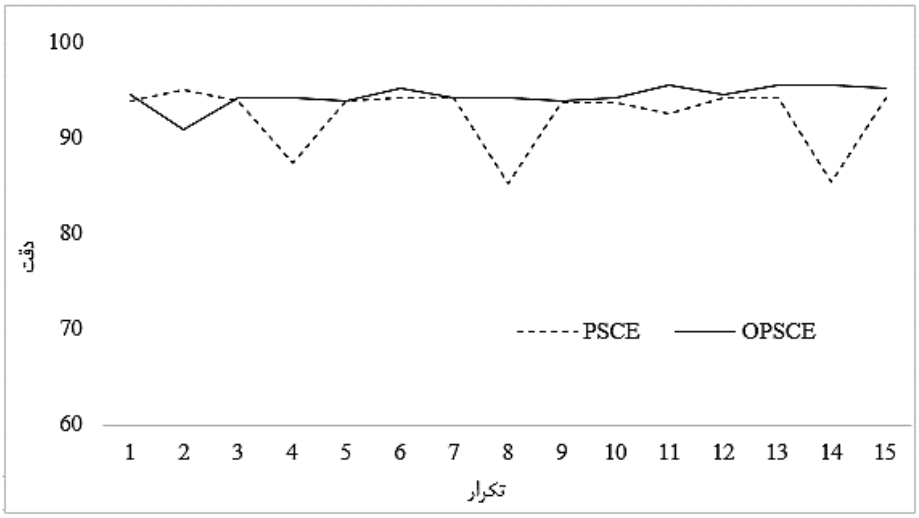
شکل ۳- مقایسه دقت الگوریتم OPSCE و PSCE روی مجموعه داده Ionosphere



شکل ٤- مقایسه دقت الگوریتم OPSCE و PSCE در مجموعه داده Bupa



شکل ٥ - مقایسه دقت الگوریتم OPSCE و PSCE در مجموعه داده Wine



شکل ٦- مقایسه دقت الگوریتم OPSCE و PSCE در مجموعه داده Iris

در جدول ۲ میانگین دقت الگوریتم OPSCE و الگوریتم PSCE با یکدیگر مقایسه و نتایج آن گزارش شده است. همان طور که مشاهده می‌شود، میانگین دقت الگوریتم OPSCE نسبت به روش PSCE در مجموعه داده‌ها ، بهبود پیدا کرده است و افزایش محسوسی داشته است.

جدول ۲- میانگین دقت در الگوریتم OPSCE و PSCE

مجموعه داده	میانگین دقت روش OPSCE	میانگین دقت روش PSCE
Ionosphere	۸۸/۴۲	۸۶/۵۰
Bupa	۹۳/۷۱	۸۷/۹۶
Wine	۹۲/۰۹	۹۰/۰۶
Iris	۹۴/۴۶	۹۲/۴۴

۶-نتیجه گیری

خوشه‌بندی ترکیبی به عنوان روشی برجسته برای بهبود نتایجی بهتر از لحاظ دقت، کیفیت و استحکام، با استفاده از ترکیب نتایج حاصل از چندین خوشه‌بندی پایه، ظهور کرده است. همچنین برای بالاتر بردن سرعت همگرایی و دستیابی به نتایج بهینه در روش‌های خوشه‌بندی ترکیبی می‌توان از الگوریتم‌های فراابتکاری استفاده کرد به همین دلیل در این مقاله راه کاری برای خوشه‌بندی ترکیبی مبتنی بر الگوریتم خوشه‌بندی ازدحام ذرات ارائه شد و مورد تجزیه و تحلیل قرار گرفت. در الگوریتم PSC بر خلاف دیگر الگوریتم‌های بهینه‌سازی ازدحام ذرات، هر ذره نشان دهنده یک مرکز خوشه است نه کل افراز. این ویژگی همراه با خاصیت خود سازماندهی، PSC را تبدیل به یک الگوریتم ساده و رقابتی و کارامد کرده است. خوشه‌بندی ترکیبی شامل دو مرحله می‌باشد: ابتدا چندین افراز پایه از مجموعه داده بدست می‌آید و سپس افرازهای پایه به عنوان ورودی دریافت و با استفاده از تابع ترکیب، یک افراز نهایی ایجاد می‌شود. که مسلما تابع ترکیب مهمترین عامل در الگوریتم‌های خوشه‌بندی ترکیبی می‌باشد.

در این مقاله از الگوریتم PSC، در هر دو مرحله تولید و تابع ترکیب استفاده شده است. در این تابع، خوشه‌بندی بر روی زیر مجموعه‌ای از نمونه‌های افرازهای پایه اجرا می‌شود. قبل از تولید افرازهای پایه، پارامترهای PSC را با استفاده از الگوریتم PSO بهینه کردیم تا خوشه‌های اولیه بهتری تولید شود. و سپس وارد مرحله تولید افرازهای اولیه می‌شویم و سپس در مرحله

هرس تکاملی با استفاده از عملگر انتخاب زیر مجموعه‌ای از افرازها انتخاب شدند (به عبارت دیگر انتخاب افرازها قبل از ترکیب نهایی). و در نهایت از الگوریتم PSC برای ترکیب افرازهای انتخابی مرحله قبل استفاده و افراز نهایی تولید شد.
الگوریتم OPSCE بر روی تعدادی مجموعه داده اجرا و نتایج آن مورد ارزیابی قرار گرفت و نتیجه آن با روش PSCE مقایسه شد. و دریافتیم به طور کلی انتخاب تعداد کمی از افرازهای پایه (زیر ۲۰٪ از تعداد کل افرازها) نتایج بهتری را در بر دارد. با ارزیابی الگوریتم OPSCE دریافتیم که تغییرات دقت الگوریتم نسبت به روش PSCE بهبود پیدا کرده است و همچنین دقت الگوریتم OPSCE افزایش محسوسی پیدا کرده است.

منابع

[1]. *A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," ACM computing surveys (CSUR), vol. 31, pp. 264-323, 1999.*

[2]. *J. V. De Oliveira and W. Pedrycz, Advances in fuzzy clustering and its applications: John Wiley & Sons, 2007.*

[3]. *A. K .Jain, "Data clustering: 50 years beyond K-means," Pattern recognition letters, vol. 31, pp. 651-666, 2010.*

[4]. *A. L. Fred and A. K. Jain, "Data clustering using evidence accumulation," in Pattern Recognition, 2002. Proceedings. 16th International Conference on, 2002, pp. 276-280.*

[5]. *S. Vega-Pons and J. Ruiz-Shulcloper, "A survey of clustering ensemble algorithms," International Journal of Pattern Recognition and Artificial Intelligence, vol. 25, pp. 337-372, 2011.*

[6]. *R. Ghaemi, M. N. Sulaiman, H. Ibrahim, and N. Mustapha, "A survey: clustering ensembles techniques," World Academy of Science, Engineering and Technology, vol. 50, pp. 636-645, 2009.*

[7]. *X. Hu and I. Yoo, "Cluster ensemble and its applications in gene expression analysis," in Proceedings of the second conference on Asia-Pacific bioinformatics-Volume 29, 2004, pp. 297-302.*

[8]. J.-P. Barthélemy and B. Leclerc, "The Median Procedur*e for Partitions," Partitioning data* sets, *vol. 19, pp. 3-34, 1993.*

[9]. *A. Fred and A. Lourenço, "Cluster ensemble methods: from single clusterings to combined* solutions,*" in Supervised and unsupervised ensemble methods and their applications, ed: Springer, 2008, pp. 3-30.*

[10]. *X. Z. Fern and C. E. Brodley, "Random projection for high dimensional data clustering: A* cluster *ensemble approach," in Proceedings of the 20th international conference on machine* learning *(ICML-03), 2003, pp. 186-193.*

[11]. *B. Minaei-Bidgoli, A. Topchy, and W. F. Punch, "Ensembles of partitions via data resampling," in Information Technology :Coding and Computing, 2004. Proceedings. ITCC 2004. International Conference on, 2004, pp. 188-192.*

[12]. *A. Topchy, B. Minaei-Bidgoli, A. K. Jain, and W. F. Punch, "Adaptive clustering ensembles," in Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on, 2004, pp. 272-275.*

[13]. *A. Topchy, A. K. Jain, and W. Punch, "A mixture model for clustering ensembles," in Proceedings of the 2004 SIAM international conference on data mining, 2004, pp. 379-390.*

[14]. *A. Strehl and J .Ghosh, "Cluster ensembles---a knowledge reuse framework for combining multiple partitions," Journal of machine learning research, vol. 3, pp. 583-617, 2002.*

[15]. *S. Dudoit and J. Fridlyand, "Bagging to improve the accuracy of a clustering procedure," Bioinformatics, vol. 19, pp. 1090-1099, 2003.*

[16]. *A. Weingessel, E. Dimitriadou, and K. Hornik, "An ensemble method for clustering," in Proceedings of the 3rd international workshop on distributed statistical computing, 2003.*

[17]. S. C. Cohen and L. N. de Castro, "Data clustering with particle swarms," in Evolutionary Computation, 2006. CEC 2006. IEEE Congress on, 2006, pp. 1792-1798.

[18]. J. Kennedy and R. Eberhart, "Particle swarm optimization 1995 IEEE International Conference on Neural Networks Proceedings," ed: Vols, 1942.

[۱۹] علیزاده، حسین و مشکی، محسن و پروین، حمید و مینایی بیدگلی، بهروز; خوشه بندی ترکیبی مبتنی بر زیرمجموعه ای از خوشه های اولیه, پردازش علائم و داده ها، سال هفتم، ۱۳۸۹، صفحه ۱۹.

[20]. *R. Baumgartner, R. Somorjai, R. Summers, W. Richter, L. Ryner, and M. Jarmasz, "Resampling as a cluster validation technique in fMRI," Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine, vol. 11, pp. 228-231, 2000.*

[21]. *V. Roth, T. Lange, M. Braun, and J. Buhmann, "A resampling approach to cluster validation," in Compstat, 2002, pp. 123-128.*

[22]. *V. Roth, M. L. Braun, T. Lange, and J. M. Buhmann, "Stability-based model order selection in clustering with applications to gene expression data," in International Conference on Artificial Neural Networks, 2002, pp. 607-612.*

[23]. *A. L. Fred and A. K. Jain, "Learning pairwise similarity for data clustering," in Pattern Recognition, 2006. ICPR 2006. 18th International Conference on, 2006, pp. 925-928.*

[24]. *Y. Yang and M. Kamel, "Clustering ensemble using swarm intelligence," in Swarm Intelligence Symposium, 2003. SIS'03. Proceedings of the 2003 IEEE, 2003, pp. 65-71.*

[25]. *L.-y. Yang, J.-y. Zhang, and W.-j. Wang, "Cluster ensemble based on particle swarm optimization," in Intelligent Systems, 2009. GCIS'09. WRI Global Congress on, 2009, pp. 519-523.*

[۲۶].حسین پور، محمدجواد و پروین، حمید; انتخاب خوشه های اولیه به کمک الگوریتم های هوشمند برای مشارکت در خوشه بندی ترکیبی، مهندسی برق و الکترونیک ایران، سال سیزدهم، صفحه ۱۶۳

[27]. *S. Chatterjee and A. Mukhopadhyay, "Clustering ensemble: a multiobjective genetic algorithm based approach," Procedia Technology, vol. 10, pp. 443-449, 2013.*

[28]. *A. Ahmadi, F. Karray, and M. Kamel, "Particle swarm clustering ensemble," in Proceedings of the 10th annual conference on Genetic and evolutionary computation, 2008, pp. 159-160.*

[29]. *A. Ahmadi, F. Karray, and M. Kamel, "Multiple cooperating swarms for data clustering," in Swarm Intelligence Symposium, 2007. SIS 2007. IEEE, 2007, pp. 206-212.*

[30]. *J. V. de Oliveira, A. Szabo, and L. N. de Castro, "Particle Swarm Clustering in clustering ensembles: Exploiting pruning and alignment free consensus," Applied Soft Computing, vol. 55, pp. 141-153, 2017.*